Department of Computer Science

Computational Approach for an Automatic Facial Appearance Outcome Measure of Cleft Lip Surgical Repair Using Digital Images

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Paul Bakaki

Supervisory Team

Director of Studies
Prof. Yonghuai Liu

Co-Supervisors
Prof. Ella Pereira
Dr. Aristides Tagalakis
Mr. Bruce M. Richard

SEPTEMBER 2023

Abstract

Cleft lip (CL) is a common congenital facial anomaly that affects several individuals worldwide and is treated through a surgical procedure. The appearance outcome following the procedure is normally qualitatively assessed by human experts. However, human experts are naturally constrained with fatigue, potential bias, replicability weaknesses, and inconsistencies. Presence of large datasets presents further challenges to human qualitative assessment.

This study aims to develop and validate novel computational techniques that can automatically, objectively, and quantitatively assess the CL treatment outcome. Consequently, the study assesses the effectiveness of CL treatment using computational techniques. Using digital imagery, this study has led to the development and validation of some computational techniques to aid with automatic, objective, and quantitative assessment CL treatment outcome.

The first approach investigated the appearance and shape of the mouth lips as a region of interest for analysis. The bisector of the line connecting the mouth corners was estimated as the vertical symmetric axis of the mouth borderline. By splitting the mouth blob into two parts, the two parts were analysed for structural similarity. Consequently, a numeric score ranging from 1 to 5 was generated and validated using Pearson correlation coefficient against human-assigned numeric scores.

Secondly, a novel technique for adaptive detection of the symmetric axis of the cropped facial images of patients after CL treatment was developed. A Gaussian filter was applied to smoothen the images to compress potential noise on the subsequent tasks. Segmentation using a bilateral semantic network was applied to detect the facial components in each region of interest in the facial image. Applying the previous approach led to improved validation metrics using Pearson's correlation coefficient.

The final approach explored transfer learning using CNNs in a regression analysis study. An investigation was completed for the impact of transfer learning on regression scoring and assessed its potential in overcoming dataset limitation challenges. Through extensive experimentation and evaluation on diverse regression scoring combinations, different numeric assessment prediction results were generated. It was demonstrated that appearance assessment through CNN transfer learning is significantly competitive and better than human expert assessment and scoring. Competitive metrics using RMSE, MAE, and Pearson correlation were generated.

Overall, this thesis presents a comprehensive computational approach for automatic appearance assessment estimation of CL treatment using digital imagery. It offers insights into the potential of advanced computational techniques, such as shape analysis and deep learning, to provide accurate and objective assessments. The findings contribute to the field of CL treatment evaluation and pave the way for further advancements in automated appearance assessment methodologies.

Acknowledgement

The completion of this thesis is due to the steady support of my Principal Supervisor, Prof Yonghuai Liu, for the persistent support, guidance, patience, and motivation throughout my PhD journey. I have enjoyed watching and learning from his immense knowledge and experience. I am forever thankful to you for the time you spared and the frequent discussions and meetings to see me through this journey.

I thank the supervisory team, Prof Ella Pereira, Dr Aris Tagalakis, and Dr Bruce Richard. You have mentored, advised me immensely, and morally picked me up when a lot failed. You are academic mentors and guardians alike. Thank you. I thank Dr Quanbin Sun and Prof Nik Bessis for their insights at the very start of this research. Prof Ardhendu Behera is highly appreciated for his intriguing intuitions towards this work.

Furthermore, I appreciate the different staff members at the Department of Computer Science, the Graduate School, and the International Office for the administrative guidance throughout the research journey.

Throughout this research work, I have encountered several researchers, some fellow students while others are seasoned and established researchers. They have been immense sources of inspiration to me. Naming them would be biased because the list is long. Thanks for your inspiration and collaboration.

This research would not have been possible without funding from the Graduate Teaching Assistantship scheme at the university. Therefore, the visionaries of the scheme are highly appreciated. The Cleft Care UK (CCUK) is appreciated for providing the research datasets used throughout this research study.

The religious groups I belong to are continuous sources of inspiration. I thank the following Rev. Deacons for their perseverance in prayer with me: Justin, Peter, Alison, Patrick, Dr Peter Ankomah, Dr Phillip K. Gichuru, Ms Pam Ashcroft, and the community at St. Francis of Assisi RC Church, Skelmersdale. Thank you for your prayers.

Special appreciation to my family: my wife Rita and daughters Cissy and Theresa, my parents, my Mum Deziranta, Mr. and Mrs. Ngobi, and siblings Noah, among the many, for your patience. You missed me and prayed for me throughout, thank you! Additional family members I continue to thank are: Mr. and Mrs. Rennie, Mr. and Mrs. Olusori, Mr. and Mrs. Mendy, Mr. and Mrs. Katali, and Mr. and Mrs. Arua for their continuous support during the period of this PhD research study.

I thank the Almighty God, the Alpha and Omega, for the divine providence, grace, and strength, without which this work would have ceased. Eternally, I dedicate this thesis to Yahweh!

Documentations Generated from the Project

The following documents are as a result of this PhD research.

Mandatory Submissions

1. Research proposal for the registration viva (March 2020)
2. Progress Report to signal the halfway mark of the research study (April 2021)

Publications

1. Bakaki, Paul, Bruce Richard, Ella Pereira, Aristides Tagalakis, Andy Ness, and Yonghuai Liu. 2021. “Shape Analysis Approach Towards Assessment of Cleft Lip Repair Outcome.” Pp. 165–74 in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 13052 LNCS.
2. Bakaki, Paul, Bruce Richard, Ella Pereira, Aristides Tagalakis, Andy Ness, Ardhendu Behera, and Yonghuai Liu. 2022. “Key Landmarks Detection of Cleft Lip-Repaired Partially Occluded Facial Images for Aesthetics Outcome Assessment.” Pp. 718–29 in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 13232 LNCS.
3. In advanced stages of preparation -Title: “**A Regression CNN Framework Using Transfer Learning for Aesthetic Assessment of Partial Facial Images after Cleft Lip Treatment”.**

List of Abbreviations

nD	n dimensions, n = 1, 2, 3 …
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
API	Application Programming Interface
CCUK	Cleft Care UK
CLAPA	Cleft Lip And Palate Association
CL	Cleft lip
CLP	Cleft lip and palate
CNN	Convolutional Neural Network
Conv	Convolutional Filter or Convolution
CP	Cleft Palate
CPU	Central Processing Unit
CRANE Database	Cleft Registry and Audit NEtwork
CT	Computed Tomography
CuDA	Compute Unified Device Architecture
CuDNN	CuDA Deep Neural Network
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
ELU	Exponential Linear Unit
ENT	Ear Nose Throat
GAN	Generative adversarial network
GAP	Global Average Pooling
GMP	Global Max Pooling
GPU	Graphics Processing Unit
HSV	Hue, Saturation, and Value
MAE	Mean Absolute Error
OFC	Orofacial cleft
OpenCV	Open Computer Vision
PCC	Pearson Correlation Coefficient
ReLU	Rectified Linear Unit
ResNet	Residual Network
RGB	Red, Green, and Blue
RMSE	Root Mean Square Error
RMSprop	Root Mean Squared Propagation
RNN	Recurrent Neural Network
RoI	Region of Interest
SSIM	Structural Similarity Index Measure
TPU	Tensor Processing Unit
VGG	Visual Computing Group
YOLO	You Only Look Once

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Overview

There is a catchy phrase in the Physical Science and Technology Manual which states that: "*image processing and analysis often require fixed sequences of local operations to be performed at each pixel of an image*". Relatedly, digital image processing is the manipulation of images with digital computers using different processing techniques at each stage of the processing pipeline (Pratt, 1994, Gonzalez and Woods, 2008). Different digital image processing techniques may output unique analysis results.

Digital image analysis can be applied to different domains (Girod, 2015). Some dominating fields include traffic flow monitoring/transportation, industrial monitoring, earth observation/satellite technology, security systems design/ monitoring and medical technology (Shih, 2017; Girod, 2015). In traffic and transportation, image analysis influences the development of intelligent transport systems and informs the planning and development of safety-aware physical infrastructure. The scene's complexity greatly influences image and video analysis in transportation and traffic management, hence the need to identify all the objects, such as traffic flow at a road junction. Different image representation colour models and segmentation techniques positively influence multiple object identification (Buch, Velastin and Orwell, 2011, Lira et al., 2016). Monitoring driver behaviour has improved road and traffic safety by influencing better car designs (Behera et al., 2020, Wharton et al., 2021). In security systems, automatic analysis of security footage eliminates bias (Chen, Surette and Shah, 2020). Altogether, considering the processing load born by security cameras, analysis of their performance is important to ensure the reliability of the captured footage (Chen, 2005). Similarly, the automation of appearance analysis and assessment is a primary outcome of this research study. Consequently, performance analysis of automatic appearance analysis and assessment should be monitored using some conventional evaluation metrics.

This has immensely contributed to the advancement of computer vision techniques to efficiently analyse all the forms of digital images through a processing pipeline summarised in Figure 1.1. Since digital images are potentially useful sources of evidence and heritage preservation, their acquisition techniques and storage

mechanisms are fundamental to quality preservation and quantitative analysis (Gonzalez and Woods, 2008).

As indicated in Figure 1.1, image acquisition and storage are vital steps towards generating (medical) image datasets, both in controlled environments and in the wild.

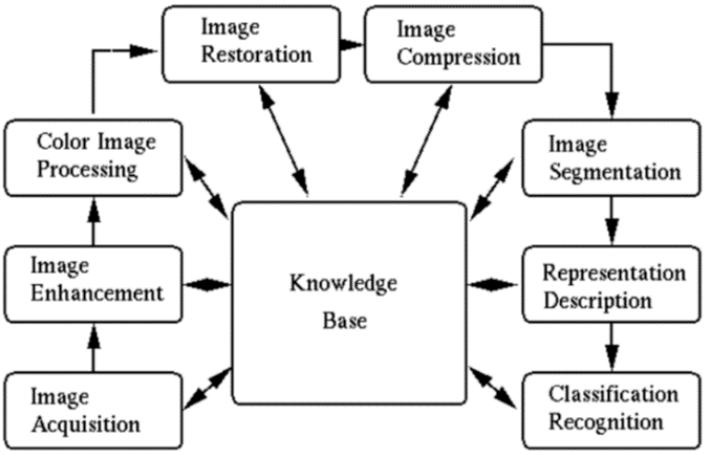

Figure 1. 1: Image processing pipeline - adopted from (Girod, 2015), contains all the fundamental image processing operations required in typical computer vision model.

In medical technology, for instance, image analysis techniques are applied to modern ways of healthcare delivery (Bankhead et al., 2017, Ker et al., 2017). Specifically, outcomes such as precision with surgical procedures, improved surgical operations that lead to faster recovery and diagnostic advisory for internal organ failure among others have emerged. Overall, a summary of cutting-edge research on machine learning applications in medical digital image analysis has been presented by (El-Baz, Gimel'Farb and Suzuki, 2017). The prevalent computed tomography (CT) scanner generates images of profound interest to medical practitioners and researchers alike to aid the development of computer vision applications and advance diagnostic research studies.

Images captured from the same scene, by the same device, may have different interpretations by different people. These possibilities arise from both experts and non-trained audiences. Three leading reasons potentially explain these inconsistencies (Aeffner et al., 2017):

1. Human beings are better at qualitative assessment, yet images are better interpreted quantitatively.
2. An individual's mental state, background and knowledge influences their interpretation of any image.

3. Visual impairment and cognitive exhaustion could lead to biased interpretation.

Therefore, misinterpretations of a given scene, as may be portrayed in a medical image such as the visual outcome of a cleft lip (CL) treatment, could have adverse implications on patients' well-being. Consequently, it is important to employ unbiased techniques for image analysis studies (Xu et al., 2014, Joskowicz, 2017).

Research advancement in image analysis techniques has contributed to introducing medical imaging and analysis systems. The existing medical image processing systems and applications such as CT scanning, X-rays, ultrasound and nuclear medicine manufacturing have revolutionised computational and mathematical research in medicine (Drahansky et al., 2016, El-Baz, Gimel'Farb and Suzuki, 2017, Litjens et al., 2017). Furthermore, increased accessibility to relevant (big) datasets has contributed to the development and validation of machine learning algorithms, leading to improved medical-based applications with better performance (Tajbakhsh et al., 2016, Razzak, Naz and Zaib, 2018). Solutions based on deep neural networks have successfully diagnosed the following illnesses and/or conditions using image analysis and classification: diabetic retinopathy, histological and microscopical elements detection, gastrointestinal (GI) diseases, tumour, and cardiac illnesses, among others (Razzak, Naz and Zaib, 2018). Large medical datasets are often required to develop and validate diagnostic and/or treatment computational models (Oakden-Rayner, 2020). In some cases, fine-tuning and adapting existing computational models leads to more effective performance outcomes and innovations. Likewise, innovations through research studies can be spurred by knowledge and adaptation of existing management frameworks. Furthermore, computerisation and optimisation of such frameworks often leads to better solutions (Oh, Yang and Yi, 2015).

The current treatment and management framework for cleft lip can be summarised in the workflow in Figure 1.2, while Figure 1.3 shows the meaning of the symbols used. Figure 1.2 comprises several decision-making steps and processes.

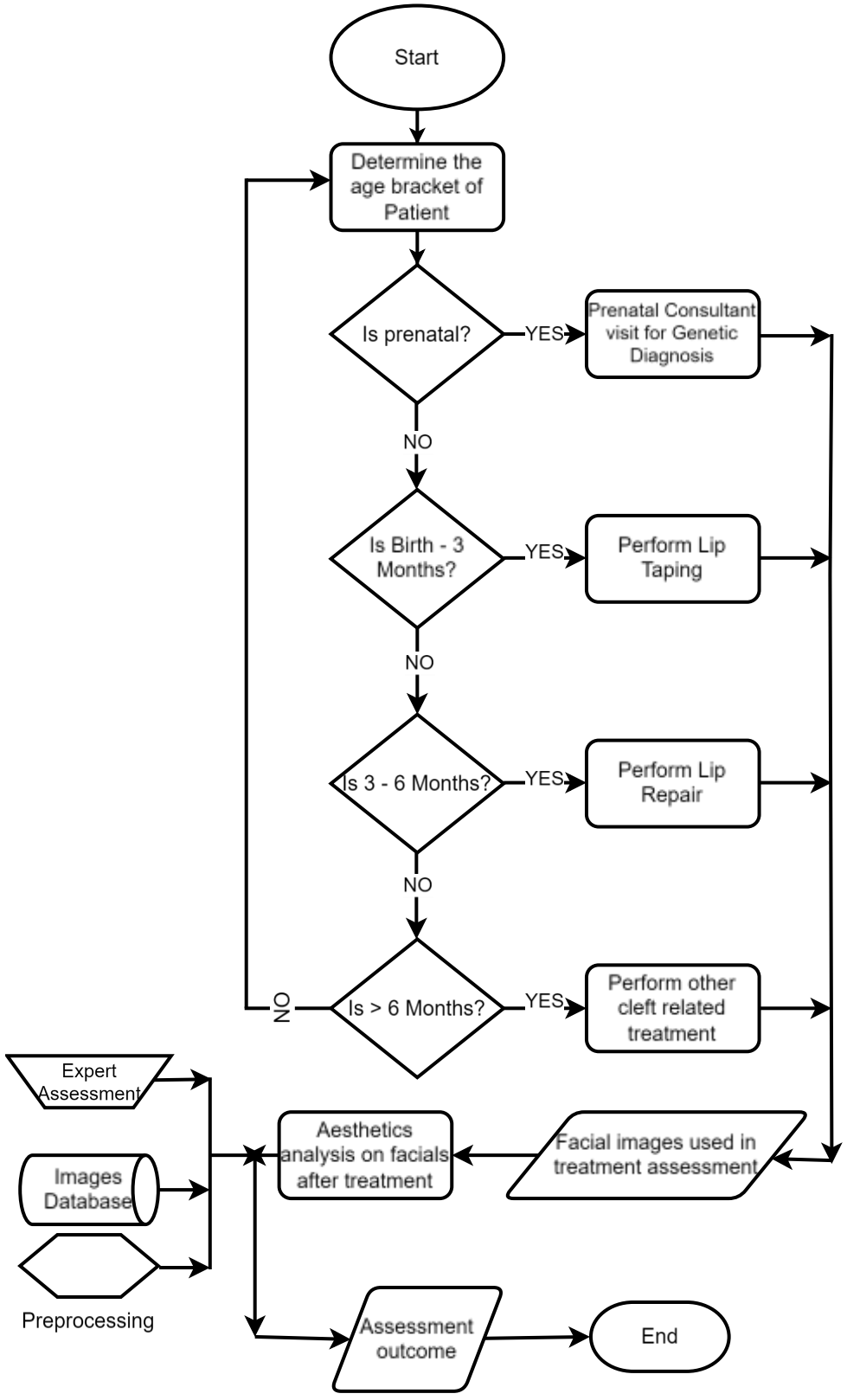

Figure 1. 2: Treatment and Management Workflow for cleft lip from prenatal stage to at least six months.

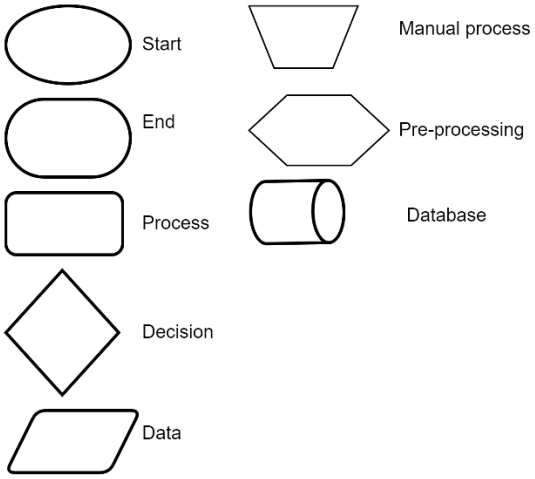

Figure 1. 3: Meaning of symbols used in flow chart design.

The steps in Figure 1.2 are subsequently explained below:

1. **Prenatal Counselling and Support**

In many contexts, ultrasound scanning of pregnant mothers is a common practice. This is intended to determine the structural growth of the foetus. This may reveal details of the foetus's facial structures, among other discoveries (Bäumler et al., 2011).

The prenatal diagnosis of cleft lip may allow early counselling and support to expectant parents. Genetic analysis and prenatal counsellors play a vital role in this exercise.

Providing accurate information about cleft lip, treatment options, and potential challenges helps parents prepare emotionally and make informed decisions (Costa et al., 2023). Genetic counselling may be offered to discuss the potential causes and recurrence risks in future pregnancies (Bronshtein, Blumenfeld and Blumenfeld, 1996).

2. **Treatment through a Multidisciplinary Team Approach**

A multidisciplinary team of specialists, such as paediatricians, plastic and maxilla-facial surgeons, nurses, speech therapists, audiologists, ENT specialists, Orthodontists, Psychologists and social workers, collaborates to provide comprehensive care (Frederick et al., 2022). The team develops an individualised treatment plan based on the severity and specific needs of the cleft lip condition.

In many cases, feeding can be challenging for infants with cleft lip due to difficulty creating a seal and proper suction. Special feeding techniques, such as using specialised bottles and nipples or utilising alternative feeding methods like nasogastric or orogastric tube feeding, may be recommended. The involvement of a feeding specialist or lactation consultant can be beneficial in assisting parents in managing feeding difficulties (Khanchezar et al., 2019).

After birth, lip taping is decisive for successful future stages of cleft lip treatment and should be performed upon diagnosis, within three months from birth.

3. Surgical Repair

A baby's organs and features are grown enough to safely endure a surgical procedure between three and six months. Therefore, surgical cleft lip repair is typically performed around 3 to 6 months of age, depending on the infant's overall health and growth (Shaye, Liu and Tollefson, 2015). Besides, patients have a high chance of full recovery and better appearance outcomes over time after surgery.

Plastic surgeons with expertise in cleft lip repair perform the surgical procedure, which involves correcting the separation of the mouth lip tissues and creating a more natural appearance (Sischo et al., 2016). The surgical technique employed may vary depending on the individual case, the surgeon's preference, and the operations' protocols or guidelines.

4. Post-Surgical Care and Monitoring

After cleft repair, care and monitoring involve attending to the needs of the patients, and assessing, or rating the treatment outcome. Therefore, close monitoring and follow-up care are essential to ensure proper healing and address any complications or concerns. These possibly influence the decisions on any potential future surgical repairs (Taib et al., 2015, Burg et al., 2016).

Regular appointments with the surgical team allow for the evaluation of surgical outcomes, management of any issues related to scarring, feeding difficulties, or speech development, and adjustment of the treatment plan as needed. In some cleft care centres, an imaging unit may be available to collect and store pre- and post-treatment visuals (in a database), to aid research studies (CLAPA, 2022).

5. **Qualitative Assessment Framework**

After surgical treatment, an appearance assessment is required to determine the efficacy of the procedure. Expert assessment is manually conducted through anthropometric measurements or visual assessment of facial images. In some cases, web-based semi-automatic approaches are used, (Bella et al., 2016). Therefore, preprocessing images into understandable content is desired, implying that the visuals should reside in a database to enable dynamic access by the assessors. Additionally, for safe and secure storage.

1.2 Background and Context

There has been great transformation of surgical and clinical practices by developing innovative medical technologies and imaging systems. Medical imaging systems have facilitated the acquisition and development of (large) image datasets. These datasets can be used to pre-plan surgical procedures, analyse surgical outcomes, and improve patient treatment, care, and well-being. If researchers are equipped with such resources, image analysis techniques can further improve surgical or clinical planning practices, progress monitoring of procedures, and evaluation of outcomes following surgical interventions (Zhou et al., 2021).

Analysis of pre-surgical or post-surgical images involves using computer algorithms to analyse, process and interpret medical images to extract clinically relevant information. Image analysis techniques can be used to improve the accuracy and efficiency of surgical procedures, reduce the risk of complications, and enhance patient outcomes (Hashimoto et al., 2018). Advanced computer algorithms can be designed to detect growths, segment body organs, and classify tissue types in medical images. Computer algorithms can also be used to analyse surgical videos and extract features indicative of surgical outcomes, such as incision repair, tissue damage repair, congenital malformation repair, and surgical duration (Ali et al., 2022).

One of the most renowned conditions with high economic and social importance is the cleft lip (CL). The Center for Disease Control and Prevention[1] explains that "A cleft lip happens if the tissue that makes up the mouth lip does not join completely before birth. This results in an opening in the upper mouth lip. The opening in the mouth lip can be

[1] https://www.cdc.gov/ncbddd/birthdefects/cleftlip.html

a small slit or a large opening that goes through the lip into the nose". CL is a common congenital deformity that affects approximately 1 in 700 live births worldwide (Zhang et al., 2019). The three potential biological conditions summarised below attempt to explain the genesis of the CL condition and sometimes the cleft palate (CP) condition.

1. During embryogenesis, the development of the face involves the fusion of multiple facial processes. The primary palate, formed by the fusion of the medial nasal processes, contributes to the lip and the anterior part of the palate. The maxillary processes on either side of the primary palate fuse with the medial nasal processes to complete the formation of the upper lip. Failure of these processes to fuse completely results in a right or left or bilateral cleft lip. (Cash, 2012).
2. Though not fully understood, the exact cause of CL is credited to a classical combination of genetic and environmental factors. Genetic studies have identified several genes that play a role in the development of the face and lip, including the IRF6, MSX1, and BMP4 genes. However, the inheritance pattern is often complex, involving genetic and environmental interactions. Some socio-environmental factors can also contribute to the formation of the CL condition. These are maternal smoking, alcohol consumption, certain medications, maternal infections during pregnancy (such as rubella), and exposure to environmental toxins (Katsaros, 2013).
3. The molecular and cellular mechanisms underlying CL involve disruptions in the signalling pathways and cellular processes that regulate facial development. These include processes such as cell migration, proliferation, adhesion, and apoptosis. Imbalances in key signalling molecules, such as transforming growth factor-beta (TGF-β) and fibroblast growth factors (FGFs), can lead to defective fusion of the facial processes, resulting in cleft lip formation (Owens, Jones and Harris, 1985, Jiang, Bush and Lidral, 2006, Jamilian et al., 2017).

Surgical intervention (s) is/are typically required to restore normal mouth lip structure and function. Hence, assessing the outcome of a CL surgical repair is a revered research topic, both in the medical and computing domains (Medina et al., 2017).

While traditional assessment methods such as visual evaluation (qualitatively) and physical measurements (semi-quantitative) provide valuable information on cleft lip repair outcomes, there is a need for more objective and quantitative methods to assess treatment efficacy (Stein et al., 2019). Computational techniques, such as 2D/3D imaging and analysis, computer vision, and machine learning, offer new ways to objectively evaluate CL repair outcomes and identify areas for improvement.

Imaging frameworks (2D/3D) and analysis have emerged as valuable tools for assessing the outcomes of CL repair surgeries. By capturing high-resolution images of the mouth lip and surrounding structures, surgeons, carers, and researchers can analyse the results of surgical interventions in a more comprehensive and quantitative manner. Computer vision techniques, such as image segmentation and feature extraction, can be applied to these images to identify subtle differences in lip structure and potential symmetry or similarity or correctness which are important indicators of treatment efficacy (Huqh et al., 2022, Jeong et al., 2022). On the contrary, researchers and clinicians are interested in understanding the quantitative differences leading to asymmetry, incorrectness or dissimilarity of the mouth lips following treatment.

Computational approaches inclined to machine learning algorithms have also shown promise in assessing cleft lip treatment outcomes. By training on available datasets of cleft lip images and patient records (such as qualitative evaluation of treatment outcome), machine learning models can learn to accurately predict treatment outcomes and identify areas for improvement. These models can also extract new features and parameters from CL images that may be difficult to measure using traditional assessment methods (Bella et al., 2016). For example, machine learning models can be used to analyse the texture and colour of the mouth lips, which can provide important information on the healing process and surgical treatment effectiveness (Chowdhury et al., 2022). Semantic analysis of CL treatment outcome images coupled with transfer learning techniques can be exploited for the development of quantitative assessment methods.

Therefore, as stated before, the development and validation of computational methods for assessing CL repair appearance outcomes is an active research area. In addition, using computational techniques has led to identifying new research questions and areas for improvement in CL treatment (Knoops et al., 2019). The use of computational

approaches for assessing CL surgical treatment outcomes has the potential to significantly improve the accuracy and objectivity of CL assessment. By developing and validating new computational techniques, surgeons, carers, and researchers can identify areas for improvement in CL treatment and ultimately improve outcomes for patients with this congenital deformity.

The care and treatment of CL patients requires a multi-disciplinary team. The communication between the different members of the care team should be simplified and precise. Using computational approaches in CL repair appearance outcome assessment has potential to improve and simplify the communication arrangements through visualisation. Visualisation is highly effective with provision and presentation of complete vision of a situation, by presenting assessment estimations in a more rational and understandable manner (Wang, Li, et al., 2020)

1.2.1 Social-economic Considerations of Maxillofacial Anomalies

Congenital malformations such as cleft lip and/or cleft palate conditions have significant economic importance due to their impact on healthcare systems, individual and family expenditures, and overall societal costs (Galloway, Davies and Mossey, 2017, Thompson et al., 2017, Salari et al., 2022). The economic implications are multifaceted and encompass various aspects of life, both in the developed and least developed contexts:

1. **Prevalence**

Different countries have contrasting prevalence of cleft lip conditions in their populations, but the different prevalence is between the ethnicities represented in those countries rather than any other factor. However, the condition is most prevalent among Chinese Asians, then Caucasians and Indo-Aryan peoples equally and least prevalent among blacks (Bloomfield and Liao, 2015, Salari et al., 2022). In the United Kingdom, 1 in every 600-700 live births is born annually with a Cleft lip and/or palate (Cleft Registry and Audit NEtwork, 2020). The CRANE database indicates that the conditions affect more boys than girls The distribution of the different maxillofacial conditions is presented in Figure 1.4, adapted from (CRANE, 2021)

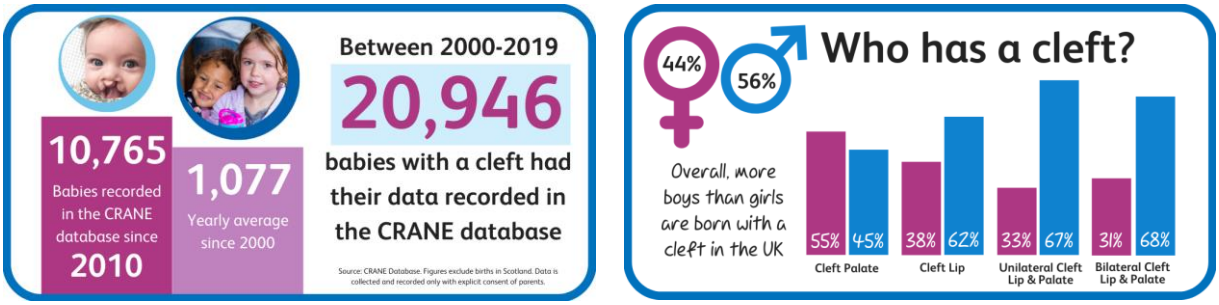

Figure 1. 4: Prevalence of different maxillofacial conditions according to the CRANE database in the UK. (CRANE, 2021)

2. Healthcare Costs

These maxillofacial malformations require comprehensive medical treatment, including surgeries, dental care, orthodontics, and speech therapy. These treatments can involve multiple surgeries over several years, leading to substantial healthcare expenses for families and healthcare systems. Because these are high-risk conditions, medical interventions such as specialist consultations and follow-up care contribute to the economic burden. Direct medical costs vary in the range of £5,000 and £15,000 per procedure, depending on one's location in the western world, (Galloway, Davies and Mossey, 2017).

3. Reduced Productivity and Income Loss

Individuals born with cleft lip and/or palate often face challenges in speech development, hearing, and dental health. These issues can affect educational attainment and employment opportunities, leading to reduced productivity and potential income loss over their lifetime (Baigorri et al., 2021). Additionally, caregivers, often parents or guardians, may need to take time off work to care for their children during treatment and recovery, further impacting family income. The study by (Baigorri et al., 2021) indicates frequent transport costs incurred by families of people affected by cleft related challenges to access speech therapy services far away from their locations. This further impacts on their economic plight.

Additionally, there may be increased number of general anaesthetics in the first 3 years of life or more hospital appointments for surgeries during the age of between 5 and 13 years (Grewal et al., 2021). This directly impacts on how much education level they may attain, potentially exposing them to low paying jobs.

4. **Rehabilitation and Psychosocial Support Services**

Cleft-affected individuals may require ongoing rehabilitation and support services, such as speech therapy and psychological counselling. These services contribute to both direct and indirect costs, as they require financial resources and time commitments. Additionally, affected individuals may experience psychological and social challenges, including self-esteem issues, bullying, and stigmatization. Addressing these emotional and social aspects often involves interventions that carry economic implications (Emeka et al., 2017, Garcia-Marin, 2021).

5. **Research and Development**

Research aimed at improving surgical techniques, treatment outcomes, and interventions for cleft lip and palate contributes to economic investment in the healthcare and medical research sectors. This increases the potential for better wellbeing of cleft-affected people and fosters collaborations, locally and globally (Zhang et al., 2019, Kassam et al., 2020, Sommer et al., 2023).

6. **Institutional Support**

Governments and public health systems such as the National Health Service (NHS[2]), may need to allocate resources to provide subsidised or free healthcare services for cleft-affected individuals, especially in regions with limited access to healthcare or financial resources. Additionally, Non-Government Organisations and charitable organisations[3,4] play a significant role in supporting individuals with maxillofacial conditions by providing medical rescue missions, subsidised surgeries, and rehabilitation services (CRANE, 2021).

1.3 Problem Statement

The surgical repair of a CL is a complex and challenging procedure that requires careful planning and execution to achieve optimal outcomes. The success of cleft lip repair is typically assessed by subjective visual evaluation, which can be influenced

[2] https://digital.nhs.uk/data-and-information/publications/statistical/compendium-public-health/current/chromosomal-abormalities-congential-malformations/incidence-of-cleft-palate-and-or-cleft-lip-crude-rate-at-birth-annual-p
[3] https://www.cleft.org.uk/
[4] https://www.clapa.com/

by individual experience, bias, and disposition (Schwirtz et al., 2018, Mulder et al., 2019).

Objective and quantitative assessment of cleft lip repair outcomes is crucial, for not only improving treatment outcomes and patient satisfaction, but also alert carers that protocols should be adhered to. Computational techniques, such as 2D/3D imaging, machine learning, and computer-aided simulations, have the potential to provide accurate and objective measurements of CL repair outcomes. However, developing and standardising these techniques for clinical practice is still in its early stages (Mosmuller, Mennes, et al., 2017a).

Therefore, there is a need for further research to evaluate the effectiveness and feasibility of existing assessment techniques for CL repair. For this reason, the greater need is the eventual development and validation of computational techniques, to aid objective and quantitative appearance assessment of CL repair outcome (Bozkurt and Aras, 2021). Such standardised protocols for the computational-based assessment cycle could be used objectively in clinical practice.

In summary, there is need for an objective and quantitative analysis and assessment pipeline of CL surgical treatment appearance outcomes with a possibility of improving patient satisfaction and treatment efficacy. Computational methods have promising potential to generate accurate, reliable, and reproducible results.

1.4 Research Goal

The study is designed to assess the effectiveness of CL treatment using computational techniques. The methodology involves using medical images of patients who had undergone CL surgery at 5 years old. Partial facial visual data is analysed and used to objectively assess surgical treatment outcome appearance using various computational techniques.

This research study uses the scientific method design to examine the relationship between CL treatment outcomes of a specific demographic while considering several clinical factors and attributes. In a scientific way, experimentation (such as algorithmic investigation and analysis), testing and observations take center stage before communicating outcomes (Kothari, 2004, Jennings, 2007, Chan et al., 2020). In this

study, the dataset consists of facial images taken after CL repair, of 5-year-old patients, from a collection of the Cleft Care UK (CCUK).

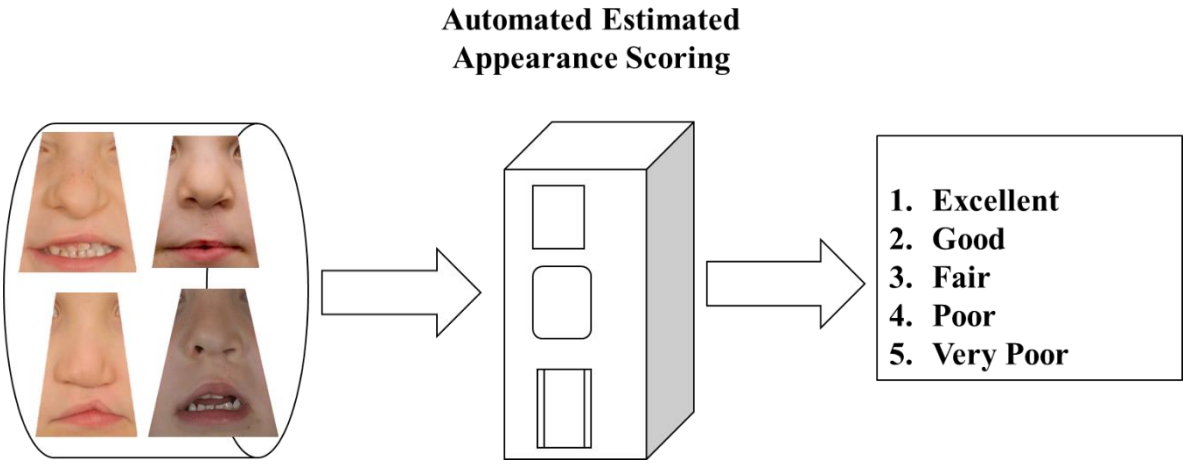

Figure 1. 5: Abstract representation of the research goal/direction. A dataset of visuals is set to be automatically evaluated into a regression result of 1, 2, 3, 4 or 5.

An image is selected from a database to predict a score in the range of 1 to 5 where 1 is 'Excellent', 2 is 'Very good', 3 is 'Fair', 4 is 'Poor' while 5 is 'Very poor'. The abstract representation of the automated process is given in Figure 1.5. Moreover, if this could become a continuous range score then it would empower research methodologies for using the score to discern differences in surgical techniques, protocols, surgeons, and Units.

1.5 Aim and Objectives

This study aims to develop and validate novel computational techniques that can automatically, objectively, and quantitatively assess the CL treatment outcome. The objective is to minimise human involvement throughout an assessment pipeline. The following objectives are applied accordingly.

1. To gain critical insights into the treatment ecosystem of cleft lip and review the assessment methods of CL treatment outcomes.
2. To identify features that influence the design of computational techniques.
3. To develop computational models that quantify the assessment of visual outcomes following CL treatment.

4. To test and validate the accuracy and effectiveness of the different developed computational techniques.
5. To compare the results of computational assessments to traditional clinical assessments of CL treatment outcomes, by identifying areas of agreement and disagreement.

1.6 Research Questions

Main Research Questions

1. Given a facial visual outcome following CL surgical treatment, in the form of a 2D photographic image, what is the possibility of designing models with automatic assessment capability?
2. Are the models capable of presenting quantitative and reproducible assessments of the facial visuals?

Sub-questions

1. What parameters are considered for the design of computational techniques associated with facial visuals?
2. Given the right parameters (from sub-question 1 above), what are the model considerations or limitations that influence accuracy and computational efficiency?
3. How does the accuracy of the different computational techniques compare regarding the assessment of CL treatment outcomes?
4. What issues affect the accuracy of computational techniques in assessing CL treatment visual outcomes?
5. How can the issues (in sub-question 4 above) be mitigated or engineered to improve the modelling process?
6. What is the comparative benefit of computational assessments techniques over traditional clinical assessment techniques?

1.7 Significance and Rationale

The socio-economic importance of CL has been underscored in several studies (Hackenberg et al., 2015, Emeka et al., 2017). CL condition is costly to treat and care

for and a source of stigma to many patients and their parents or guardians in their communities (Crerand et al., 2020). Appearance assessment of CL treatment outcomes holds significant importance in the field of maxillofacial surgery and computer science. The current image assessment methods are lengthy to perform, require many human scorers and have poor inter-rater reliability due to subjectivity and bias. In addition, they are practically ineffective for audit of treatment success and are now only used in large research studies around the world. (Sharma et al.2012). By leveraging imagery availability and machine learning techniques, computational methods offer great promise to enhance CL treatment outcome assessment's accuracy, efficiency, and objectivity.

This research study offers the hope of finding an automatic, objective, fast and reliable method of assessing post operative CL facial images and assigning an outcome score for the success of the appearance. This would be a valuable contribution to influence clinical practice and patient care. Furthermore, this study significantly contributes to the medical and computer science research domains. The following reasons explain why this research study is of significant interest:

1. Objective Evaluation: Traditional methods for appearance assessment of CL treatment outcome often rely on subjective visual evaluation by care teams, which can introduce variability and subjectivity (Mosmuller et al., 2013, Mosmuller, Mennes, et al., 2017b). Computational techniques, on the other hand, exhibit potential to offer objective, reproducible, and standardised measurements. These methods provide more reliable and reproducible assessment metrics by quantifying key parameters such as lip shape, symmetry, and nasal appearances, and hidden or semantic features (Roy, Yamasaki and Hashimoto, 2018, Talebi and Milanfar, 2018).
2. Research Advancements: The integration of computational techniques in CL treatment outcome assessment opens new avenues for research and innovation. Researchers can explore novel image analysis algorithms and develop automated assessment tools (Xu et al., 2021). In some cases, design of diagnostic tools and treatment planning, can be used to investigate the long-term effects of different surgical techniques. These advancements contribute to the overall knowledge base in the field and steer continuous improvements in patient care.

3. Additionally, computational techniques have the potential to facilitate early intervention and prediction of treatment outcomes. Machine learning algorithms can analyse datasets of CL cases to identify patterns and factors that influence treatment success (Hassaballah et al., 2019). This information can be used to develop predictive models that help carers to make informed decisions and improve long-term treatment planning.
4. Research studies on assessment of CL treatment outcomes bring numerous benefits such as provision of objective evaluation metrics, fostering research advancements and enabling early interventions and outcome predictions. Specifically, by leveraging computational techniques, carers and researchers are equipped with the potential to make significant strides towards optimisation of CL repair procedures, practices, and outcomes. This creates a hopeful window for understanding the relevant factors and improving the overall well-being of affected individuals (Leopoldo-Rodado et al., 2021).

1.8 Outline of Research Contributions

This research study has developed the understanding and application of computational techniques in evaluating the appearance of CL treatment outcomes. The key research contributions of this multidisciplinary PhD research study are outlined below:

1. Development of Novel Computational Assessment Framework

This study developed a novel computational assessment framework tailored for CL treatment. By integrating advanced image processing technologies, machine learning techniques, and computer-aided and mathematical models, a comprehensive framework has been established to evaluate the outcomes of CL surgical repair automatically and objectively. This framework incorporates both quantitative and qualitative assessment measures, providing a more accurate and holistic evaluation of treatment outcomes.

2. Application of Machine Learning and Deep Learning Techniques

One of the major contributions of this research is the innovative application of state-of-the-art machine learning and deep learning techniques in the analysis of CL treatment outcomes for appearance assessment. By training and deploying classical

models on the given datasets of post-operative images, the study has demonstrated the potential of machine learning algorithms to fully automate the assessment process and improve accuracy. Potential for more accurate results is vindicated for transfer learning approaches. Precisely, this research has developed a reliable appearance assessment pipeline for CL post-treatment partial facial visuals.

3. **Improvement in Cleft Lip Treatment Evaluation**

The computational assessment framework developed in this study has made significant improvements in the estimated appearance assessment of cleft lip treatment outcomes. Refinements in the research could yet add valuable insights into the factors influencing surgical success, such as lip symmetry, scar formation, and overall appearance outcomes. These findings, some published, and others currently under review, contribute to the development of evidence-based guidelines and will give a sound scientific basis to surgeons in choosing and optimising treatment strategies to achieve improved CL surgical treatment outcomes.

4. **Linking Traditional and Computational Approaches**

This study successfully bridged the gap between traditional manual assessment methods and computational CL treatment outcome evaluation techniques. A more comprehensive and objective assessment approach has been established by combining clinical expertise with computational tools. This integration can enhance the accuracy, efficiency, and consistency of CL treatment outcome assessment, enabling a more standardised and reliable evaluation.

5. **Contribution to Literature to aid Future Research Directions**

This research has identified several gaps in the existing literature and highlighted areas requiring further investigation. The study has shed light on the need for more robust and diverse datasets, advanced imaging technologies, and standardised evaluation metrics used in computational assessment of CL treatment. Based on these findings, future research directions could explore multi-modal data fusion techniques, integrating 3D imaging technologies. Additionally, there is need for addressing specific challenges related to scar prediction, soft tissue generalisation and long-term treatment outcomes for generic maxillofacial treatment outcome.

6. Design of Mathematical Models for Regression of SSIM into Scores

This study demonstrated the possibility of converting a quantifiable measure such as the known structural similarity index into a regression outcome as an appearance score. Therefore, other quantifiable attributes can be experimented using the developed models. Our scientific understanding of the attributes of partial facial appearance were modelled in six different equations (equations 7-9 and equations 12-14) with robust and incremental results. With such regression results produced instantly, planners for surgical procedures have the chance to improve their practices because they would evaluate surgical procedures outcomes objectively with minimal human involvement.

1.9 Thesis Structure

The rest of the thesis is structured as follows:

Chapter 2 presents a comprehensive literature review of CL treatment and potential assessment methods. It covers the following CL assessment methods: traditional (or manual or qualitative) methods, computational (semi-automatic) techniques, and evaluation of CL repair outcomes. By diving into existing research, gaps in knowledge and research opportunities are further identified.

Chapter 3 is the methodology section, describing the research design and different approaches used in the study. Dataset description, utilized pre-processing techniques before features extraction, are presented. A discussion of baseline machine learning algorithms, model frameworks used for CL assessment, and evaluation metrics is also presented.

Chapter 4 presents the Shape Analysis approach for assessing and evaluating CL treatment outcomes. The feature extraction techniques for the traditional computational method are discussed. Application of the mouth lips as the region of interest (RoI) is the pivot of this section. Using symmetry and similarity measures, CL treatment assessment is quantitatively computed using the designed mathematical models embedded in a computer program.

Chapter 5 discusses the hybrid computational approach used for CL treatment outcome assessment. In this chapter, the different features of the outcome appearances are detected per potential RoI in the CL appearance outcome to aid with

adaptive facial asymmetric/ symmetric detection. In this chapter, CL outcome assessment measures the degree of symmetry or asymmetry by using an adaptive method to determine the axis involved. Three different models are presented to quantify the CL surgical treatment outcome assessment.

Chapter 6 presents a deep learning computational approach that uses transfer learning techniques to create and evaluate a deep learning model. Extraction of stochastic features aids this approach in a semi-supervised manner for evaluating CL treatment outcome assessment.

Chapter 7 discusses an in-depth analysis, interpretation, and synthesis of research findings in chapters 4 to 6. By doing this, a comparison is made between the results of this work and existing research studies/literature. Clinical implications and practical applications of this research study are also discussed. Additionally, this chapter presents the Conclusion and Future Studies. The research objectives and key findings are summarised. Highlights of the research significance and contributions are given. Furthermore, this chapter hints on suggestions for future research directions while offering recommendations for enhancing computational assessment of CL treatment.

Chapter 2 Literature Review

2.1 Introduction

This section critically analyses existing research and knowledge relevant to this research study. The purpose of this chapter is to help contextualise the research problem, identify the research gaps, and rationalise the proposed research, as presented in the following subsections.

2.2 Epidemiology of Cleft Lip (CL)

Cleft lip (CL) is a congenital malformation that affects the facial structure of individuals. The cleft is a failed fusion of tissue plates in the embryo at around 4-6 weeks after conception. Consequently, there is an opening in the upper lip[5], as seen in Figure 2.1.

Figure 2. 1: Cleft lip in a subject. Normally, this is a cut in the upper lip whose fusion did not before birth.

It is one of the most common birth defects, with a prevalence ranging from 1 in 500 to 1 in 2,500 live births globally, depending on ratio identity and location (Mossey and Modell, 2012, Salari et al., 2022). The exact causes of CLP are not fully understood,

5 https://www.cdc.gov/ncbddd/birthdefects/cleftlip.html#:~:text=A%20cleft%20lip%20happens%20if,the%20lip%20into%20the%20nose.

but it is believed to be a complex interplay of genetic and environmental factors (Dixon et al., 2011). The CL condition can be classified as either syndromic or non-syndromic.

There is a growing body of research on the genetic basis of CL, with many genes identified as potential risk factors. Studies have shown that variations in genes involved in craniofacial development, such as IRF6, MSX1, and PVRL1, are strongly associated with an increased risk of CL (Khandelwal et al., 2013, Leslie and Marazita, 2013, Takechi et al., 2013). Other studies have also implicated environmental factors, such as maternal smoking and alcohol consumption, in the development of CL (Dixon et al., 2011).

Treatment of CL typically involves a multidisciplinary approach, with a team of healthcare professionals from various specialties working together to manage the condition. Surgical repair is the backbone of treatment for CL, intending to improve the function and appearance of the mouth (Raghavan et al., 2018). This enhances infants' ability to (breast) feed and interact with their mothers and/or carers. Other interventions, such as speech therapy and orthodontic treatment, may be necessary to address the associated complications of CL (Cai et al., 2018, Isiekwe and Aikins, 2019). The significant advancements in the surgical methods used to repair CL have improved outcomes. However, access to this specialised care remains significantly costly, therefore accessing such services remains a challenge in many parts of the world (Swanson et al., 2017, Bennett et al., 2018, Murthy, 2019)

2.2.1 Cleft Lip Demographics

The World Health Organization (WHO) reported that the prevalence at birth of orofacial cleft (OFC) varies worldwide, in the range of 3.4–22.9 per 10,000 births for cleft lip and palate (CLP) (Mossey and EE, 2003).

The prevalence of CLP and/or OFC has been found to vary based on ancestry, with the highest incidence rates observed amongst Asian (Tibeto-mongoloid) populations (0.82–4.04 per 1000 live births), intermediate rates amongst Caucasians (0.9–2.69 per 1000 live births), and the lowest rates amongst African populations (0.18–1.67 per 1000 live births) (Mossey and Modell, 2012, Salari et al., 2022, Wang et al., 2023). Despite this prevalence, it was discovered that the "burden of orofacial clefts falls disproportionately on the countries with the smallest surgical workforce or lowest

Socio-Demographic Index, rather than those with the highest prevalence of disease", (Massenburg et al., 2021).

However, several studies disagree on whether OFC/CLP is more prevalent among females or male populations(Ahmed, Bui and Taioli, 2017, Eshete et al., 2017, Swanson et al., 2017).

2.3 Methods and Challenges in Evaluation of CL Repair Outcomes

CL repair is a complex and challenging surgical procedure that requires understanding the underlying anatomy, tissue characteristics, and healing mechanisms (Massie et al., 2016, Bekele, Ekanem and Meberate, 2019). For this reason, sometimes pre-surgical planning using computer-aided software is desired and emphasised. Various objective and subjective measures can evaluate the procedure's success, including clinical assessment, functional outcomes, and patient-reported satisfaction (Klassen et al., 2021).

CL treatment assessment methods can be categorised into two broad categories, direct clinical and indirect clinical assessment. The different respective sub-categories are summarised in Figure 2.2.

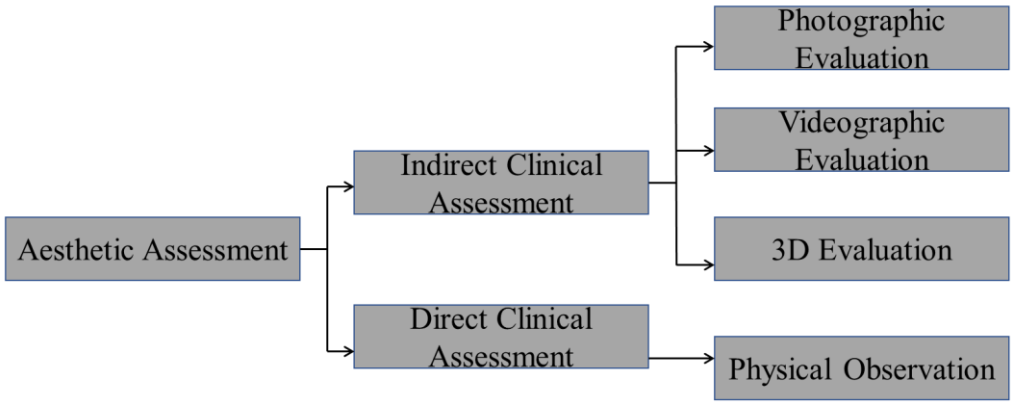

Figure 2. 2: Classification of CL Treatment Outcome Assessment Methods. There are 2 general approaches: indirect and direct approaches.

2.3.1 Direct Clinical Assessment (DCA)

This method involves the physical examination of the patient who underwent the CL surgical repair by an expert or assessor. Because a multidisciplinary care team is involved in the care and management of patients with CL condition, several experts

such as surgeons and orthodontists could be involved in the physical examination. DCA is also called a “live assessment” session (Sharma et al., 2012).

In this method, direct quantitative measurement generates metrics and dimensions of the different facial parts. This is categorised as an objective DCA tool (Sitzman and Allori, 2014).

DCA also involves visual inspection and palpation of the repaired lip, as the most used method for evaluating the appearance outcome of the surgery (Mcelroy et al., 2017). However, it is subjective and relies on the expertise and experience of the evaluator, leading to inter- and intra-observer variability (Nahai et al., 2005). Functional outcomes, such as speech and feeding, are also important measures of CL repair success under the DCA approach (Kummer, 2014).

The additional challenge is the potential conflict of interest should the assessment team include members of the CL repair team.

2.3.2 Indirect Clinical Assessment (ICA)

2.3.2.1 Overview

ICA utilises images of patients after CL treatment. The images are usually 2D facial images used as subjects for evaluation. Precisely, the region of interest (RoI) is the mouth (consisting of the upper and lower lips) (Al-Omari et al., 2003, Sharma et al., 2012).

Objective assessments, such as nasometry, electropalatography, and videofluoroscopy, have evaluated speech outcomes in CL patients (Kuehn and Henne, 2003, Nahai et al., 2005, Kummer, 2014). Similarly, feeding outcomes can be evaluated using objective measures such as the infant feeding assessment yardsticks (Gopinath and Muda, 2005).

Patient-reported outcome measures (PROMs), such as the cleft evaluation profile (CEP) and the Cleft Q system[6], have been developed to assess patient satisfaction and quality of life after CL repair (Mulder et al., 2019). PROMs provide a more patient-centred evaluation approach and are increasingly used in clinical practice.

[6] https://qportfolio.org/cleft-q/

Despite these evaluation methods, there are persistent challenges in assessing CL repair outcomes. The lack of standardised evaluation protocols to assess appearance of CL treatment outcomes makes it difficult to compare research results across different studies (Mosmuller, Mennes, et al., 2017a) . In addition, the appearance assessment of outcomes in patients with CL can be more challenging due to the additional complexity of the surgical procedures and associated comorbidities (Mosmuller, Mennes, et al., 2017a, Mulder et al., 2019).

2.3.2.2 Photographic Assessment

Most prominent ICA methods utilise digital imagery analysis or photographic evaluation techniques. This can be conducted qualitatively or quantitatively. Figure 2.3 summarises the breakdown.

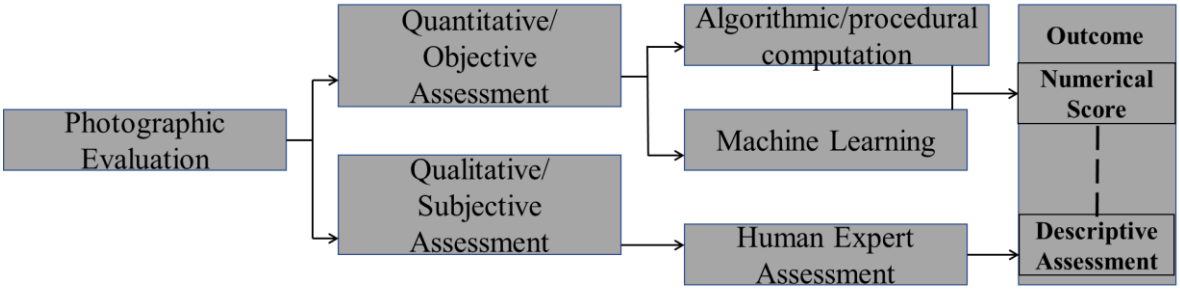

Figure 2. 3: Photographic Evaluation Methods. Divided into qualitative and quantitative methods.

Qualitative assessment requires human experience and emotion to judge the CL repair outcome. Assessment is normally descriptive but could be quantitatively coded.

Quantitative evaluation uses a wide range of image features for assessment. The nature of the assessment task presents different feature needs, either fewer features or infinitely many and potentially advanced features. Consequently, assessing the different appearances following CL treatment using traditional computational algorithms and machine learning approaches is possible.

The following research studies were identified as closely related to this PhD research. These approaches quantitatively assess CL treatment outcome using ICA digital imagery.

1. Analyse It Doc (A.I.D)

Analyse It Doc (A.I.D.) is an anthropometric analysis software for images of the facial nasolabial region. The software evaluates facial morphology, symmetry and appearances using standardised photos from Adobe Photoshop (Adobe Inc, San Jose, California, USA) (Adobe Inc, 2021) or ImageJ software (National Institute of Health, Stapleton, New York, USA) (Schneider, Rasband and Eliceiri, 2012). The photos must be standardised in different views such as frontal, lateral, and submental views (Pietruski, Majak and Antoszewski, 2017).

A.I.D has modules for subjective and objective assessment/evaluation of appearance outcomes. Therefore, this approach is semi-automatic. Additionally, the objective assessment module using the A.I.D uses only a few landmarks. The system has got two major applications: examination of individual patients and analysis of voluminous multiple evaluators from multiple treatment centres. Although the system can be used collaboratively, few validation studies have been reported using this approach.

Therefore, CL repair treatment outcome assessment is a secondary consideration for the A.I.D software. Besides, A.I.D's manual photogrammetric inputs indicates its semi-automatic approach (Pietruski, Majak, Debski, et al., 2017). Further, the features are manually marked on the frontal and lateral views. However, it outputs strong inter/intra-rater reliability. This is attributed to the multiple views of measuring several parameters from which the raters' reliability is computed. For example, there is significantly better accuracy of repeated linear and angular measurements in submental view, frontal view, and lateral view.

2. SymNose

SymNose is a computer-based program developed to aid CLP research (Pigott and Pigott, 2010) and was developed to aid quantitative outcome analysis for 2D digitised images. The UK standard for assessing the cleft lip repair is from colour 2D Anterior posterior, worms' eye and lateral images taken under standard conditions when the child is 5 years old. The system creates semi-objective parameters such as lip and facial symmetry.

Different studies have used SymNose to generate objective outcomes, specifically absolute scores through rating images from best to worst (Freeman, Mercer and

Roberts, 2013, McKearney, Williams and Mercer, 2013, Russell, Kiddy and Mercer, 2014, Mosmuller et al., 2016, Bella et al., 2016, Deall et al.,2016). An upgraded version of SymNose, SymNose 2 (Pigott and Pigott, 2016), has improved capabilities of scar quantification and a thin lip correction. The lip-aspect ratio (LAR) algorithm was responsible for this correction. The LAR algorithm was applied to a study where the appearance range of normal symmetry for facial features was also defined (Kornmann et al., 2019).

SymNose helps annotate outcomes for identifying regions of interest and the resulting symmetry for assessment preparation studies. This semi-automatic approach has been widely used in several studies and requires preparation and comparison of several parameters, especially in bilateral CL and palate cases.

3. Cleft Lip and Palate Network (CLPNet)

CLPNet utilises artificial intelligence (AI) and machine learning (ML) techniques to provide recommendations, insights, and decision support for healthcare professionals involved in CL treatment. This approach presents a preoperative advisory platform that uses deep learning algorithms to output high and accurate predictions for surgical markers and incisions to ensure an excellent appearance outcome following CLP treatment (Li, Cheng, et al., 2019). The study is purely pre-surgery advisory with no outcome assessment performed.

Table 2.1 summarises the different studies reviewed, indicating methods, datasets, and other valuable considerations.

Table 2. 1: Comparison of Related Studies

Research Study	Criteria								
	Image Acquisition Method	Image Nature (2D/3D)	Age of Subjects (in years)	Image Orientation	Input pre-processing technique	Maxillofacial Defect	(Implementation) Platform	Practitioner/ Researcher	Assessment/ Evaluation protocol
Analyse It Doc	High Resolution camera	2D	Universal	Frontal, Lateral, Submental	Standardized	Unilateral CLP	(Not stated) At least Windows 7, Mac	Practitioners and Researchers	Objective/ Subjective
SymNose	Scanner, Camera, Slide projector	2D	10	Frontal, Basal	Mouse, Digitizing pad	Unilateral CLP Complete BCLP	At least Mac 10.4	Practitioners and Researchers	Control group, Objective,
CLPNet	Face detector	3D	Flexible	Flexible	Professional labelling, centre crop	Complete CL	(Not stated)	Practitioners and Researchers	Control group

2.4 Computational Methods for CL Treatment Outcome Assessment

Over the years, there has been a growing interest in using computational techniques for cleft lip repair assessment. These techniques have been used to evaluate the outcomes of CL repair procedures and improve the precision of treatment planning. A range of computational methods, including 2D imaging, 3D modelling, machine learning, and computer-aided simulations, have been explored in the literature.

One common method is 2D imaging, which involves capturing photographs of the patient's face and analysing them to determine the severity of the cleft and the extent of the repair required. Many qualitative approaches as discussed above require capture of the facial images. For instance, (Gong and Yu, 2012) used a 2D imaging-based technique to assess the outcomes of CL repair in a group of infant patients. They found that the method provided accurate and reliable results and could be used to monitor the progress of the repair over time.

Another approach is 3D modelling, which enables the creation of a virtual model of the patient's face that can be manipulated and analysed in detail. This method has been used to investigate the effects of different surgical techniques on the outcomes of CL repair. For example, (Riedle et al., 2019) used 3D modelling to compare the outcomes of two different surgical techniques for CL repair. They found that one technique produced better results than the other and suggested that it should be used more widely.

Machine learning has also been used to analyse large datasets of CL repair outcomes. This method involves training algorithms to recognise patterns and make predictions based on the data. For instance (Chada, n.d., Riedle et al., 2019, Haque et al., 2021) proposed computer-based models and machine learning to identify factors influencing CL repair procedures' outcomes. They found that the severity of the cleft, the patient's age, and the surgical technique used were important factors that affected the outcomes.

Finally, computer-aided simulations have been used to simulate the effects of different surgical techniques on the outcomes of CL repair. This method involves creating a virtual model of the patient's face emphasising the affected region and manipulating it to simulate different surgical scenarios. For example (Gong and Yu, 2012, Riedle et al., 2019) used computer-aided simulations to investigate the effects of different

surgical techniques on the symmetry of the repaired lip. They found that the simulations provided a useful tool for predicting the outcomes of different surgical approaches.

Given a dataset of images following CL treatment, it is natural to determine and compute their appearance scores using contemporary computational approaches on advanced computing resources. The general discussion of the advanced approaches is given below.

2.4.1 Machine Learning

Machine learning is a subdomain of artificial intelligence (AI) whose focal point is on developing algorithms and models to enable computers to learn from and make predictions or decisions based on given datasets. It is a fast-evolving field that has found relevance in various domains, from healthcare and finance to self-driving cars and natural language processing (Kreuzberger, Kuhl and Hirschl, 2023).

The term ‘machine learning’ was coined by Arthur Samuel in 1959 (Samuel, 1959). Machine Learning is defined as the study of computer algorithms that improve and evolve in knowledge, automatically through experience (Alzubi, Nayyar and Kumar, 2018).

There are two commonly applied categories of Machine Learning (Alloghani et al., 2020):

1. Supervised Learning: In supervised learning, algorithms learn from labelled datasets, by making predictions or classifications.
2. Unsupervised Learning: Unsupervised learning deals with unlabelled datasets and requires clustering or dimensionality reduction.

There are several algorithms’ examples and techniques aligned with the above-mentioned machine learning categories (Rajoub, 2020):

1. Linear Regression: Used for predicting continuous values.
2. Decision Trees: Used for classification and regression tasks.
3. K-Nearest Neighbour: Is a non-parametric, supervised learning classifier, which uses closeness to make classifications or predictions about the clustering of an individual data point.

4. Deep learning: Is a subset of machine learning which involves artificial neural networks with many layers, and it has proven highly successful in many computer vision tasks like image recognition, image interpretation and classification.

Despite the huge potential, machine learning faces some challenges and considerations as discussed below (Paleyes, Urma and Lawrence, 2022):

1. Bias and Fairness: Machine learning models can perpetuate biases present in training data, leading to unfair and / or discriminatory results.
2. Ethical and Privacy: Issues arise regarding collecting and using personal data for machine learning. Normally ethical approvals take long and if approved, it is for only a shorter period.
3. Transparency: The 'black box' nature of some models raises concerns about accountability.
4. Interpretability: Complex machine learning models can be challenging to interpret and explain, which is critical for trust and accountability.
5. Data Quality: High-quality, diverse, and representative data is crucial for training effective models.

Although there are some challenges, machine learning continues to advance, with ongoing research and future developments in academia and industry, driving innovations that shape the global technological landscape. Some future directions include (Gaur et al., 2019, Mohammad-Rahimi et al., 2021):

1. Explainable AI (XAI): Developing models that provide understandable explanations for their decisions.
2. Federated Learning: Training models on decentralized data to preserve privacy.
3. Quantum Machine Learning: Exploring the potential of quantum computing to enhance machine learning algorithms.
4. AI Ethics: Continued focus on ethical considerations, fairness, and responsible AI development.

2.4.2 Deep Learning

Deep learning is a subfield of machine learning and artificial intelligence (AI) that focuses on the development and training of deep neural networks (DNNs) (Goodfellow, Bengio and Courville, 2016). DNNs are complex and hierarchical

computational models inspired by the structure and function of the human brain (Lecun, Bengio and Hinton, 2015). Deep learning has gained significant attention and popularity due to its remarkable success in solving complex tasks, particularly in areas such as computer vision, natural language processing, and speech recognition (Havaei et al., 2017, Samek et al., 2021).

As stated before, artificial neural networks (ANNs) are at the core of deep learning, which consist of interconnected nodes or neurons organized in layers. These networks can have many layers, giving rise to the term "deep" learning. DNNs are designed to automatically learn hierarchical representations of data, with each layer learning increasingly abstract features (Miikkulainen et al., 2018).

Some of the commonly used deep learning architectures are:

1. Feedforward Neural Networks (FNNs): These are the simplest form of DNNs, with information flowing from input to output layers in one direction.
2. Convolutional Neural Networks (CNNs): Specialised for processing grid-like datasets, such as images and videos. They use convolutional layers to learn spatial features.
3. Recurrent Neural Networks (RNNs): Designed for sequential data, like time series or natural language. They maintain hidden states to capture temporal dependencies.
4. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU): Architectures within RNNs that address the vanishing gradient problem and are capable of learning long-term dependencies.

Some of the common deep learning applications are discussed in (Samek et al., 2021) and summarised below:

1. Computer Vision: CNNs are widely used for tasks like image classification, object detection, and image generation (e.g., GANs).
2. Natural Language Processing (NLP): Transformers, a type of deep learning architecture, have revolutionized NLP, leading to advancements in machine translation, chatbots, and sentiment analysis.
3. Speech Recognition: Deep learning models have achieved human-level performance in speech recognition tasks, enabling voice assistants like Siri and Alexa.

4. Autonomous Vehicles: Deep learning is used for perception and decision-making in self-driving cars.

Deep learning has a few notable considerations and challenges, in addition to the those discussed for general machine learning (Razzak, Naz and Zaib, 2018, Galván and Mooney, 2021).

1. Training: Training deep neural networks typically involves backpropagation, where the model's parameters are adjusted using gradient descent or its variants to minimise a loss function. Therefore, large-scale labelled datasets and powerful hardware, such as GPUs and TPUs, have effectively trained deep learning models. Further, cloud computing services have been instrumental in helping researchers access superior computing resources.
2. Overfitting: DNNs are prone to overfitting, where the model performs well on the training data but poorly on unseen data.
3. Vanishing and Exploding Gradients: These issues can hinder the training of DNNs with many layers.
4. Data Requirements: Deep learning models often require vast amounts of labelled data, which may not always be readily available.

Ever since deep learning started demonstrating groundbreaking results in various domains, it has continued to be a driving force in AI research and application development. Some of the ongoing advancements, breakthroughs, and future directions are discussed in (Galván and Mooney, 2021):

1. Explainable Deep Learning: Developing methods to make deep learning models more interpretable and transparent.
2. Continual Learning: Enabling deep learning models to adapt to latest information over time.
3. Quantum Deep Learning: Exploring the intersection of deep learning and quantum computing.

2.5 Summary of Knowledge Gaps

The above-described CL treatment outcome computational assessment measures have got some deficiencies as summarised below:

1. **Lack of standardisation**

There is currently no standardised protocol for using computational techniques in CL repair assessment. This makes it difficult to compare results across different studies and hinders the development of a robust evidence base (Kassam et al., 2020, Alighieri et al., 2021).

2. **Limited validation**

Many computational techniques used in appearance assessment of CL repair outcomes have not been adequately validated. This makes it difficult to determine their accuracy and reliability and raises concerns about the validity of the results obtained. This is partly because “despite advancements, there is variable consensus on technique, timing, and sequence of clefts…related repair procedures” (Naidu et al., 2022), hence affecting treatment outcome assessment validation efforts.

3. **Limited application**

Many computational techniques used in CL repair outcome assessment are not widely available and some are too complex for routine use by the healthcare teams (Frank-Ito et al., 2019) and researchers. This limits their potential impact on clinical practice, patient treatment outcomes and improvement by other researchers.

4. **Limited generalisability**

Many studies have focused on specific populations, such as infants or patients with unilateral CL, complete CL, or otherwise. This limits the generalisability of the findings and raises questions about their applicability to other populations.

5. **Heterogeneity of CL Research Network Teams**

Multidisciplinary research teams are sometimes hindered by communication deficiencies of key ideas for specialised domain concepts. This hinders development and validation of tools and techniques. For example, the significant difference in the assessment outcomes of different experts for the same subjects is one of the indicators of how challenging different teams can harmonise assessment tasks for treatment outcomes, (Patcas et al., 2019).

6. Limited integration

There is a need for greater integration of computational techniques into routine clinical practice (Huqh et al., 2022). This will require changes in clinical workflows, training of clinicians, and investment in infrastructure. For example, an online crowdsourcing approach compared to speech language pathologists can’t easily be adapted (Sescleifer et al., 2020).

To address the identified knowledge gaps, there was need to conduct further research.

Chapter 3 Methodology

3.1 Introduction

This section presents the research approach used in this study. Different researchers can differently define Research Methodology. “Research methodology is a system of models, procedures and techniques used to find the results of a research problem”, (PANNEERSELVAM, 2014). Another definition is that “Research methodology is a way to systematically solve the research problem … has many dimensions and research methods constitute a part of the research methodology”, (Kothari 2004:8). The two definitions are complementary and applicable to this research study. Visual datasets are analysed by extracting distinctive features to aid interpretation and better assessment of outcomes following cleft lip treatment. For purposes of advancing knowledge and understanding, a range of techniques, strategies, and procedures (research methods) are employed and investigated.

3.2 Experimental Research Design

Research design is a critical component of research methodology. Research design refers to the blueprint or plan that outlines how a study has been conducted. It involves decisions about the type of research, data collection methods, and data analysis techniques (Creswell, 2003, Seltmann, 2014).

Evident from the dataset is the availability of facial image outcomes following surgery. From these images, computations can lead to determine symmetry, proportions/ shapes, and anthropometric measurements. Human experts qualitatively assessed facial image outcomes. However, the semi-structured qualitative assessment was converted to a numeric Likert scale where 1 is ‘Excellent’, 2 is ‘Good’, 3 is ‘Fair’, 4 is ‘Poor’ and 5 is ‘Very Poor’.

Subsequent sections in this chapter present details of the computational analysis methods through the development and evaluation of different computational algorithms, for appearance analyses of the post-surgical facial dataset. The developed algorithms involve regression analysis, symmetry assessment, landmark detection, anthropometric measurements, and proportion/shape analysis and deep learning.

A validation approach is through correlation analyses between quantitative assessment outcomes performed automatically and those by human assessors. However, the deep learning regression analysis approach explores several validation metrics.

Applying the developed computational algorithms to non-cleft lip partial facial images would give the research study a solid foundation. However, no satisfactory data set of facial visuals of corresponding non-cleft 5-year-olds with ethical permission was available to us.

3.3 Image Processing and Analysis Pipeline

The collected images were processed and analysed using various traditional and advanced image manipulation techniques. The image processing techniques used in the study can be classified as image pre-processing, segmentation, feature extraction, and classification, among others. To maximise outcomes of advanced image manipulation operations, pre-processing is necessary (Dharavath, Talukdar and Laskar, 2014). Blurring, morphing, pixel equalisation, standardisation, luminosity adjustments, among others are key towards reducing the noise on images (Miljkovi, 2009). Figure 3.1 depicts the detailed facial analysis flow, from input to output.

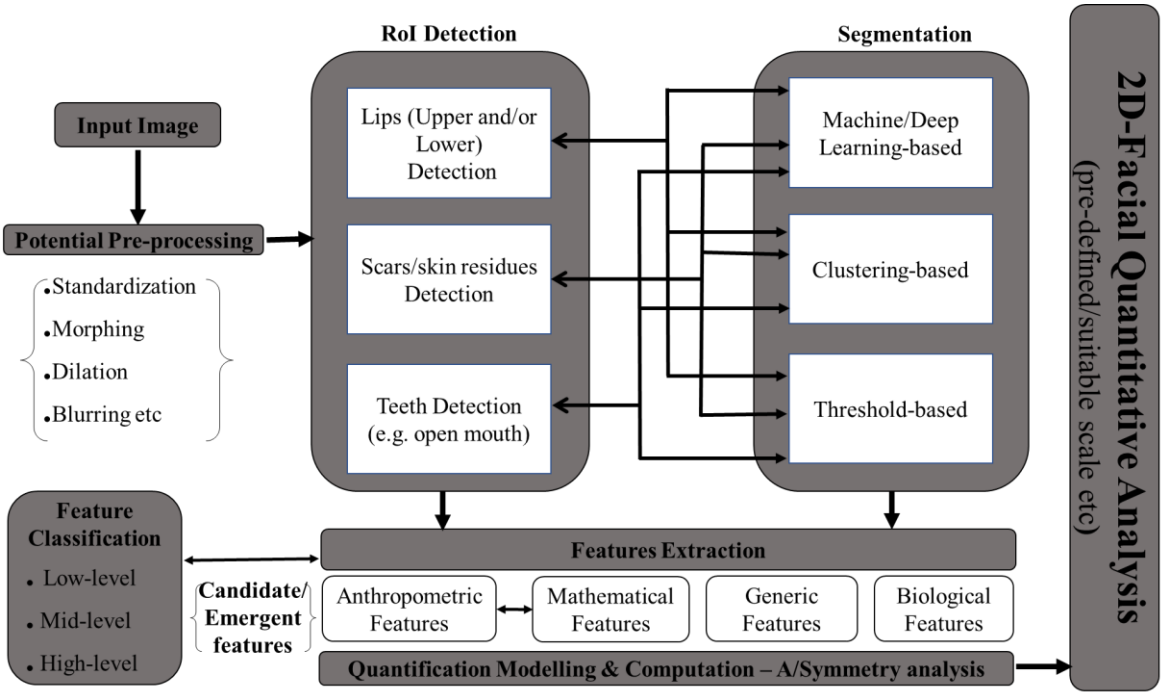

Figure 3. 1: Representation of a fusion of different blocks responsible for detailed quantitative facial analysis. Preprocessing, RoI detection, Segmentation, Features extraction and Modelling.

Basically, Figure 3.1 indicates that subsequent chapters will present image pre-processing techniques to aid with RoI detection through segmentation, before performing any feature extraction tasks.

3.3.1 Brief Image Pre-processing

This is a technique for formatting the input images before analysis or inference or model creation. The main reasons for pre-processing are quality enhancement, noise reduction/removal, interpretability improvement, reduction of computational complexity, and preparation for further processing or analysis. Therefore, pre-processing is an essential step in computer vision and image analysis research (Szeliski, 2011). A detailed explanation of pre-processing operations considered has been discussed in (Gonzalez and Woods, 2002, Szeliski, 2011, Dharavath, Talukdar and Laskar, 2014, Nixon and Aguado, 2019).

In this research, the following preprocessing operations were applied as detailed in the subsequent chapters: Resizing and Scaling, Gray Scaling, Denoising, Contrast Enhancement, Image Normalisation.

However, the image cropping/slicing technique has been used across the different chapters and is briefly discussed below.

Cropping involves selecting a specific region of interest within an image while discarding the remaining parts. It is useful for removing unwanted backgrounds, focusing on specific objects, or resizing images to a standard aspect ratio. Cropping can be performed manually or using automated techniques based on object detection or segmentation algorithms. One of the approaches used for CL treatment assessment, crops the facial images into three thirds. A common implementation technique is the array slicing. Some results are presented in Figure 3.2.

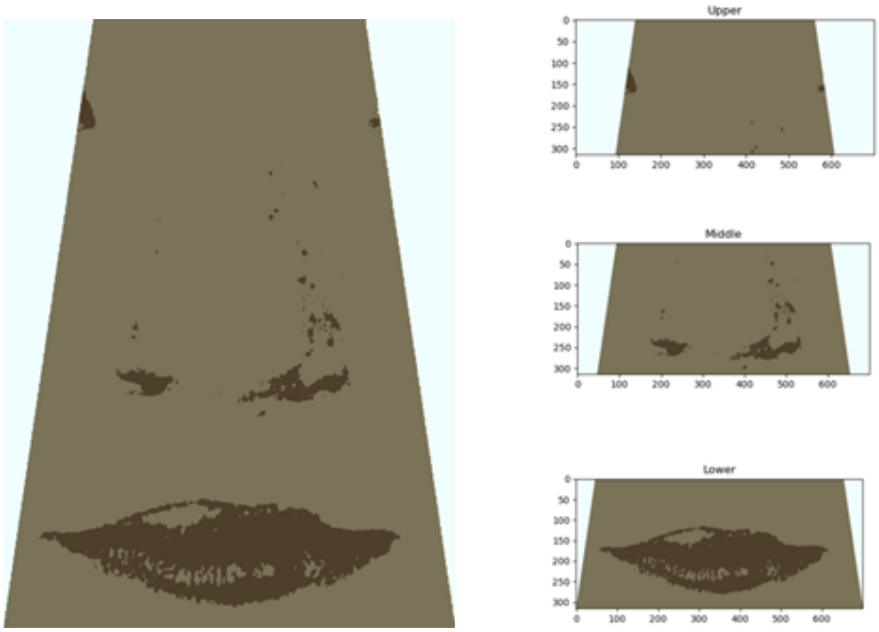

Figure 3. 2: Image (left) sliced into three segments i.e., upper, middle, and lower segments (right)

3.3.2 Segmentation Overview

Image segmentation is a fundamental task in computer vision that involves dividing an image into meaningful and coherent regions or segments. In simpler terms, segmentation helps users identify the object (wanted/meaningful section) and background in an image. It plays a crucial role in various applications, such as object recognition, image understanding, medical imaging, and autonomous driving (Wang, Wang and Zhu, 2020, Wang et al., 2022). This study innovatively applied/adapted some segmentation approaches to develop and validate computational methods for appearance assessment of CL treatment outcomes. A discussion of segmentation algorithms used is presented in the respective subsequent chapters. The general approaches are Thresholding-Based Segmentation, Edge-Based Segmentation, Clustering-Based Segmentation and Deep Learning Segmentation Methods.

3.3.3 Features Extraction and Detection

After segmentation, unique features in the region of interest (RoI) can be identified. There are four broad categories of features:

1. **Anthropometric Features**

These would naturally involve direct and physical measurements of the different facial feature dimensions (such as thickness, length) where possible (Sunderland, 1995). This facilitates analysis of facial anatomical features and assess the recovery extent, especially for the upper lip. CL affects the upper lip and is usually the part of the mouth region to undergo surgical repair. The regrowth of the development of that region is closely monitored if treatment is to be assessed.

Specialists use specialised tools such as callipers or anthropometers to capture anthropometric data (Norton, 2019). Anthropometry can violate a patient's privacy, may lack standardisation and can be tedious, hence should be minimised or completely avoided (Wang et al., 2000). Segmentation not only automates this process but also ensures objectivity. The different features are stored as continuous numeric values, in a mathematical formulation. Consequently, we obtain quantifiable record of measurable facial attributes.

2. **Geometric Features**

Following successful segmentation, mathematical features can be generated. These types of features refer to quantifiable measurements and attributes of the region of interest such as the human face.

Normally, these attributes can be represented, analysed, and interpreted using mathematical equations (or models). In this study, the following mathematical features of the face have been studied:

a. **Symmetry**

   This attribute is foundational to many facial appearances studies including assessment of surgical treatment (Passalis et al., 2011). Analysis of some features of the mouth region leads to landmark-based symmetry. This technique is employed to measure and quantitatively determine CL treatment assessment using models.

b. **Rational Facial Composition**

Ratios and proportions can be used to illustrate the relative sizes and relationships between different facial attributes. The golden ratio can be used to compare the considerably acceptable distance between facial features such as the eyes, mouth corners or between the eyes, nose, and mouth regions (Prokopakis et al., 2013) or the general face (Hashim et al., 2017).

c. **Facial Landmarks**

These are categorised as some specific points on the face, belonging to known regions such as the eyes, nose, or mouth (Wang et al., 2018). For example, eye corners, nose-base, nose tip, mouth corners, and philtral ridges, respectively are potential facial landmarks. Automatic detection of these features is a backbone of deep learning to aid geometric and mathematical algorithms to assess spatial relationships, hence assessment of a potential treatment remedy.

3. Biological Features

Biological factors such as genetics, anatomy and physiological processes determine the physical characteristics and structures of the human face (Jagadish Chandra et al., 2012). Biological features are intrinsic to individuals while contributing to their unique facial appearance in addition to playing a crucial role in treatment outcome assessment and comparison (Ritz-Timme et al., 2011). For example, eye corners and nose shape, mouth/lip structures are unique among individuals. Understanding the structure, growth, and redevelopment dynamics of these features after a surgical procedure like in CL treatment case, can lead to design, development, and validation of robust frameworks.

Biological features are easy to differentiate between them if using physical measurements is applicable as a progressive characteristic (Bonidia et al., 2021). If the goal is to assess beauty, attributes such as thick lips, smaller eyes, long pointed nose would be ideal (Rennels and Cummings, 2013, Kar et al., 2018). On the contrary, the CL condition recovery process following treatment differs among individuals. Therefore, the assessment framework for the outcome should be robust to take into consideration the different biological features.

4. Generic Features

Generic facial features refer to the common and typical characteristics that are shared by most elements in the dataset. Whereas these features are not particular to categories of the dataset classes, they are representative of the classical facial traits (Vo and Le, 2016).

Therefore, features such as eye shape, nose shape, nose size, mouth/lips thickness, facial symmetry, and skin tone vary significantly across the different elements of our dataset. Because facial features are influenced by a combination of genetic, environmental, and cultural factors, it is instrumental that generic features are incorporated for scalability and robustness during model design.

Subsequently, generic features can be dynamically considered as anthropometric, mathematical, or biological by adaptation (Gu et al., 2017). For that reason, clinical assessment standards for successful CL surgical repair can be varied and harder to model.

The region of interest (RoI) is characterised by at least one of the feature categories described above. Once the appropriate features have been detected, the next step is to transform the features parametric properties into quantifiable result that form an assessment of a CL repair outcome.

3.3.4 Quantitative Modelling

Scientifically, quantitative modelling refers to the approach employed towards building mathematical or computational models used for description and analysis of (complex) systems (Jonkers and Franken, 1996). Normally, the models involve the use of numerical and statistical techniques to represent the relationships, interactions, and behaviour of variables within a system. Quantitative models can significantly predict phenomena in the hard sciences such as in physics, biology, engineering as well as in social sciences and economics (Series and Sterman, 2003, Sterman, 2006, Berhe and Makinde, 2020). In prediction and estimation of the success of the CL repair, a quantitative assessment model is designed based on the features of the repaired visual outcome.

The following steps constitute the process for quantitative model construction and have been applied to modelling of biological complex systems (Brodland, 2015):

1. **Problem Definition**

Define the problem that the model aims to address by clearly understanding the objectives, scope, and context of the modelling task. Furthermore, it is critical to identify the key variables, assumptions, and constraints involved.

2. **Conceptual Modelling**

Representation of the relationship and interactions between the identified variables follows. This is normally done using visuals such as flowcharts, sequence diagrams or pseudo codes. The goal is to realise any emergent interactions and challenges.

3. **Data Description**

The different variables should be explicitly described and defined to eliminate any ambiguities. Datasets arrangements and partition should be defined at this stage. Data for building and testing the model should be clearly defined. Where possible, gaining insight into the dataset patterns using statistical approaches should be conducted.

4. **Formulation**

The available dataset and extracted features influence mathematical equations or simulation techniques or machine learning approaches towards formulation of model equations or algorithms. Equations formulation and machine learning are preferred approaches when visual datasets are involved, like in CL treatment outcome assessment using partial facial images.

5. **Implementation and Validation**

Finally, the model is implemented using appropriate software tools or programming languages through translation of the model equations or algorithms into executable code. It is conventional to ensure that the model is efficient, robust, and user-friendly. After successfully implementing the model, evaluation of the model performance using test datasets follows.

3.4 Dataset Description and Ethical Considerations

The Cleft Care UK (CCUK) dataset used in this research study consists of anonymised facial images of 5-year-old children following CL treatment. Anonymity of the facial

images not only maintains children's privacy but also ensures objectivity of the human assessors who would potentially be destructed by other facial features.

This dataset was specifically curated for CL treatment outcome assessment studies. This dataset has been used by human assessors and for the development of alternative computational algorithms for further facial analysis and CL treatment objective assessment. The British Dental School at the University of Bristol undertook a national cross-sectional survey of children with a cleft lip and palate in 2013, called Cleft Care UK (CCUK). The intention was to assess the impact of reconfigured cleft care in the United Kingdom 15 years after an initial survey the Clinical Standards Advisory Group (CSAG) report in 1998, had informed government recommendations on centralising care. CCUK was a research study to repeat the previous audit of outcomes and assess if the re-organisation had improved the quality of cleft care. Figure 3.3 shows few dataset samples.

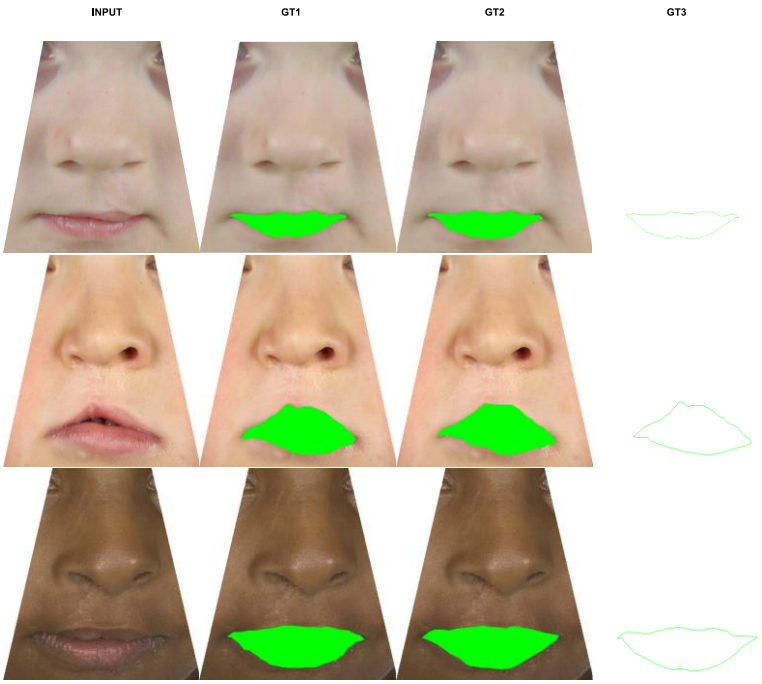

Figure 3. 3: Sample dataset images. Input, first column and respective ground truth images (GT1, GT3 and GT3) generated by three experts in subsequent columns.

The survey involved 5-year-old children with non-syndromic unilateral cleft lip and palate (Persson et al., 2015). The CCUK was a multi-centre cross section study that identified 359 eligible 5-year-old children from different cleft care audit clinics but only recruited 268 in the survey. Eventually, 250 facial images were captured due to some

images failing to meet the minimum inclusion rules (Sell et al., 2001). Clinical measures of speech, recordings, and photographs, among others were acquired by the national research team with main outcome measures being facial appearance, facial growth, speech, and wellbeing outcomes (R Al-Ghatam et al. 2015, Ness et al. 2015).

Permission to use the 2D images of the CCUK dataset was granted by the British Dental School at the University of Bristol. A team of caregivers and clinicians has utilised a different dataset of post operative CL repair outcome images of 5 year olds for qualitative assessment purposes (Bella et al., 2016). Consequently, a benchmark dataset for evaluation of computation algorithms for CL treatment assessment was generated.

In the subsequent chapters, the CCUK dataset will be applied and analysed alongside three ground truth datasets: GT1, GT2, GT3 generated by different human experts. In the context of computational modelling and deep learning, 'ground truth' refers to the accurate and reliable labels or annotations assigned to the training data used to train and evaluate a model (Richter et al., 2016). The model could be an ordinary computational or deep learning model. Ground truth usually serves as a reference or benchmark against which the model's predictions or outputs are compared (Kondermann, 2013).

3.4.1 Considerations of Ground Truth

1. **Introduction to Ground Truth**

'Ground truth' is a term commonly used in computer science domains such as machine learning, computer vision, and data science, among others. It refers to the authoritative and reliable source of data or information that is used as a reference or benchmark for evaluating the performance of algorithms, models, or systems (Martín-Morató and Mesaros, 2021).

Ground truth data serves as the reference or gold standard against which the results of experiments, measurements, or predictions are compared (Cardoso et al., 2014). It represents the true and accurate values or labels for a given problem.

Ground truth data has been used for different experiments in the following ways in this study:

1. Supervised Machine Learning: In supervised learning tasks, such as classification or regression, ground truth data consists of correctly labelled examples that the model tries to learn from. During model training and evaluation, the model's predictions are compared to the ground truth to assess its accuracy (See Chapter 6).
2. Computer Vision: Ground truth can include manually annotated object boundaries, object labels, or pixel-wise segmentations in images. Computer vision algorithms are evaluated based on how well they match the ground truth (See chapter 5).
3. Data Collection: Ground truth data is often collected through manual annotation or by experts in the field who have authoritative knowledge about the subject matter. This data is meticulously validated to ensure its accuracy.
4. Evaluation: Ground truth is essential for evaluating the performance of models, algorithms, or systems. Metrics like accuracy, precision, recall, F1 score, or mean squared error are computed by comparing the model's predictions to the ground truth (See chapters 4 to 6).

Other applications of ground truth data are in natural language processing (NLP) and geospatial analysis.

Obtaining high-quality ground truth data can be expensive and time-consuming. Human annotation can introduce inter-rater and intra-rater variability challenges. These challenges can significantly influence the quality and reliability of ground truth data, which, in turn, impacts the performance and validity of models, algorithms, or systems that rely on this data (Kondermann, 2013).

Subjective biases, and discrepancies can arise when multiple annotators are involved (inter-rater issues). However, there are also cases where the differences in judgement or annotations is made by the same rater (intra-rater variability). There is need to observe if a single annotator's judgments are inconsistent over time. Ensuring the reliability of ground truth is critical. Therefore, ground truth should be continuously updated or reviewed over time to account for changing conditions or to accommodate new datasets (Kanclerz et al., 2022).

To mitigate inter-rater and intra-rater issues affecting ground truth, it is fundamental to establish rigorous annotation protocols, provide clear guidelines, and conduct quality control checks.

The generated ground truth is both visual and numerical. Human experts qualitatively and semi-quantitatively rated visuals based on a scale. This was converted to numerical scores in the range of 1 to 5. This is critical for the evaluation of algorithms presented in chapters 4 to 6. Despite there being 250 images, experts' evaluation was more consistent with about 25 images. It is these images that make a great part of the ground truth used for evaluation.

2. **Iterative Enhancement**

Ground truth can be refined and improved over time. As more knowledge is gained, experts may revisit and update the existing labels to enhance the accuracy and quality of the ground truth dataset. This iterative process helps in continuously improving the performance of the computer vision and deep learning models (Cardoso et al., 2014).

3. **Influence on Building Models**

The quality and reliability of ground truth have a direct impact on the performance of computational models and deep learning models. Inaccurate or inconsistent ground truth can lead to biased models or misleading results. Therefore, careful attention should be given to the creation and maintenance of high-quality ground truth datasets. Likewise, the level of domain knowledge of the experts impacts the quality of the ground truth (Barr et al., 2020).

4. **Training and Supervised Learning**

In supervised learning, the deep learning model is trained using a dataset in which each input sample is associated with a corresponding ground truth label. These labels are manually assigned by human experts or obtained from reliable sources. In this study, the boundaries of the mouth and lip region are significant, among other features.

5. **Accuracy and Performance Evaluation**

Ground truth is crucial for evaluating the accuracy and performance of a trained machine/ deep learning model. By comparing the model's predictions to the ground

truth labels, metrics such as accuracy, precision, recall, and F1 score can be computed to assess the model's effectiveness (Li et al., 2018).

6. Prevalence of Annotation Methods

Ground truth labels can be obtained through various annotation tools and methods, depending on the nature of the task and available resources. Manual annotation involves human experts meticulously labelling the data, which can be time-consuming and costly. Alternatively, semi-automatic or automated annotation methods, such as crowdsourcing or using pre-existing labelled datasets, can be employed to accelerate the labelling process (Foncubierta-Rodríguez and Müller, 2012).

7. Labelling Challenges and Other Considerations

Creating accurate ground truth labels may face certain challenges. Ambiguity in the data, inter-annotator variability, or subjective interpretations can introduce discrepancies in ground truth annotations. To mitigate such challenges, careful annotation guidelines, consensus among annotators, and quality control measures are often implemented (Kondermann, 2013, Arhin et al., 2021).

3.5 Requirements of the Intervention

The solution required for the evaluation and assessment of cleft lip repair aims to minimise human intervention. Human evaluation is peppered with prominent limitations such as fatigue, subconscious bias, lack of reproducibility, slow evaluation process hence taking longer per evaluation. Additionally, human visual interpretation is unique and can be a source of nonuniformity between specialists (Palmer, Schloss and Sammartino, 2013, Bennett et al., 2019). This research handles hundreds of images. Much as the dataset is deemed inadequate for computational modelling, human evaluation would last longer and would harbour inconsistencies. Therefore, computational interventions are expected to meet the following minimum thresholds.

1. Reproducibility and Consistency

Assessment and evaluation of cleft repairs are critical operations. These operations could easily be indicators for the level of commitment and skill of the surgeons. Following surgical repair, it is significant to evaluate the outcome in a consistent manner. If there exist any evaluation doubts, then reproduce the evaluation to aid the next course of action in the treatment plan/ cycle. This saves professional reputation

and resources. Computational methods can be automated and called upon repeatedly, whenever they are needed (Low, Bentley and Ghosh, 2020). This is unlike human evaluation team whose assessment requires an assembly which is hard to convene.

2. Instant Evaluation Outcome

The urgency of the evaluation of a cleft repair is important as is its consistency. Computational methods have the capability to produce instant evaluation of multiple cleft repair outcomes in a single operation command. This is unlike human specialists. From a research and practical perspective, this has potential to serve a great purpose in specialised equipment built for evaluation purposes only. Alternatively, a mobile application can be conceived for this purpose. Again, it is impossible having human specialists improvised as a mobile platform.

3. Diverse Cleft Evaluation and Dataset Augmentation

Several cases of cleft lip exist. Given a proper dataset of repairs from all the possible cases, a computational model is a better and reliable alternative. Once a model has been trained on an appropriate dataset that has been improved upon by augmentation and transfer learning (Wang et al., 2021), it serves a better purpose than several human specialists. Typically, specialist surgeons or carers handle a single aspect of the facial malformation. Therefore, the classification of different cleft lip conditions is possible with the power of modelling, unlike with human subjective and qualitative evaluation. Eventually, computational modelling together with the influence of artificial intelligence and machine learning helps with creation of a robust, faster, and scalable assessment mechanism.

Chapter 4 Shape Analysis Towards Cleft Lip Treatment Assessment

4.1 Introduction

This is the study of recognition and quantification of geometric properties and characteristics of objects or structures. This entails mathematical and computational analysis of shapes to extract meaningful information for shapes comparison, shapes classification into categories, and shape variations study (Loncaric, 1998a).

Shape analysis is applied in various disciplines: computer vision, computer graphics, image processing, medical imaging, biology, and engineering. It plays a central role in specific tasks such as object recognition, shape matching, shape deformation, shape registration, and shape-based segmentation (Tabia and Laga, 2017, Arnaudon, Holm and Sommer, 2019). Shape-based segmentation has been innovatively applied in the representation of facial regions of interest (FRoI) by mapping of boundaries using contours.

To successfully carry out shape analysis, we need to understand the following crucial aspects:

1. **Shape Representation**

Shapes can be represented using different mathematical representations, such as point clouds, contour curves, meshes, implicit functions, or parametric models. The choice of representation depends on the nature of the objects and the specific analysis requirements (Arnaudon, Holm and Sommer, 2019). Contour curves are key to the representation of the facial features and RoIs given their non-uniform structure.

2. **Shape Descriptors**

These are mathematical features or properties that capture the characteristics and provide quantitative representations of a shape. Such features include curvature, area, volume, moments, or Fourier coefficients and greatly facilitate shape comparison, classification, and retrieval (Rahmann, 2000). OpenCV can use inbuilt functions and procedures to retrieve shapes. Otherwise, contours and edges detection are fundamental for shape retrieval.

3. **Matching of Shapes and their correspondences**

Shape matching aims to find correspondences or similarities between shapes. Matching techniques can be based on geometric features, such as point correspondences, contours, or surface features (Klingenberg, 2015). When a shape is partitioned into different segments, it is possible to compare the segments using different transformations and/or quantify the similarities/dissimilarities.

4.2 Context and Problem Definition

The shape analysis approach leverages on the fact that digital images contain useful features that facilitate analysis and research studies. Such features (both high-level and low-level) can be extracted from facial images and analysed to support automatic assessment of CL treatment outcome. This approach is based on low-level features of the lips and/or mouth region. The mouth boundary is detected following successful segmentation, as proven by the ground truth. The CL condition distorts the shape of the mouth. Therefore, following surgical treatment of the cleft lip, is it possible to determine that proper/considerable restoration of low-level features of the mouth region happened? Segmentation is a renowned technique for identification of the mouth region from a given facial image, using distinct lip colour and skin texture (Rohani et al., 2008, Saeed and Dugelay, 2010, Shoba and Sam, 2020). For example, the shape of the mouth is determined by some basic features such as the mouth corners, philtrum, and the upper and lower lips boundaries (upper and lower vermillion borders) (Carey et al., 2009, Kar et al., 2018). The alignment of the different features on either side of a potential symmetric axis is considered a measure of treatment outcome success or failure.

The challenge is reduced to determining the symmetric axis using the mouth region features. Consequently, the overall structure similarity of the mouth region, folded over the symmetric axis, should be computed to quantify the treatment outcome assessment.

4.3 Materials and Methods

The following components and steps constitute the pipeline of this method: mouth detection, symmetrical axis determination, similarity measurement, and numerical score estimation. Mouth detection is vital for clear determination of the visual features of lips, vermilion lines and mouth corners from a given partial facial image.

4.3.1 Dataset and Tools

The data set consists of four different categories of 25 partial facial images each. For ethical reasons, the facial images partially only reveal the nose and mouth/lips. In addition, it was also intended that human assessors are not biased by any other facial features. The dataset in this chapter consists of only the 25 images with a higher scoring or evaluation agreement among the human evaluators/ raters. Consistency and uniformity in assessment of the ground truth visual data is vital for reliability of ground truth numerical data. The latter category is also referred to as human numerical scores. However, in conducting the preliminary experiments, just like the human raters, all the 250 images were tested.

The first category constitutes the 25 raw facial appearance images following CL treatment. This category is presented to human assessors, either in hard copy format or digitally for outcome assessment estimation. The second, third and fourth categories are ground truth (GT) images generated by 3 different human experts. They have been coded as GT1, GT2, and GT3 respectively. The different human experts manually draw/trace the mouth/lip region boundary using the open-source ImageJ software (Abràmoff, Magalhães and Ram, 2004, Schindelin et al., 2015).

The ground truth categories of the dataset serve as validation images for the segmentation and the assessment prediction mechanism. Subsequently, Human numeric scores (HNS) were generated through a subjective appearance assessment process aided by statistical coding of assessor's description of the individual images in the raw dataset.

In this method, all the images of the 4 categories are automatically assessed and a numeric score is then generated. The automatically generated score is coded as the automatically estimated numeric score (AENS). This process is implemented using Python Programming Language[7], version 3.7. Compatibility tests were made for versions 3.8 – 3.10. Different software packages and libraries in the Python development ecosystem were also used:

1. Open Computer Vision Library (OpenCV)[8];

[7] https://www.python.org/
[8] https://opencv.org/

2. Matplotlib is used for creation of static, animated, and interactive visualisation with Python programming language.
3. Keras open-source library, used for development, compilation, and execution of deep neural network-based segmentation algorithms.

Computation of AENS is performed on GT1, GT2 and GT3, hence AENS can be respectively appended to the different dataset categories as GT1-AENS, GT2-AENS and GT3-AENS.

4.3.2 Applied Image Pre-processing Techniques

The following techniques were applied to the dataset before basic features were extracted.

1. **Resizing and Scaling**

Resizing an image involves changing its dimensions while preserving the aspect ratio whereas scaling changes the size of an image with little regard for the aspect ratio. Resizing and scaling are extensively used to standardise image dimensions, reduce computational complexity, and ensure compatibility across different algorithms and models. In this research, the images in our dataset needed uniform dimensions to perform objective quantitative analysis and scoring. Because the different images in the CCUK dataset were generated with different dimensions, resizing the different images was a natural step to take. For example, some images' dimensions were 498 by 487, 500 by 526, and 712 by 683, among other non-uniform dimensions. Additionally, if resizing and scaling are appropriately conducted as seen in Figure 4.1, image resolution is not compromised. Resizing helps to reduce the computational power due to reduced number of pixels. However, context-aware image resizing is becoming popular to reduce on computational load before the main computational steps are undertaken (Avidan and Shamir, 2007). In this chapter all the images were resized to 296 by 320 as opposed to a square dimension because a human face is not square. Besides, resizing can help with reducing on the image background, hence easily getting or separating the object or desired RoI from the input image.

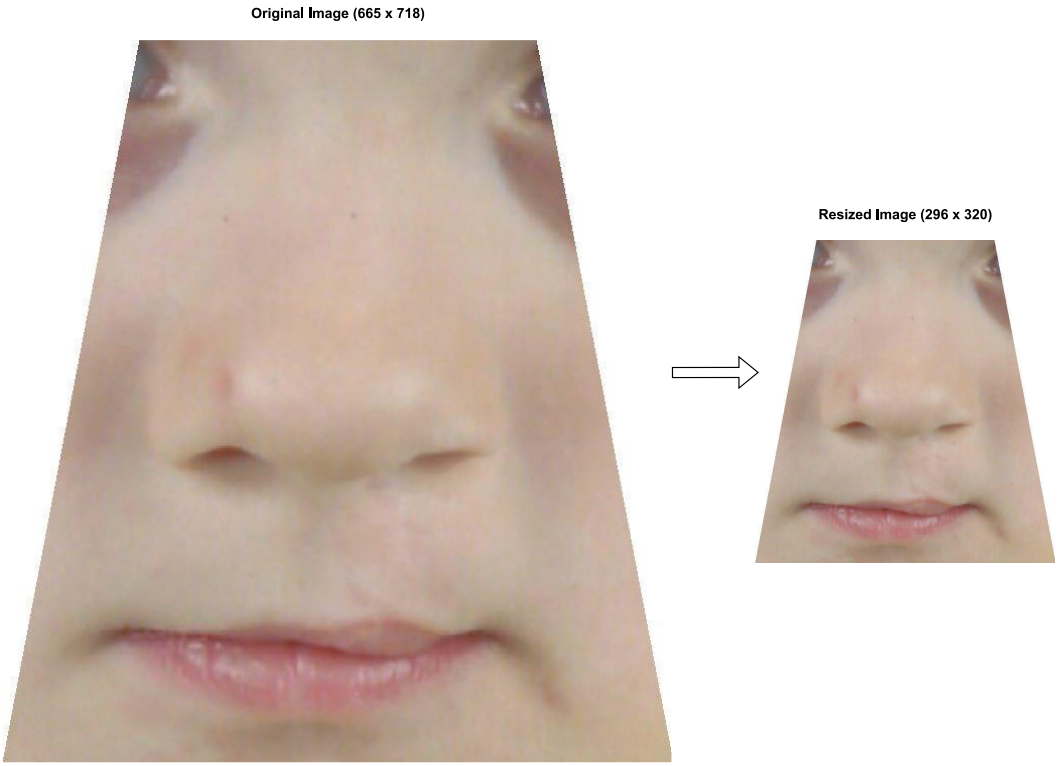

Figure 4. 1: Appropriately resized and rescaled sample from the dataset. Left is the original input image with a higher resolution while the right image is carefully resized and scaled down image whose resolution has been maintained.

2. Gray Scaling

This is the transformation of a colour image into a single-channel image where each pixel represents the intensity of light. The image pixel values are either 0 or 1 (0 or 255). Ordinarily, the average of the different colour bands is computed to associate binarisation as seen in Figure 4.2. Open Computer Vision (Open CV) (Intel, Santa Clara, California, USA) (Pulli et al., 2012) among other packages can be dynamically used to facilitate these processes. Grayscale images simplify processing tasks by reducing the computational complexity and removing colour-related information that may not be relevant to certain applications such as edge detection or texture analysis.

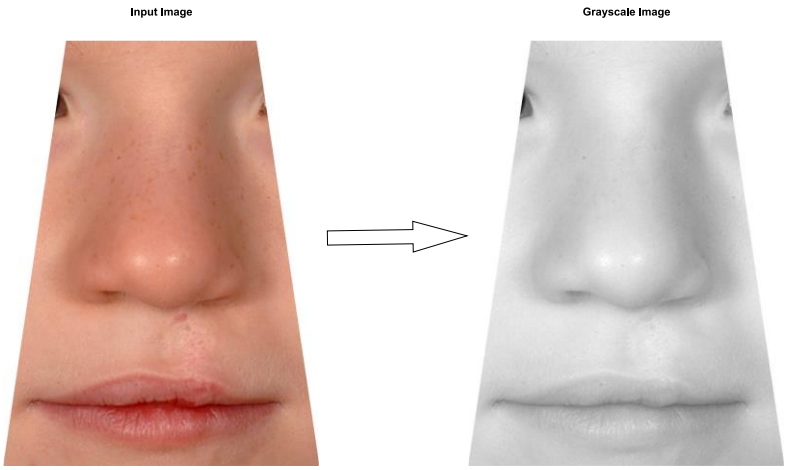

Figure 4. 2: Input colour image - left and grayscale image - right.

3. Denoising

Noise can substantially reduce image quality and affect subsequent analysis tasks. Denoising techniques aim to reduce unwanted differences caused by sensory limitations (from poor storage or heterogenous sources), compression, or other sources of noise. Common denoising methods include Gaussian blur, median filtering, and total variation denoising. Figure 4.3 shows the Gaussian blurring performed using different filter sizes. These techniques help improve image clarity and enhance the accuracy of image analysis algorithms such as the automatic appearance assessment approaches.

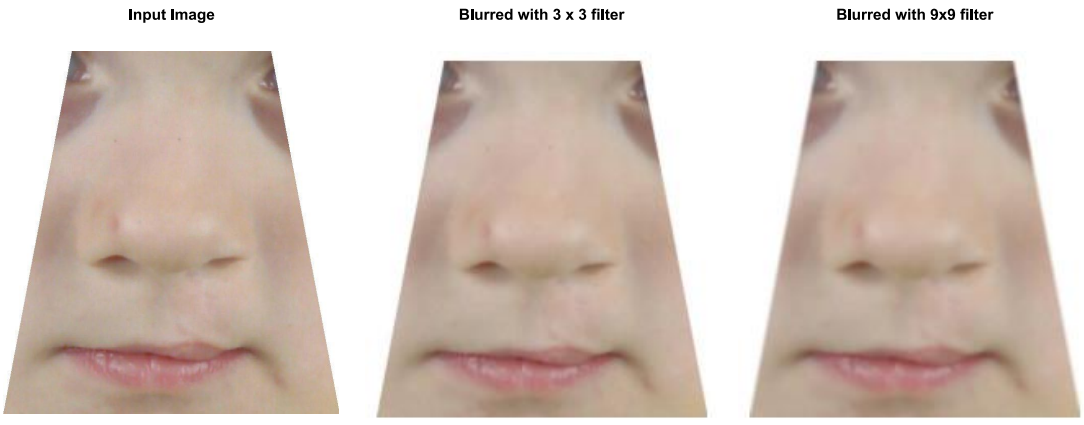

Figure 4. 3:Denoising using the Gaussian approach with different filter sizes of 3 (middle) and 9 (right).

4.3.3 Feature Description and Detection

All the facial images have been anonymised for ethical and other reasons as stated previously. On this note, it is impossible to detect other facial features, implying that only limited features can be identified. The focus is on detecting the features of the mouth region through segmentation. The anatomy of the human mouth region consists of the following key parts: the vermillion border (upper and lower), oral commissures (left and right) and the philtra ridges (left and right, separated by philtrum) (Carey et al., 2009, Berlin et al., 2014).

Similarly, according to the anatomy of the human face, ideals of facial beauty indicate that the mouth region should be in the lower third of a given facial image (Prendergast, 2011, Hashim et al., 2017). Because the skin colour and the lips may be indistinguishable, contrast enhancement and selection of suitable colour transform is inevitable. To mitigate this, the segmentation method should consider the semantics of individual pixels, first discussed in 1987 (Gritzman, Rubin and Pantanowitz, 2015). While traditional techniques which perform segmentation as a binarisation task usually under- perform at medical imagery analysis tasks (Kuruvilla et al., 2016, Wang, Wang and Zhu, 2020), the deep learning based semantic segmentation method (Yu et al., 2018), has been employed in this specific approach. Nevertheless, residues such as scars, open mouth and runny nose still influence the segmentation outcome. Semantic segmentation enhances edge detection by creating a sharper contrast between the surrounding skin and the mouth region, hence facilitating shape identification and feature extraction. The ideal mouth region mainly consists of soft tissue features defined below:

1. PR_L and PR_R: The philtra ridges identified as one of the upper most extreme pixels on the left-hand and right-hand sides of the philtrum, found along the upper mouth boundary, respectively. Also, PR is short form for philtrum ridge.
2. OC_L and OC_R : The left-hand and right-hand side mouth corners identified as the most extreme pixels on the left-hand and right-hand sides, located along the mouth boundary, respectively. OC is short form for oral commissure.
3. VB_U and VB_B are a list of pixels constituting the upper and lower mouth region boundaries, stretching between OC_L and OC_R. VB stands for vermillion border. Consequently, the two lists are defined as sets, Equation 1 and 2:

$$VB_U = \{u_1, u_2, \ldots, u_{\mathrm{m}}\} \quad \text{(Equation 1)}$$

$$VB_B = \{b_1, b_2, \ldots, b_{\mathrm{n}}\} \quad \text{(Equation 2)}$$

Where u_{i} and b_{j} are pixels in each 2D grayscale image I, $1 \leq i \leq m$, $1 \leq j \leq n$, $|VB_U| = i$, $|VB_B| = j$,

4. The mouth boundary B is a combined list of VB_B and VB_U . Collectively, it is also known as the largest non-nested detected contour in the face, represented in Equation 3 below:

$$B = VB_B \cup VB_U \quad \text{(Equation 3)}$$

Where $VB_B \cap VB_U = \{OC_L, OC_R\}$, $PR_L, PR_R \subset VB_{\mathrm{U}}$, $OC_L, OC_R \in B$, and n, m are the list sizes of VB_B and VB_U respectively.

5. The line that links OC_L and OC_R is not always parallel to the horizontal plane of the image. The line's possible orientation angle θ to the horizontal plane dictates the magnitude of rotational transformation (Figure 4.4). If $\theta < 0$, rotate anticlockwise; otherwise, rotate clockwise. Such orientation may influence how human subjects visualise and assess the different facial images.

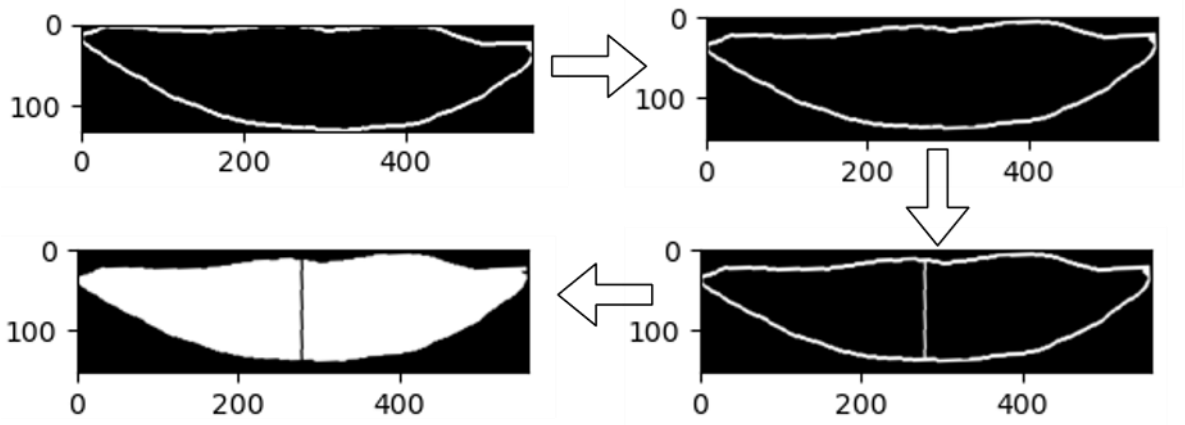

Figure 4. 4: This is an example for boundary extraction, rotation, and symmetry axis detection of a cropped mouth lip image. Top row - left: mouth corners are at different elevations from the horizontal axis. Top row - right: After anticlockwise rotation mouth corners are at the same elevation. Bottom row shows the symmetric axis (black and white).

In Figure 4.1, the top row – left: Mouth corners are at different elevations from the horizontal axis. Top row - right: After anticlockwise rotation mouth corners are at the same elevation. Bottom row shows the symmetric axis (black and white).

4.3.4 Symmetric Axis Detection and Measurement

Symmetry is defined as 'Harmony of Proportions' (Weyl, 1952). In his works on Mathematics, Prof. Hermann Weyl dwelt on geometric concepts of symmetry such as bilateral, translation, rotation, among others.

The Oxford Advanced Learners Dictionary defines symmetry as "the exact match in size and shape between two halves, parts or sides of something" or "the quality of being very similar or equal". Arguably, the term asymmetry can be inferred as the contrary. Several studies have indicated that aesthetically pleasing objects have a higher degree of symmetry (Penton-Voak et al., 2001, Little and Jones, 2003, Little, Jones and Debruine, 2011, Bella et al., 2016). Symmetric axes offer guidance towards taking measurements for size, similarity, equality, and categorisation of the shape, sides of the different halves, parts, or sides (Wei et al., 2022).

Several approaches have been previously used in the general detection of symmetry. Related methods are discussed in (Deng, Loy and Tang, 2017). However, those techniques utilised many more local and invariant object features with higher contrasts. This approach utilises basic lip and mouth features instead, like the perception of human assessors. The midpoint D, computed in Equation 4 is a position where the vertical symmetric axis is plotted through the image plane.

$$D = (OC_L + OC_R)/2 \quad \text{(Equation 4)}$$

A vertical straight line plotted through D and crossing the lower and upper mouth boundaries ensures slicing the mouth region into two shapes, left-side shape, sh_l and right-side shape, sh_r. The evenness or variance is computed and categorized using the structural similarity index measure, denoted by S (Wang et al., 2004). S is an aggregated rational number ranging between -1 and 1 for colour images or 0 and 1 for binary images. sh_l and sh_r are considered as independent and unique shapes over which to compute S. S is an aggregate of luminance l, contrast c, and structure s, as expressed in Equation. 5 below:

$$S(sh_l, sh_r) = [l(sh_l, sh_r)^{\alpha} \times c(sh_l, sh_r)^{\beta} \times s(sh_l, sh_r)^{\gamma}] \quad \text{(Equation 5)}$$

Where $\alpha = 1, \beta = 1$, and $\gamma = 1$ for easier implementation. Since the dimensions of sh_l and sh_r should be similar, sh_r is vertically flipped along the vertically plotted symmetric axis. Setting the default statistical parameters of l, c and s (Wang et al., 2004) gives the usable form of the parameter S in Equation 6 below:

$$S(sh_l, sh_r) = \frac{(2\mu_{sh_l}\mu_{sh_r} + C_1)(2\sigma_{sh_l sh_r} + C_2)}{(\mu_{sh_l}^2 + \mu_{sh_r}^2 + C_1)(\sigma_{sh_l}^2 + \sigma_{sh_r}^2 + C_2)} \quad \text{(Equation 6)}$$

Where μ_{sh_l}, σ_{sh_l}, μ_{sh_r}, and $\sigma_{sh_l sh_r}$ are the mean and standard deviations of pixels in shapes sh_l and sh_r respectively, $\sigma_{sh_l sh_r}$ is the standard deviation of the pixels in sh_l and sh_r,

$C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$, $k_1 = 0.01$, $k_2 = 0.03$, $L = 2^2 - 1$ and p is the number of bits per pixel.

4.3.5 Quantitative Modelling for Outcome Assessment

The next step is to quantitatively assess the CL treatment appearance outcome. This is accomplished by computing the structural similarity index measure S by converting S to a numeric score in the range of 1 and 5, where 1 = 'Excellent', 2 = 'Good', 3 = 'Fair', 4 = 'Poor' and 5 = 'Very Poor'. The mathematical model is a function f of S. Consequently, $f(S)$ should fulfill the following boundary and monotonicity functions: $f(0) = 5$, $f(1) = 1$ and $f(S)$ is monotonically decreasing. Therefore, $f(S)$ is the AENS.

Modelling is the formation of a relationship between variables. Normally, between independent and dependent variables. For example, in prediction of prevalence of illnesses, several factors are modelled (Tiwari, Deyal and Bisht, 2020). In this work, S can be independently computed from each visual image following successful symmetry detection. Therefore, defining the mathematical models below is intended create a relationship between AENS and S. The former is the dependent variable while the latter is the independent variable. Consequently, there is need for conversion of S into AENS using different mathematical formulae. The range of AENS must be between 1 and 5.

The three models in Equation 7, 8, and 9, were innovatively designed to aid a comparative study of the relationship between S and $AENS$ (also $f(s)$), resulting into the treatment outcome assessment as a number between 1 and 5.

The design of these functions is based on several assumptions. The major assumption is that the relationship between S and AENS could be linear or non-linear.

$$f(S) = 5 - 4S \quad \text{(Equation 7)}$$

$$f(S) = 5 - 4S^3 \quad \text{(Equation 8)}$$

$$f(S) = 1/(0.2 + 0.8S^2) \quad \text{(Equation 9)}$$

Equation 7 is derived by assuming a linear relationship between AENS and S.

With a purpose to model an implicit linear relationship between S and $f(S)$. The former is independent while the latter is dependent on the former.

We needed to find a constant value (a coefficient) that represents the rate at which $f(S)$ decreases in case S increases.

Additionally, there is need to assume the intercept which is a baseline value of $f(S)$ in case $S = 0$.

Based on $y = mx + c$ as a general linear equation representation, the gradient, $m = -4$, and $c = 5$.

This leads to arriving at $f(S) = 5 - 4S$.

The accuracy of equation 7 to predict AENS has been presented in subsequent sections.

In a similar way, other degrees of mathematical models were derived from $f(S)$ and its relationship to S.

Figure 4.5 presents three scenarios which have been considered for the generation of sh_l and sh_r, to facilitate further comparison and definition of the two shapes.

Scenario 1: Parameters are calculated over the entire mouth blob.
Scenario 2: Parameters are calculated over the entire mouth boundary only.
Scenario 3: Parameters are calculated over the upper lip blob only.

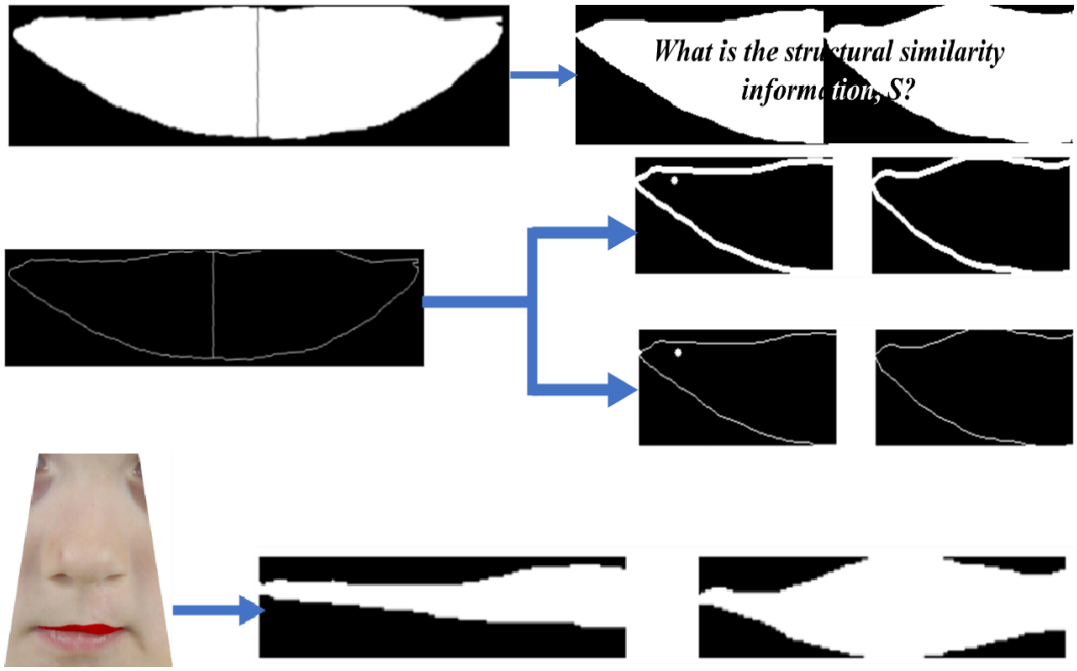

Figure 4. 5: Different scenarios for parameter calculation. Top: Scenario 1 where the entire mouth region blob, consisting of upper and lower lips has been split into right and left blobs (sh_l and sh_r respectively). sh_r has been flipped. $Middle$: Scenario 2 with the boundaries defined with different thicknesses of 1 and 3 pixels, respectively. $Bottom$: Scenario 3.

In Figure 4.5, Top: Scenario 1 where the entire mouth region blob, consisting of upper and lower lips has been split into right and left blobs (sh_l and sh_r respectively). sh_r has been flipped. Middle: Scenario 2 with the boundaries defined with different thicknesses of 1 and 3 pixels, respectively. Bottom: Scenario 3.

4.4 Implementation Summary

This section presents the summarised/ procedural implementation of the solution discussed in this Chapter. Figure 4.6 is the most crucial algorithm (or function) that anchors the implementation of the discussed solution.

1. Convert image to grayscale.
2. Find contours in grayscale image.
3. For each of the contours:
   3.1 Calculate the area of contour and store it as a list.
4. Find the index of the contour with the largest area in the list.
5. Create an empty image with the same size as the grayscale image.
6. Draw the largest contour on the empty image.
7. Find the extreme points (i.e., left most, right most, topmost, bottom most) of the largest contour,

8. Calculate the center of mass using the left most and right most extreme points.
9. Calculate the angle of orientation based on the line joining leftmost and rightmost points.
10. Use angle of orientation to aid plotting a line perpendicular to a specific line and draw an arbitrary long bisector/ mouth radius.
11. Plot potential vertical symmetric axis through the center of mass.
12. Determine the upper most points which may correspond to the philtral ridges. One of the points is the upper most pixel along the boundary while the second point can be geometrically projected.
13. Calculate the perspective transformation matrix.
14. Warp the image using the perspective transformation matrix to obtain the straightened image (warped) with width and height.
15. If the rotation angle of bounding rectangle is less than -45 degrees:
    15. 1 Rotate the warped image (warped) by 360 degrees and orientation angle and the rotation angle of the bounding rectangle.
    Else:
    15.2 Rotate the warped image by the rotation angle of the bounding rectangle less the angle of orientation.
16. Return the warped image.

Figure 4. 6: Most significant algorithm for extraction of the region of interest for generation of the largest boundary (or contour) of the mouth region.

The above algorithm essentially detects the largest useful boundary (the contour) in the facial image. This is the boundary with the largest area. The identified contour is used on the corresponding facial image from which other parameters such as mouth shape similarity or dissimilarity are measured. Other algorithms are presented in Figure 4.7 and Figure 4.8.

1. Use the computed median and any auto-determined upper and lower threshold for appropriate edge detection.
2. Perform dilation to close some gaps between objects edges.
3. Select from contours list in the image of the contour with the largest area and draw it.
4. Draw the order of the bounding box points: bottom left, top left, top right, bottom right.
5. Draw the width and height of the detected bounding box (usually rectangular).
6. Determine coordinates of the points in the bounding box points. This is after the rectangular shape has been straightened (at zero degrees to the horizontal).
7. Computer the perspective transformation matrix.
8. Directly warp the rotated rectangle to get the straightened or horizontally aligned rectangle.

Figure 4. 7: Algorithm for cropping of the face image for the region of interest.

When analysing the structural shape of the mouth, elimination of other facial components is performed through cropping out the mouth. Algorithm in Figure 4.4 can be used for this purposed. Before cropping, the region of interest should be semantically segmented using the basic building block of a deep learning network, whose skeleton is presented in the algorithm in Figure 4.5.

1. Perform a suitable convolution operation with padding.
2. Create a suitable convolutional layer with the specified number of input planes, output planes, and stride (usually, the default=1).
3. Set padding to 1 to maintain the spatial dimensions of the input.
4. Initialise the Network Basic Building Block with input channel, output channel, and stride.
5. If input and output channels are not the same or stride is not 1, create a down sample operation.
6. Else go back to step 1.
7. Calculate the forward pass of the Basic Building Block.
8. If down sample is not None, apply the down sample operation to a given output.
9. Perform an element-wise addition of shortcut and residual.
10. Create a layer with Basic Building Block units.
11. Initialise the layers list with a Basic Building Block that inputs appropriate input and output channels and the specified stride.
12. Repetitively add specific Basic Building Block numbers to the Layers container
13. Create a sequential container with the Layers in the list and return it.
14. Finally, involve a Residual Network building block and append the above to return a module for use in an adapted semantic segmentation network.

Figure 4. 8: Basic building block for an adapted network to semantically segment the facial image's mouth region.

4.4.1 Complexity Analysis

Complexity analysis of different computational interventions determines the efficiency and performance characteristics of either the blueprint or the implementation(Bichler, 2017). Complexity analysis helps with understanding how the runtime or memory usage of the implementation scales with input size (Erciyes, 2014). During analysis, the following activities were completed:

1. Identification of key operations such as loops and recursive operations.
2. Counting of key operations which considers the number of times each identified operation is executed based on the input size.

3. Derivation of time complexity as the sum up the counts of operations and express the total as a mathematical function of the input size, such as n.
4. Analyse the memory usage of the code, counting variables, data structures, and recursiveness. Additionally, the space usage is expressed as a function of input size, n.
5. Comparison with known classes, a comparison was made between the derived complexities with known complexity classes such as linear, quadratic, logarithmic.

In carrying out the above activities, the representation of efficiency metrics is performed using the Big O notation. Big O notation is a system of measurement for determining an algorithm's efficiency (Rutanen et al., 2013). Furthermore, this metric gives an estimate of the duration for a code segment (implementation) to run on different sets of inputs. In some cases, experiments may be performed to measure the effectiveness of a code segment scalability with an increase in input size. The following steps were used.

1. Image Blurring

This operation applies a blur filter to the image and involves iterating over each pixel in the image and computing an average based on neighbouring pixels. The time complexity is $O(n.m)$, where n and m are the height and width of the image, respectively. The space complexity is $O(1)$ since the operation is performed in a uniform memory space.

2. Thresholds Calculation and Adapted Canny Edge Detection

Thresholds were automatically calculated in an adaptive manner as was edge detection using Canny's algorithm. This operation involves iterating over each pixel in each image. The time complexity is $O(n.m)$, the space complexity is $O(1)$.

3. Morphological Operations

Both dilation and erosion operations were performed to close and close out any pixels' gaps stemming from edge detection at the boundaries. Therefore, both operations involve iterating over each pixel in the image. In this research, the operations are performed three times, so the time complexity is $O(3.n.m) = O(n.m)$, the space complexity is $O(1)$.

4. **Finding Contours**

There is need to find all possible contours in the edge-detected image. The intention is to eventually get a contour with the largest area.

The time complexity depends on the number of contours found and their sizes. In the worst case, the time complexity can be $O(n.m)$, and the space complexity is $O(n.m)$ to store the contours.

5. **Calculate Area**

There is need to loop over each contour to calculate its area. If there are c contours, the time complexity is $O(c)$, and the space complexity is $O(1)$, since it only involves storing the areas.

6. **Selection of Contour with Largest Area**

The selection of the contour with the largest area involves finding the maximum area from the list of areas. If there are c contours, then the time complexity is $O(c)$, and the space complexity is $O(1)$.

7. **Calculate Center of Mass using Contour Points**

Given the largest contour, its points are traversed (through a loop iteration) with the aim of calculating the center of mass. If the largest contour has p points, then the time complexity is $O(p)$, and the space complexity is $O(1)$.

8. **Compute the Symmetric Axis using Extreme Points**

Extreme points are determined in one operation after which the symmetric axis is computed. This is a basic step that involves constant access and operations. Therefore, both the time complexity and space complexity are $O(1)$.

9. **Model Construction and Score Calculation**

These operations involve constant time complexity $O(1)$ because they are based on individual pixel values, derived from the structural similarity index measure.

10. **Image Split and Resize**

Before any structural similarity index measure is computed, the region of interest is split into two. Both image splitting and resizing depend on the dimensions of the

images. Let d_1 and d_2 be the dimensions of the two images. The time complexity is $O(d_1.d_2)$ and the space complexity is $O(d_1.d_2)$ because new images are created.

Overall Time Complexity

The dominant time complexity is $O(n.m)$, due to the image processing operations involving iteration over each pixel in the image. If the image dimensions are the same, as may be the requirement in most image processing frameworks, then $n = m$. This implies that the time complexity is $O(n^2)$.

Overall Space Complexity

The overall space complexity is $O(n.m)$, due to the storage of contours and intermediate images. Using a similar analogy, the overall space complexity is $O(n^2)$.

4.5 Outcomes of Shape Analysis

This section presents both the qualitative and quantitative experimental results of the automatically programmed rating (PR) method and are compared with others, when applicable.

4.5.1 Preprocessing Summary

A 3 by 3 filter was used to denoise and smoothen the images. Additionally, before performing any edge detections, lower and upper thresholds were generated using the NumPy library functions of image median and minimum and maximum. Finally, the morphological operations of erosion and dilation through three iterations each were used to close any gaps between object edges. In order to minimise the computational power needed, some images are processed as grayscale (Vidal and Amigo, 2012, Ballabeni et al., 2015). Some of the outcomes of pre-processing are in Figures 4.1, 4.2 and 4.3 above.

4.5.2 Image Segmentation

Following preprocessing is segmentation of the partial facial images using different approaches as presented below.

1. **Thresholding-Based Segmentation**

Thresholding techniques involve selecting a global threshold value and assigning pixels above or below that threshold to different segments. This simple yet effective method is widely used, especially in binary segmentation. Otsu's thresholding

technique uses a global threshold value, but it is not chosen (Otsu, 1979). The threshold, T, is determined automatically. In the simplest form, the Otsu algorithm outputs a single intensity threshold that separates pixels into two classes, foreground, and background. Figure 4.9 shows some outcomes from Otsu thresholding segmentation. Automatically determined Otsu threshold is 199 (left) and 221 (right)

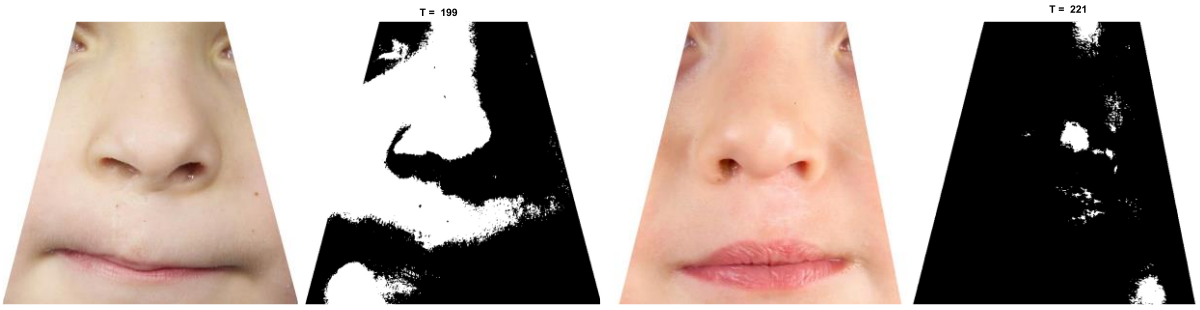

Figure 4. 9: Results from Otsu thresholding where threshold T has different values: second figure where T=100, third is where T=125 and fourth is where T=150. T=125 shows a better result for the mouth region.

However, this technic works accurately for bimodal images. The bimodal images are those images whose histogram has two peaks. The threshold value is the approximate value of the middle of these two peaks.

Additionally, Figure 4.10 presents the segmentation outcomes for another facial image using some traditional approaches.

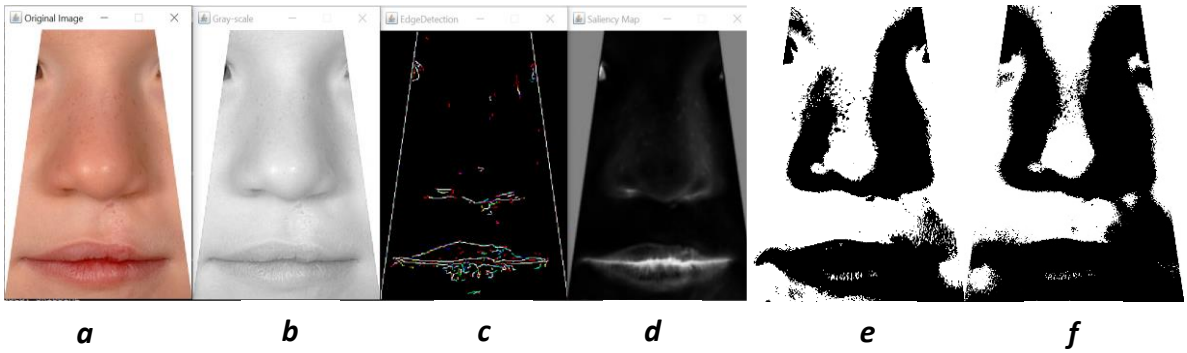

Figure 4. 10: facial features segmentation outcomes using traditional approaches. a: input image, b: grayscale image, c: canny edge detection, d: saliency map detection using maximum symmetric saliency detection result, e: Otsu segmentation result, f: Moment-preservation segmentation.

Clearly, the outcomes in Figures 4.9 and 4.10 are not suitable for analysis.

In Figure 4.10, there are the following outputs: a: input image, b: grayscale image, c: canny edge detection, d: saliency map detection using maximum symmetric saliency detection result, e: Otsu segmentation result, f: Moment-preservation segmentation.

However, result c and d portray the shape of the mouth region with results in e and f showing a vague mouth shape. Such is the binarisation challenge using traditional algorithms.

2. Clustering-Based Segmentation

This approach utilises clustering algorithms, such as k-means (MacQueen, James and others, 1967, Hartigan and Wong, 1979) and mean-shift (Comaniciu and Meer, 2002), to group pixels into clusters based on their similarity in feature space. These methods allow for unsupervised segmentation and are effective in scenarios where the number of segments is either unknown or varies or both.

Classical clustering algorithms operate by partitioning the feature space, which can include colour, texture, or other relevant image features. The algorithm steps include feature extraction and post processing. These flexible techniques are widely used in various applications such as image analysis and pattern recognition (Chen et al., 2015, Xia et al., 2016). Figure 4.11 shows the outcomes of the selected clustering algorithms.

In the CCUK dataset, overall, K-means is a better clustering approach because the number of clusters are easily set compared to the mean shift-based cluster outcomes. Our dataset easily accepted a fixed number of clusters of 3. Mean shift failed to automatically generate a consistent number of clusters around which datapoints could converge.

K-means (KM) and mean shift (MS) usually produce unsatisfactory results, though better than the traditional approaches. A comparative presentation between the MS (spatial $bandwidth = 20$, $colour\ bandwidth = 7$), KM ($k = 3$) and bilateral real-time semantic network (SN) segmentation is given in Figure 4.11.

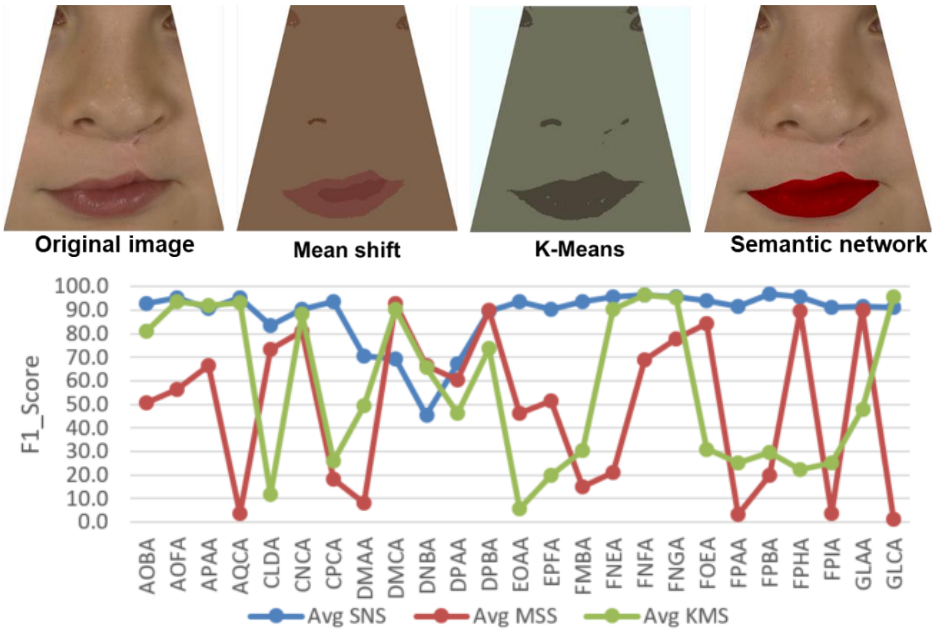

Figure 4. 11: More segmentation results: Top row: 1st: input image, 2nd Result from mean shift segmentation, 3rd is k-means segmentation outcome and 4th: bilateral semantic network segmentation. Bottom row: the average $F1\_Score$ for the 4 categories of the dataset of 25 images is calculated and plotted against the image ID. The average from SN segmentation ($Avg\ SNS$) is better than the $F1\_Score$ for MS and KM segmentation results.

The performance measurement used is F1 $Score$ percentage: the higher the better. The F1 score is a metric commonly used in machine learning and statistics to assess the performance of (mainly classification) models, especially when dealing with imbalanced datasets. It combines two fundamental evaluation metrics, precision, and recall, into a single score to provide a balanced measure of a model's accuracy (Vujović, 2021).

Precision is the ratio of correctly predicted positive instances (true positives) to all instances predicted as positive (true positives + false positives) while Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive instances (true positives) to all actual positive instances (true positives + false negatives).

F1 Score: The F1 score is the harmonic mean of precision and recall and is calculated using the following formula:

$$F1\ Score\ =\ 2\ *\ (Precision\ *\ Recall)\ /\ (Precision\ +\ Recall)$$

The harmonic mean gives more weight to lower values, making the F1 score a suitable metric for situations where you want to balance precision and recall. It ranges from 0 to 1, with higher values indicating better model performance.

Clustering-based approaches (such as KM and MS) yielded worse outcomes with discontinuous areas and boundaries compared to deep learning methods (such as SN). Gaussian blurring, morphing and dilation were usually used to mitigate such issues. The segmented mouth region (the required RoI) is found in the bottom third of the facial image. Standardisation with a bounding box was also used to reduce the background from the image as seen in Figure 4.1, above.

For each image in the evaluation dataset of 25 images, the F1 score was calculated against the three segmentation results. For example, the segmentation result for AOBA from KM, MS and SN was used to calculate the F1-score against the AOBA input image to measure accuracy. However, AOBA has 3 versions of GT1, GT2 and GT3 from which to compute the F1-score. Therefore, the average F1-score (Table 4.4) was calculated from Tables 4.1, 4.2 and 4.3 and used to plot the graphical visualisation in Figure 4.11.

Table 4. 1: KM Segmentation-based F1 Score between predicted result (PR) and all ground truth datasets.

Set_ID	PR_GT1	PR_GT2	PR_GT3
AOBA	0.815648065	0.807862004	0.806662
AOFA	0.93320467	0.945721553	0.927642
APAA	0.908377733	0.944606788	0.910623
AQCA	0.94625209	0.928272493	0.922093
CLDA	0.108813418	0.11788113	0.12464
CNCA	0.887036643	0.85367664	0.914286
CPCA	0.260777009	0.225259705	0.29327
DMAA	0.550027168	0.534708824	0.394108
DMCA	0.875859385	0.891853014	0.942704
DNBA	0.900602186	0.870074983	0.194787
DPAA	0.261911565	0.269644224	0.860335
DPBA	0.74425136	0.736581147	0.733232
EOAA	0.057533887	0.058298996	0.0596
EPFA	0.176339721	0.188451004	0.236048
FMBA	0.299097455	0.30419909	0.307775
FNEA	0.888163071	0.921539513	0.901553
FNFA	0.966946174	0.961932127	0.96837
FNGA	0.969148432	0.946422043	0.946868
FOEA	0.315040538	0.304043716	0.311125
FPAA	0.262378386	0.246870983	0.243172
FPBA	0.311234531	0.280639802	0.29368
FPHA	0.222842974	0.225425141	0.225105
FPIA	0.28082756	0.239577911	0.238855
GLAA	0.483187845	0.479348816	0.475409
GLCA	0.949514732	0.949579103	0.967063

Table 4. 2: MS Segmentation-based F1 Score between predicted result (PR) and all ground truth datasets.

Set_ID	PR_GT1	PR_GT2	PR_GT3
AOBA	0.494092014	0.516165862	0.507945043
AOFA	0.570350252	0.575049263	0.545726737
APAA	0.679237793	0.653155941	0.664297612
AQCA	0.034911495	0.034168066	0.036595753
CLDA	0.750890614	0.739470712	0.711944975
CNCA	0.829286608	0.796070629	0.806266973
CPCA	0.187896978	0.167763158	0.194493184
DMAA	0.087337986	0.09223847	0.062444908
DMCA	0.931687243	0.932992618	0.920016692
DNBA	0.885341497	0.894374206	0.215710511
DPAA	0.742285425	0.775460931	0.289780266
DPBA	0.898543689	0.889666308	0.911403705
EOAA	0.465725962	0.456149068	0.460879511
EPFA	0.533431301	0.530143411	0.48441358
FMBA	0.154897203	0.137407478	0.158753084
FNEA	0.216843579	0.206396675	0.210046619
FNFA	0.692235457	0.689118557	0.690590869
FNGA	0.770102684	0.790837368	0.776221513
FOEA	0.846807172	0.838568268	0.84336059
FPAA	0.030351759	0.030953862	0.030539849
FPBA	0.189383706	0.201624505	0.206126875
FPHA	0.897982028	0.89681709	0.893944363
FPIA	0.035299341	0.034209646	0.035300211
GLAA	0.902059827	0.90896858	0.88216728
GLCA	0.012155434	0.012172907	0.01183122

Table 4. 3: SN Segmentation-based F1 Score between predicted result (PR) and all ground truth datasets.

Set_ID	PR_GT1	PR_GT2	PR_GT3
AOBA	0.922593654	0.929714601	0.937361431
AOFA	0.95611049	0.945148938	0.960660114
APAA	0.881634719	0.894875205	0.950560316
AQCA	0.939808501	0.951709091	0.961450855
CLDA	0.845787164	0.820419886	0.83271284
CNCA	0.892463555	0.911219663	0.911476679
CPCA	0.936068892	0.913825118	0.963893623
DMAA	0.917746155	0.917882118	0.278684821
DMCA	0.641430766	0.690865226	0.751957149
DNBA	0.202237914	0.217638392	0.938616071
DPAA	0.903187409	0.885069955	0.229224989
DPBA	0.885734676	0.909310373	0.896859021
EOAA	0.922231502	0.944904972	0.938755497
EPFA	0.892086331	0.885092418	0.936974106

FMBA	0.935641461	0.928154124	0.944181078
FNEA	0.955355332	0.952948935	0.966313541
FNFA	0.958210863	0.965622412	0.969076523
FNGA	0.970849138	0.963240743	0.937671191
FOEA	0.92736166	0.936578928	0.959624244
FPAA	0.940172444	0.919717405	0.891144557
FPBA	0.967337204	0.967680418	0.965182713
FPHA	0.95808807	0.949489796	0.957186922
FPIA	0.914362639	0.934328194	0.880944713
GLAA	0.911015288	0.913205021	0.927808086
GLCA	0.916082675	0.913603902	0.905149169

Table 4. 4: Average F1 Scores from GT1, GT2 and GT3.

Set_ID	Avg SNS	Avg MSS	Avg KMS
AOBA	93.0	50.6	81.0
AOFA	95.4	56.4	93.6
APAA	90.9	66.6	92.1
AQCA	95.1	3.5	93.2
CLDA	83.3	73.4	11.7
CNCA	90.5	81.1	88.5
CPCA	93.8	18.3	26.0
DMAA	70.5	8.1	49.3
DMCA	69.5	92.8	90.3
DNBA	45.3	66.5	65.5
DPAA	67.2	60.3	46.4
DPBA	89.7	90.0	73.8
EOAA	93.5	46.1	5.8
EPFA	90.5	51.6	20.0
FMBA	93.6	15.0	30.4
FNEA	95.8	21.1	90.4
FNFA	96.4	69.1	96.6
FNGA	95.7	77.9	95.4
FOEA	94.1	84.3	31.0
FPAA	91.7	3.1	25.1
FPBA	96.7	19.9	29.5
FPHA	95.5	89.6	22.4
FPIA	91.0	3.5	25.3
GLAA	91.7	89.8	47.9
GLCA	91.2	1.2	95.5

The F1 score results in Table 4.4 indicate that the average F1 score using SN segmentation (Avg SNS) is better than the average F1 score for using MS segmentation (Avg MS) and KM segmentation (Avg KM). Additionally, the blue plotted graph in Figure 4.11 supplements this observation. Therefore, SN segmentation is a

more reliable segmentation algorithm. There is notable variation in SN or KM or MS segmentation F1 Score values among the different datasets. This is attributed to potential inter-expert variability in generation of the ground truth visuals (Kondermann, 2013) as well as intricate variabilities of the segmentation algorithms themselves (Wang, Wang and Zhu, 2020).

4.5.3 Validation of Shape Analysis Assessment Approach

Successful segmentation is followed by identification of features and points such as boundaries (vermillion borders), extreme points (philtra ridges, oral commissures/mouth corners) and the boundaries enclosures. However, S is computed alongside the different features of the mouth region. After computing S based on the three scenarios stated above, S is converted into a numeric score (AENS). Consequently, AENS and HNS are contrasted using Pearsons Correlation Coefficient (PCC), where the higher the better.

Standardisation through image orientation and alignment using the bounding box improves the value of S, hence the AENS. The indicator of this observation is that mouth orientation has potential to affect and influence human assessors whose visual perception may be compromised. This is an interesting finding that is worth further investigation in clinical setting and ground truth collection. Table 4.5 shows this observation. The value of S is significantly improved upon before and after standardisation of the mouth orientation in Scenario 1.

Table 4. 5: Comparison of S before and after standardisation of the mouth orientation in Scenario 1.

$Category$	$Range\ of\ S$		$Average\ of\ S$	
	$Before$	$After$	$Before$	$After$
PR	$0.24 < S < 0.82$	$0.55 < S < 0.89$	0.60	0.79
$GT1$	$0.30 < S < 0.84$	$0.49 < S < 0.89$	0.60	0.72
$GT2$	$0.28 < S < 0.84$	$0.39 < S < 0.86$	0.60	0.69
$GT3$	$0.35 < S < 0.82$	$0.51 < S < 0.88$	0.64	0.72

The performance metrics of shape analysis are presented in Figure 4.8 over 3 scenarios, 3 models (defined in Equations 7, 8 and 9) and 2 options of the symmetric axis crossing position, D and D_2. D and D_2 are respectively defined in Equations 4 and 10.

$$D_2 = d(OC_L + OC_R)/2 \qquad \text{Equation 10}$$

Where an inward shift factor d of 5% has been experimented as the most effective. Naturally, mouth corners are normally not easily detected accurately due to imaging noise and shadows (Zhang et al., 2012).

Figure 4.12 also shows PCC results for HNS vs AENS for the different dataset categories. In a nutshell, PCC results for different scenarios over the different transformation models. Odd row: symmetric axis plotted at D; Even row: symmetric axis plotted at D2. Top two rows represent Scenario 1; Middle two rows represent Scenario 2 and Bottom two rows represent Scenario 3.

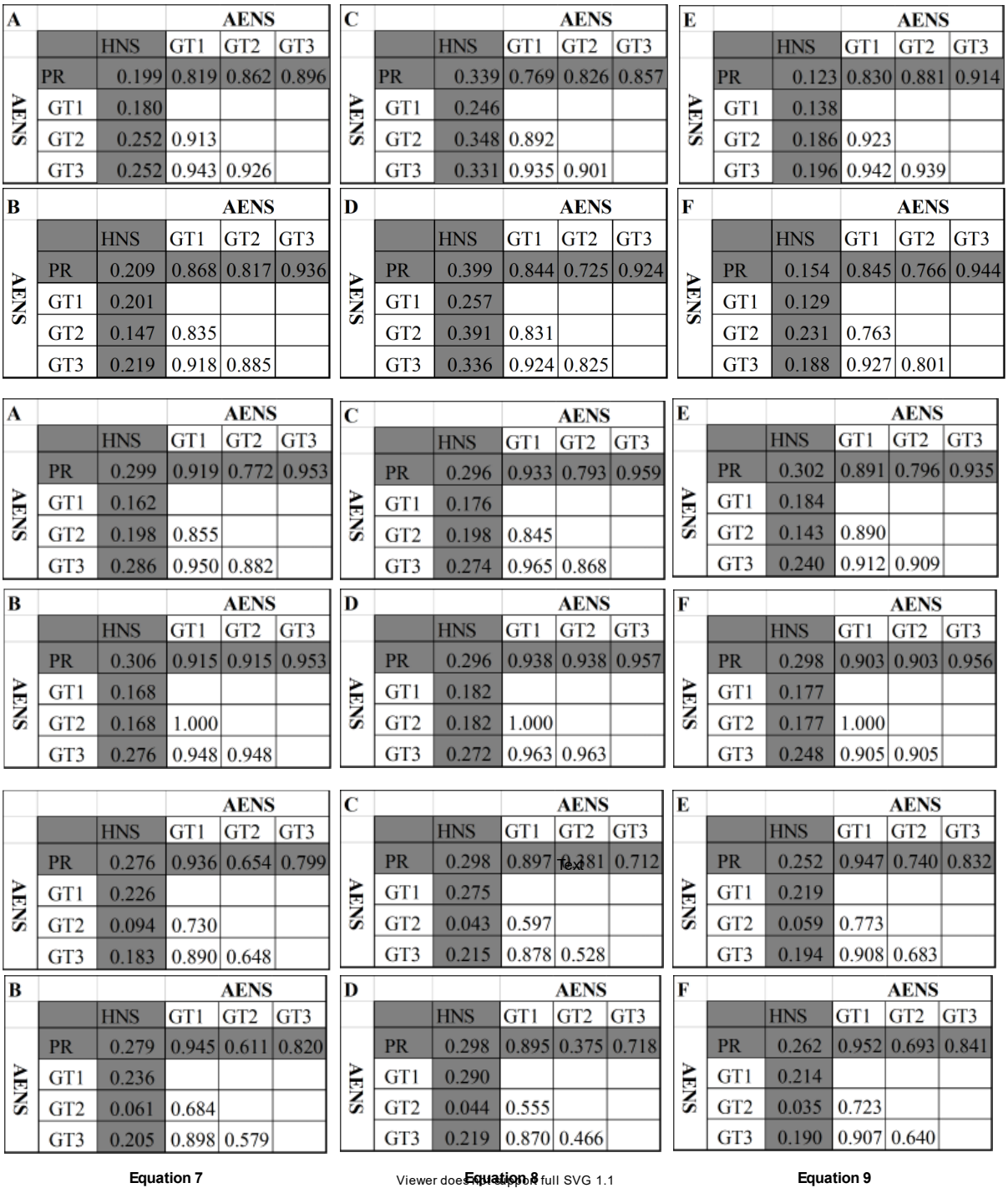

Figure 4. 12: This figure shows different tables for the PCC of the AENS vs AENS for different dataset categories, GT1, GT2, GT3 and the PR. Results showed are also for PCC of HNS vs AENS for the different dataset categories. In a nutshell, PCC results for different scenarios over the different transformation models. Odd row: symmetric axis plotted at D; Even row: symmetric axis plotted at D2. Top two rows: Scenario 1; Middle two rows: Scenario 2; Bottom two rows: Scenario 3.

The most significant PCC is between PR_AENS and HNS because the intention of automation is to discover whether the fully automatic assessment mechanism agrees with the human assessment approach.

The highest PCC is about 40% ($Scenario\ 1, Table\ D, Figure\ 4.12$) and the lowest is about 15% ($Table\ F, Figure\ 4.12$). This is partly due to the inconsistency among human assessors in assigning scores between different images. Overall, shifting the mouth corners inward improves the most significant PCC across the three models. However, the model in Equation 8 is the most robust, implying that the mapping between shape structural similarity measurements (S) and appearance assessment scores ($AENS$) is non-linear. In sharp contrast, the PCC between PR_AENS and either $GT1\_AENS$, $GT2\_AENS$ or $GT3\_AENS$ is significantly higher, as high as 94% ($Table\ F, Scenario\ 1$). Implying that automatic segmentation of the mouth regions as accurate as the human drawn mouth boundaries.

In Scenario 2, the most significant PCC is about $31\%, (Table\ B)$. There is little difference in the various correlations over different setups, indicating that the mouth boundaries may not be as predictive as expected. This is somewhat contradictory to the practice that focuses on the vermilion border lines and thus requires further investigation. Scenario 3 has produced the lowest PCC in the category of PR_AENS of 38% ($Table\ D$). A similar trend has been noticed for, $GT1\_AENS$, $GT2\_AENS$ and $GT3\_AENS$ with the same scenario. This is a clear indication that the process to determine the RoI is still challenged, thus can be improved upon.

It is noted that determining the symmetric axis using fewer features is a potential limitation of this approach. Consequently, Chapter 5 presents a hybrid approach with improved results. Additionally, the benchmark for the validity testing of this technique is based on a single approach, spearheaded by human experts.

4.6 Summary of Shape Analysis Framework

Shape analysis is an automatic appearance assessment approach for CL treatment outcomes that utilises lips and mouth features. These low-level features are considered appealing to humans and can be distinguishable to aid with appearances judgement. The features include oral commissures, philtra ridges and the vermillion border. Once the mouth region has been detected using the bilateral semantic segmentation network method and split through the midpoint of the horizontal line linking the mouth corners, the two ensued blobs are analysed for potential evenness or unevenness. To this end, the widely used structural similarity index measure (Wang et al., 2004) is employed. The measure is a rational number, which is then converted

non-linearly to a numeric score in the range of 1 and 5, like the Asher-McDade 5-point Likert Scale used by human experts. A numerical similarity computation following a symmetric axis computation is a better objective appearances assessment of the repaired lips compared to the qualitative measures proposed in (Pigott and Pigott, 2010, Deall et al., 2016b, Pietruski, Majak and Antoszewski, 2017). The discussed experimental results disclose that the automatically estimated numeric scores have relatively low correlation coefficients with human assigned score but have high correlation coefficients with those estimated from the human manually drawn mouth regions.

It is also noted that inward shift of the mouth corners by 5% improves the accuracy of the midpoint D_2 and offers an alternative for a symmetric axis position to combat the challenging nature in identifying the mouth corners with improved appearances assessment scores. Chapter 5 presents more accurate estimations of the symmetrical axis using hybrid approaches and different measurement between the two sides of the mouth regions is significantly improved. Some of the weaknesses of the approach presented in this Chapter are addressed in Chapter 6.

Chapter 5 Adaptive Symmetry from Key Landmarks Using the Hybrid Approach

5.1 Introduction

Surgical treatment of the CL condition is meant for complete or partial restoration of key facial features. Because these features define the appearance of an individual's face, they can be referred to as facial landmarks (Naqvi et al., 2022). Therefore, given that individuals are different, facial landmarks are better described biologically or genetically.

Biologically, facial landmarks are specific anatomical points or features on the human face that have consistent locations and are used as reference points for various analyses, measurements, and descriptions. These landmarks represent distinct anatomical structures or specific locations on the face and play a crucial role in understanding facial morphology, development, and variation (Shier, Butler and Lewis, 2007).

These landmarks, among others, are used in various fields, including anthropology, genetics, craniofacial surgery, and facial recognition technology, to study facial morphology, growth, and development, as well as to understand facial characteristics and their variations within and across populations (Kukharev and Kaziyeva, 2020, Naqvi et al., 2022).

In this research, partially occluded facial images are used, implying that the landmarks can only be found at the eyes & brows, nose, or mouth region. In analysis studies, facial landmarks are key anatomical points of reference to aid taking measurements and application in domains such as anthropometry, 2D/3D imaging studies, preparation for surgical procedures (Fink and Neave, 2005). Detection of some important features is of paramount importance regarding understanding the restoration and recovery of some regions following the CL surgical treatment.

The mouth lip beauty is a targeted outcome measure. The obvious distortion of the lip morphology hinders detection and identification of key features, considered essential for beauty. The features depicted from the facial appearance outcome significantly aid towards categorization as success or failure of a cleft repair. Eventually, this aids any

audit of different cleft repair practices by assessing the restoration of the mouth lips (Hashim et al., 2017, Kar et al., 2018).

5.2 Identification of Features from Partially Occluded Facial Images

Occlusion in computer vision started in the 1960s when Guzman proposed to detect faint lines in polyhedral drawings (Hoiem, Efros and Hebert, 2011). Consequently, it has been a subject of study in computer vision for detection of hidden facial features using convolutional neural networks with an attention mechanism (Li, Zeng, et al., 2019); facial appearance and shape learning to robustly detect facial features using an occlusion-adaptive deep network (Zhu et al., 2019); and cascaded pose regression (CPR) (Dollár, Welinder and Perona, 2010, Burgos-Artizzu, Perona and Dollar, 2013). However, the notion of face detection before any features are identified is a persistent component and not appropriately applicable to datasets whose facial features and shape has been significantly occluded.

Determining facial features in images/videos is predominantly premised on a detected face Therefore, face detection is a major component of facial feature identification studies. Facial anonymisation of appearance outcomes is a convention for cleft lip related studies (Lee et al., 2019). It is logically commendable and ethically a best practice for unbiased outcome assessment audit of different practices. Anonymization obstructs biased human assessment from any eye colour, ears shape, hair etc, unlike computer-based assessment (Shkoukani, Chen and Vong, 2013, Lee et al., 2019). Consequently, the images used during outcome assessment bear significant partial occlusion.

Some of the facial image features of significant importance include inner eye corners (i.e. inner canthus, lacrimal punctum and inner canthal distance), nose features (tip, ala, root, and nasal base) and mouth features (upper/lower lip vermillion, oral commissure, vermillion border) (Hall et al., 2009, Hennekam et al., 2009). Presence of these features in the facial images symbolises beauty. Therefore, computer vision tools aim to detect and locate these features, hence assess beauty using symmetry and other suitable shape defining parameters (Sharma et al., 2012). Deep learning-based methods have robust prediction capabilities to detect occluded features from visual data such as facial images.

Most facial features occur in group or pairwise classification and can be used to determine the symmetric or asymmetric nature of a facial appearance. Key mouth features are used in (Bakaki et al., 2021) to determine the symmetry axis from which shape analysis is applied for facial appearance assessment. Scars and other skin residues from surgical repair and photography effects can naturally cause features occlusion and influence appearance outcome assessment. Given this fact, deep learning techniques and regression studies have registered success regarding feature detection.

The approach used in this section disregards face detection because all visuals used in this study are anonymised facial images. Therefore, a deep learning-based approach has been used for detection of facial features from cropped images for the analysis and assessment of CL treatment outcome.

One study on occluded facial landmarks detection uses statistical regression analysis framework where the facial image is partitioned into nine equal portions with anticipated landmarks positions (Dollár, Welinder and Perona, 2010). It has been applied to normal facial images in several datasets such as WFLF (Sagonas et al., 2013) and COFW (Burgos-Artizzu, Perona and Dollar, 2013). A more robust approach (RCPR) introduced in (Burgos-Artizzu, Perona and Dollar, 2013) operates under difficult occlusion with the intention to improve the performance in (Dollár, Welinder and Perona, 2010).

The general assumption that occlusion is casually created using external objects such as spectacles, caps, hair styling, religious attires etc is not conclusive. This study introduces and investigates a unique case of CL images where occlusion is introduced by the surgical treatment procedure and ethical norms instead.

5.3 Approach and Implementation

In this section, a detailed step by step discussion is presented. The aim is to detect as many feature points as possible. Additionally, the objective is to classify and group the detected feature points in the three apparent segments of the facial images: upper third, middle third and bottom third, represented as the periorbital region, nose region and lips/oral region respectively (Erian and Shiffman, 2011, Hashim et al., 2017), please refer to Figure 5.1.

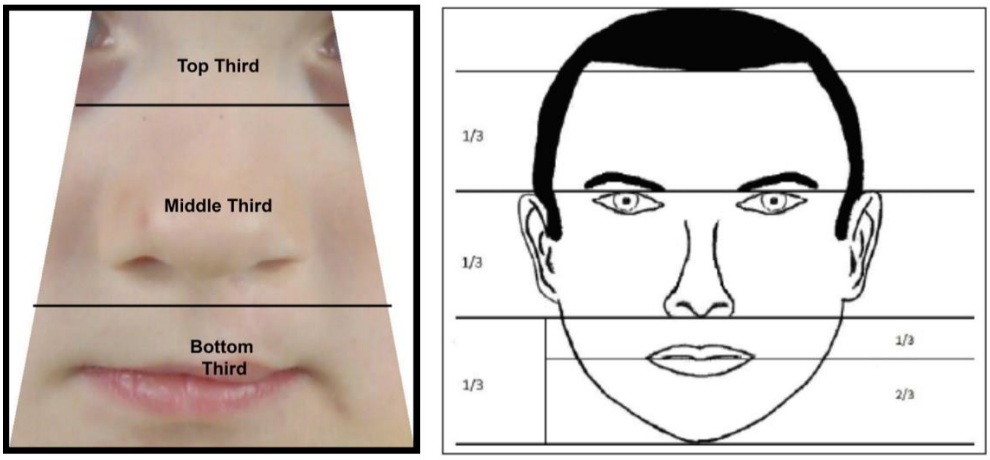

Figure 5. 1: Illustration and adaptation of horizontal thirds under occlusion of a cropped face ($left$) and full face ($right$)

Successful categorisation of these features is crucial towards the determination of the most befitting symmetric axis of the face. To this end, a deep learning-based method is used to detect the facial feature points of interest in the three regions: pre-processing, feature detection, symmetrical axis estimation and numerical score estimation. These steps are detailed below.

1. **Pre-processing**

It is fundamental that preprocessing is conducted on the different outcome facial images using an appropriate filter. Filters have an enhancement and smoothing effect to facilitate generation of better segmentation results (Frery, 2013). Several filters such as Gaussian, Laplacian of Gaussian, median and the others (Frery, 2013) can be used for this purpose. However, given the nature of our dataset's visuals, a $3\ by\ 3$ Gaussian filter was the best choice because image pixels are evenly distributed despite any image degenerative conditions with each element set to 1. Without the Gaussian filtering, less features are detected. Other filters can be designed. Examples of other filters are:

$mean\ f\ = \frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$ and a vertical Sobel filter: $f\ = \begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}$

Following filtering is segmentation such as using edge-based segmentation. This technique detects and traces boundaries or edges between different regions in an image. The Canny edge detector is a widely used method that identifies significant edges based on gradient information. Other edge-based methods, such as the Sobel

operator and Laplacian of Gaussian, are also frequently utilized (Canny, 1986). Different operations have different outcomes when used in different colour systems such as HSV or RGB, as seen in Figure 5.2.

Laplacian edges produce better results across HSV than RGB colour spaces, but struggle on grayscale images. SobelY edges are second best in performance, following our experiments. This is attributed to presence of more horizontally inclined features than vertically pronouncing features. The grayscale image from an HSV segmented image has better visibility than its RGB counterpart. This is attributed to the fact that HSV separates luma, or the image intensity, from chroma or the colour information (Hu et al., 2021). In subsequent studies, exploration of the image intensity components without their colour components will be studied. Besides, HSV is more human natural friendly space compared to RGB.

Some of the outputs from these filters are presented in Figure 5.2.

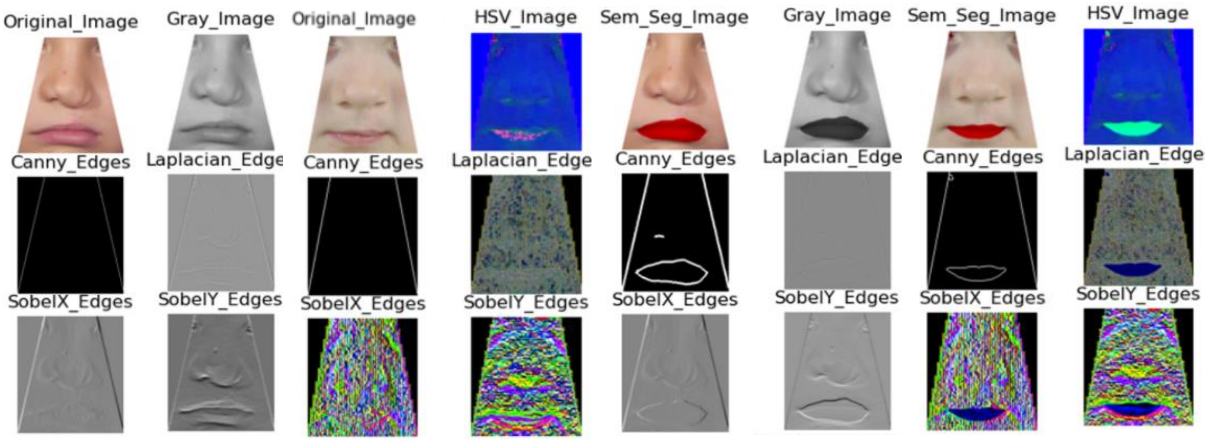

Figure 5. 2: Different filters used to visualize features for potential classification. $Left\ four$ columns show that facial features are not clearly localized. $Right\ four$ columns show clearer features from the same filters after segmentation using an ML approach.

The left four columns show that facial features are not clearly localised using any of the filters in the RGB and HSV colour spaces. The best result in the first four columns stems from applying Sobel Y filter to a grayscale image. The outcome image is not a plausible choice for any computational method. The last four columns have had semantic segmentation transformation on the input image. The mouth boundary for the images using different filters in clear from all colour spaces apart from grayscale image in column 6, row 2 with Laplacian edge filters. Therefore, the best choice of images is using the RGB with canny edge filters. It produces a clearer mouth boundary

outcome from both grayscale and RGB images. The outcome is also computationally lighter to process.

2. Recognition of Salient Regions using Semantic Segmentation

Convolutional neural networks (CNNs) have revolutionised image segmentation. Fully Convolutional Networks (FCNs), U-Net, and Mask R-CNN are popular architectures used for pixel-wise segmentation tasks. These methods achieve state-of-the-art performance by leveraging large-scale annotated datasets and end-to-end learning. For example the bilateral semantic segmentation network performs segmentation at pixel level to differentiate between skin colour patterns of the human facial features using rich spatial information and sizeable receptive fields (Yu et al., 2018). Semantic segmentation can be applied in real time in medical images diagnostics, autonomous vehicles training and traffic management (Garcia-Garcia et al., 2017).

Bilateral Segmentation Network works with two parts, namely, the Spatial Path (SP) and Context Path (CP). The SP component is devised to confront with the loss of spatial information while the CP component is designed to shrinkage of the receptive field for intensity values. The design of the two paths is such that for SP, only three convolution layers are stacked to obtain the 1/8 feature map, which retains affluent spatial details. However, for CP component, a global average pooling layer is appended on the tail of Xception network. By so doing, the receptive field becomes the maximum backbone network. This technique has mitigated previous semantic segmentation interventions that compromise spatial resolution for real-time speedy segmentation.

Additionally, the segmentation network uses the attention refinement module to refine features at every stage of processing by employing the global average pooling. This is used to capture global context by computing the required attention vector used to guide features learning. Next, the feature fusion module sums up or fuses the features from CP and SP components (Yu et al., 2020). This module is part of the network architecture. Finally, batch normalisation is applied to balance the scales of the features.

Whereas the principal loss of used to monitor the output of the whole bilateral semantic network, the SoftMax loss is used to monitor the loss from the CP component.

Using the bilateral semantic segmentation, a desired region of interest can be segmented, Figure 5.3 has some outcomes. In the three rows in Figure 5.3, the first column is the input image, the second column represents the outcome of segmenting the mouth and eye corners, however, all eye corners are missed in row 1, column 2, one eye corner is missed in row 2 column 2, while all eye corners are segmented in row 3 column 2. The third column shows that only the upper lip can be segmented while the fourth column indicates that the network can be configured to segment only the upper lip and eye corners. Missing some eye corners is attributed to non-uniform generation of the dataset by experts, during manual cropping and limited diversity in the dataset used for network training.

The key benefit of bilateral network segmentation is its capacity to consider both spatial and intensity information simultaneously. This makes it robust in preserving edges and minute details, which is essential in many image analysis tasks. The integration of neural networks further enhances its segmentation capabilities.

Figure 5. 3: Bilateral semantic network segmentation outcome: second column, segmentation of the lips and eye corners. In some images, the eye corners could not be segmented. Third column shows that only the upper lip can be segmented, or the upper lip and the eye corners as seen in the fourth column.

Despite its advantages, bilateral network segmentation can be computationally rigorous, especially when applied to large images. Fine-tuning neural networks for specific segmentation tasks may also require a substantial amount of labelled data. The segmentation network has been trained on large images (1024 by 1024) which presents a computational challenge where thousands of images are involved, hindering customisation and transfer learning efforts.

Salient regions are separated through semantic segmentation because facial images present segmentation challenges using ordinary techniques due to low contrast (Oliveira et al., 2016). Inner canthus and oral region features are most salient. Figures 5.3 and 5.4 show that not all the key regions will be detected due to poor anonymization procedures.

Figure 5. 4: Segmentation results. Mouth region properly detected in all. From Left to Right: first - right eye corner not detected, fourth - all eye corners not detected.

Further, skin colour tone and scars from surgical treatment complicates the detection of any features (Sandy, Kilpatrick and Ireland, 2012). The nose region is not segmented semantically but through any edges that may be detected. Consequently, the head orientation during photo-taking (either looking straight or downwards) influences the detection of the nose region edges and feature points following luminance contrast. Both the bilateral semantic network segmentation algorithm (Yu et al., 2020) and high-resolution network segmentation (Wang, Sun, et al., 2020) produce appropriate segmentation results. The former is a faster and less resource intensive approach. We utilize the detailed module and semantic module of the bilateral segmentation network to acquire the image's low- or high-level features and the semantics of each pixel, respectively. The two modules are combined through a

real-time fusion module. The outcome is a clearly segmented mouth region, and the eye canthus, where possible. Figure 5.3 shows the red-segmented mouth region and red-segmented eye canthi (Middle two). For the different scenarios, to be discussed in the subsequent subsection of this chapter, different features are therefore considered as inputs to the top network layers.

Segmentation therefore aids the detection of the mouth region and the inner canthi, but it usually completely missed the nose region. The nose was not considered a key CL surgical outcome, hence excluded from training the segmentation network. Besides, the baseline dataset for the bilateral network should have been annotated to incorporate the nose, (Liu et al., 2015), without the two nostrils. Eventually, the two nostrils would be considered significant to determine the symmetry of the partially occluded faces. To this end, we propose to apply Canny edge detection (Canny, 1986). This further results into more feature points with higher accuracy for detection of the mouth and eye regions, Figure 5.5, middle. Each of the regions should have feature points to aid with symmetry detection. For eyes, the interest lies with the closest inner canthus distance and the median distance while for the mouth region, the philtrum, vermillion borders, and oral commissures are desirable. Within the nose region, the tip, nostrils, and their base are of interest.

Figure 5. 5: Features identified per horizontal partition. Left: Largely disorderly without segmentation. Middle: Shows improved features mapping and detection after segmentation. Right: Classified per horizontal partition.

Figure 5.5 (left) shows that identification of feature points can be more complicated before segmentation. After segmentation, as seen in Figure 5.4 (Middle), it is easier to identify many features/ feature points in the different sections of the facial image, especially the mouth region, nose region, and eye corners. Figure 5.5 (right) indicates better feature classification by location in the facial image. For clarity, feature points in

the mouth region (bottom third) are coloured blue, features coloured green lie in the nose region (middle third) while features coloured light blue lie in the eye corners (top third).

3. Identification of Connected Components

Because of the partially occluded and anonymised nature of the dataset and other outliers like skin residues, scars etc, some feature points within the key regions are disconnected. Implying there is need to identify connected components for stability to aid the determination of the symmetry. This is aided by filtering the detected feature points using Canny edge detector with the lower and upper thresholds set as decrease and increase by 0.33 of the median of pixel intensities of the whole image. The integral features lie along the following facial parts: eye corners (or inner canthi), nose tips, nose base (or nose root), nostrils, mouth boundary, philtrum and oral commissures. Once the feature points are detected, different colours, other than red (for the predicted set, PS) and green for the ground truth set (GT1, GT2, and GT3), are assigned to the feature points per horizontal third to aid visualisation (Figure 5.5, right).

4. Linking Points as Contours

The next step is linking the respective feature points as contours. Successfully determining the feature point's perimeter suggests presence of a closed area or contour. At this stage, all possible contours have been identified as re-sampled contours from fully connected shapes without self-intersection, following the library implementation of the Douglas-Peucker algorithm (Wu, Silva and Márquez, 2004), Figure 5.4, left. Some contours may be very small but necessary for the location of the position of the features of interest. For example, a detailed execution shows that the green feature points representing identification of the nose region features (Figure 5.5, right) are better visibly displayed.

5. How Partitioning Works

Partitioning is conducted for the feature points on the detected contours into the three horizontal thirds based on their heights. Additionally, the feature points are classified into respective horizontal third segment. Figure 5.5, center and right illustrate the features' locations using distinct colours for each of the three horizontal thirds (Blue for bottom third, Green for middle third, light blue for top third).

6. Center of Mass using Contours

Determine the centres of mass of the contours in each of the three horizontal thirds. In Figure 5.6, the colours of the feature were changed to accommodate axes and the different center of mass colours. Yellow was used for upper third features, pink for middle-third features and light blue for bottom third features. Similarly, the same colours were used for plotting the respective axes passing through the center of mass of the different three thirds.

The average of the features in the top third (yellow) was computed and a relatively thick blue dot plotted in the top third as the center of mass. Eventually, a yellow vertical line was plotted through the blue dot as the potential symmetry for the top third region. A similar procedure was followed for the middle third and bottom third. That accounts for the yellow vertical line, pink vertical line, and light blue vertical line.

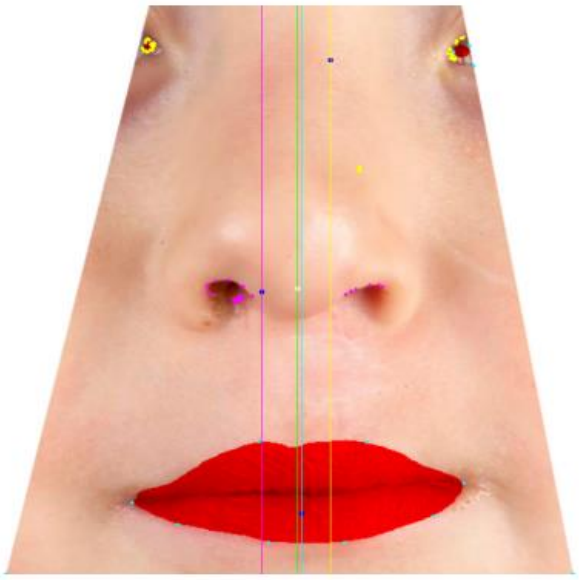

Figure 5. 6: Potential symmetric axes plotted based on component positions and their averages. Each plotted potential symmetric axis has been assigned a different colour that corresponds to that of the detected feature points in the different three third region.

Next is computation of the overall centre of mass, which is the average of the three centres of mass or the average of all the features in the different three thirds of the facial image. The thick white dot in the image is the overall center of mass. Finally, a green vertical line is plotted through the white dot.

7. Adaptive Choice of Ideal Symmetric Axis

Given four plotted potential axes of symmetry (Figure 5.6), only one axis should be considered as an optimal one. We determine the Manhattan distance of each of the feature points from each of the potential symmetric axes through Equation 11. Due to aggressive features detection, it has been experimentally proven that there are enough points from which to determine the symmetric axis.

$$dist(axis_k) = \sum_j \sum_{i=0}^{n} | axis_k - p_{ij} |, k = 1, 2, 3, 4 \quad \text{(Equation 11)}$$

Where n is the number of detected contours in the image, $axis_k$ is the potential vertical axis of symmetry, $p_{ij}$is a feature point j in the i^{th} contour. The symmetric axis is finally determined as the one with the minimum Manhattan distance. In Figure 5.7 (left), the symmetric axis was determined as the light blue line while in Figure 5.7 (right), the symmetric axis was determined as the green line. From the four potential axes of symmetry, different facial images returned different symmetric axis. Overall, there is no definite trend of a favoured symmetric axis, hence the adaptive computation.

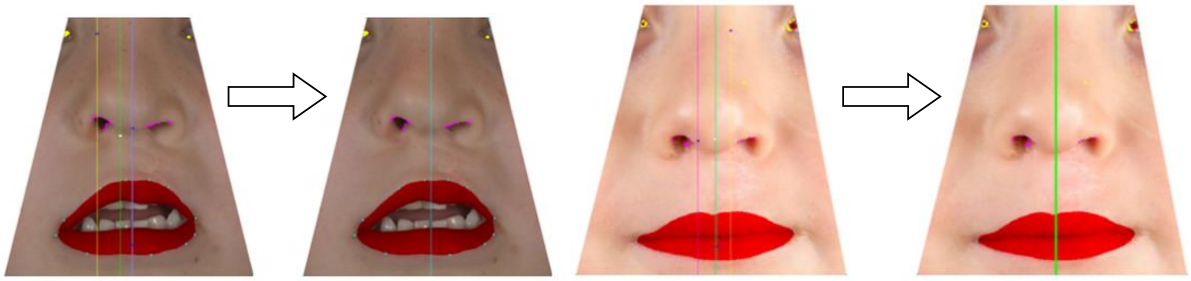

Figure 5. 7: Most suitable symmetric axis selected using average Manhattan distance. Following Figure 5.6, the most suitable symmetric axis is selected from either 2 or 3 detected axes. Green colour is assigned to the most suitable line of symmetry.

8. Chosen Symmetric Axis

The determined symmetric axis is the basis for dividing the mouth lips region into two sections to aid lip shape analysis (Bakaki et al., 2021) under different appropriate scenarios. Lip shape analysis aims to determine how evenly or unevenly shaped the lip region is on either side of the symmetric axis. Generally, some studies have applied shape analysis to describe human perception features in medical images using contrast improvement ratio (CIR) (Kimori, 2013). (Loncaric, 1998b) reviews other shape analysis techniques applied to images. In this study, a structural similarity index measure (Wang et al., 2004) is preferred to automatically determine how agreeable

the human visual perception is of the mouth lips, nose, or a combination of both. Shape analysis is reduced to a structural comparison between the two mouth lip sides using the symmetric axis as a basis.

5.3.1 Implementation

The algorithm in Figure 5.8 summarises the implementation framework for this assessment mechanism. The accompanying assumptions are made in case some images in the dataset return exceptions.

Assumptions:

1. The rule of three thirds as applied implies that the mouth region lies in the bottom third of the input facial image.
2. Due to the empirical possibility that the rule of three thirds may not apply to some images, this has been automatically mitigated in a way that sub-dividing the facial image does not happen.

1. Input an image.
2. Preprocess the image using smoothening with a 3-filter Gaussian blur.
3. Semantically segment the image if applicable, otherwise, do not.
4. Divide the image into three thirds based on the image height.
5. Using the computed median, automatically determine lower and upper thresholds before appropriate adaptation of the Canny edge detection algorithm. The output of this step should be edges and points in each three third segment.
   5.1 For a non-segmented image, the number of detected edges/features per three third component is minimal and not as orderly.
   5.2 In step 5, differentiating the features using different colour codes is highly advisable.
6. Perform morphological functions (dilation and erosion) to close gaps between object edges.
7. Select the contour with the largest area in each three third segment and determine its centroid. This is also the average point of the feature points detected in step 5.
8. Plot the perpendicular of through the centroid determined in 7 above for each of the three thirds and for the overall features in the image.
9. Where possible, differentiate the perpendicular lines using different colour codes, and should correspond to colour coding in step 5.2.
10. Adaptive shape analysis follows using structural similarity index measures using three different scenarios for an RGB image.
    10.1 Mouth and nose region
    10.2 Mouth region only
    10.3 upper lip only.
11. Convert the different similarity measures into different numeric numbers between 1 and 5 using three predefined mathematical models.
12. Generate appropriate Pearson correlation metrics.

Figure 5. 8: Implementation framework or algorithm for key landmark detection using the three thirds adaptive symmetric axis detection.

5.4 Outcomes and Discussion

Automatic shape analysis can be presented as a culmination of the human visual perception of a given object. The facial images consist of eye corners, nose base, curvatures, mouth lips boundaries, and philtrum section as the key features. Each of these features is detected and drawn as an independent shape or structure with at least a position (also called feature point). Given the nature of the dataset, the detected features are more than the expected ones due to facial speckles for example, implying that the likelihood of missing the feature point is reduced.

The distribution of the number of key feature points per horizontal third of the facial images from the public CCUK dataset is shown in Figure 5.9. The four subcategories of the public dataset are: (i) Predicted Set (PS) – a set of images obtained through the proposed algorithm above, and (ii) three expert-generated Ground Truth datasets GT1, GT2 and GT3. However, GT3 has not been considered in this study because it offers only a single feature (the mouth lip boundary). Ground truth sets are generated by manual annotation of the mouth lip region using ImageJ, an open-source software.

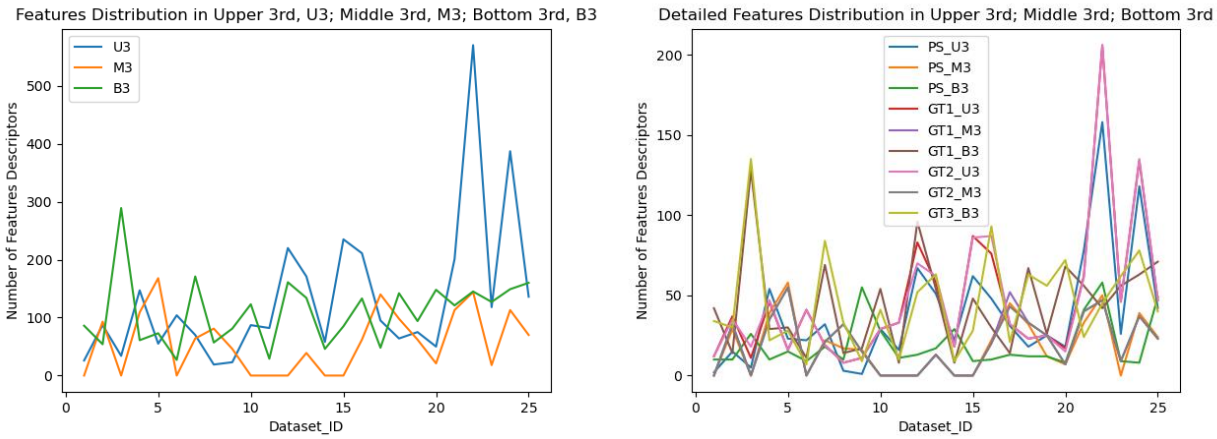

Figure 5. 9: The number of features detected across the 3 different upper, middle and bottom thirds: U3, M3 and B3 of each of the 3 considered sub-datasets.

More features are detected in the upper and bottom horizontal thirds (blue and green, seen in Figure 5.9, left), probably as expected. The middle horizontal third has fewer features because a physical dataset inspection reveals that the camera angle was not straight when taking most of the images. The other possibility is that the participants were anxious and didn't look straight during the camerawork exercise. This is mostly the case because most children with the orofacial cleft conditions have low self-esteem

among other social concerns, and may indirectly object to taking their pictures (Mosmuller, Maal, et al., 2017, Mulder et al., 2019). This observation has impacted image analysis experiments and potentially promoted discussions of multi-dimensional $3D, 4D$, image analysis techniques, seen as mitigation measures. However, resources used for multidimensional data capture are costly (Ayoub et al., 2011). Besides, historical data is only available in $2D$ format, a case of the CCUK dataset.

Additionally, some of the residues (caused by scars and running rose) potentially obscured the philtrum, oral commissure, vermillion, and nose base lining detection (orange, seen in Figure 5.10, left). Implying that features in the middle third can be discarded because they are even biologically difficult to delineate when determining the symmetric axis. Given that we can use minimal data to generate a consistent outcome, this approach is a feasible alternative to deep learning techniques that detect facial features directly from the cropped images.

Figure 5.9, right, presents a more detailed breakdown of features points distribution per dataset subcategory per horizontal third.

Shape analysis was performed through three scenarios to assess the proportionality of the appearance outcome following surgical repair.

Scenario 1: Mouth region only (Figure 5.10, left). The physical surgical repair to the cleft on the upper lip is usually taken, in consideration of its alignment with the lower lip. Hence, considering the whole mouth region is a natural occurrence when performing appearance outcome assessment. Similarity on either side of the symmetric axis through the mouth region is expected for features such as commissures and philtrum.

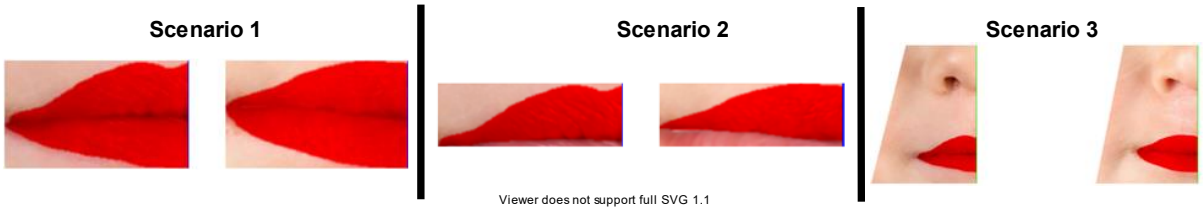

Figure 5. 10: Visualization of mouth region in Scenario 1 (left), upper lip in Scenario 2 (middle) and both nose and mouth regions in Scenario 3 (right)

Scenario 2: The Upper lip, (Figure 5.10, middle) is the actual region of the mouth that is surgically repaired. Therefore, investigating its structural feature (dis)similarity is a trivial and fair choice.

Scenario 3: Combination of the nose and mouth lips region (Figure 5.10, right). Whereas cleft surgical repair is usually performed on the upper lip, human appearance outcome assessment naturally occurs with the awareness of other neighbouring features (Deall et al., 2016a). The closest feature available and applicable to our dataset is the nose region. Observing any (mal)alignment between the nose and mouth is almost trivial.

These scenarios facilitate shape and structural computation and comparison using colour images, as presented before human assessors (Mosmuller et al., 2013). This also implies that the efficacy of this approach can be determined by comparing the shape and structural computation with the human-generated numeric score (HNS) by human assessors. In (Bakaki et al., 2021), binary images for the mouth lip region were used. After computing the structural similarity index measure, s, it is converted to a numeric score between 1 and 5 as discussed below.

5.4.1 Mathematical Modelling

Mathematical modelling is the process of applying mathematical techniques and concepts to describe and understand real-world events, systems, or processes as they occur during and/or following an experiment (Zorich and Paniagua, 2016). Conversion of the different structural similarities' measures, s, into numeric values represents a complex situation where carefully constructed mathematical equations with symbols and relationships is desired. Therefore, this is done with the intention of gaining quantitative insights through numeric predictions and analysis (Leao et al., 2020). Different forms of mathematical model representations were considered.

1. Linear Models: These describe relationships between variables using linear equations. In a linear model, the relationship between the variables is assumed to be proportionate and additive, implying that a change in one variable is directly proportional to a change in another variable, and the overall effect is considered cumulative (Pinheiro et al., 2007).
2. Nonlinear Models: They describe relationships among variables using nonlinear equations. More complex and flexible relationships between variables

may be designed. Curves (such as polynomials) and exponential functions better represent nonlinearity (Pinheiro et al., 2007).

3. Fractional Models: They are a class of mathematical models that involve fractional derivatives of non-integer orders. This is suitable where a phenomenon exhibits big memory requirements and long-range dependences.

Therefore, the three (3) models below were designed to convert s into a numeric score:

Model 1 $(M1)$: $f(s) = 5(1 - s^2) + s$ (Equation 12)

Model 2 $(M2)$: $f(s) = exp((1 - s)ln5)$ (Equation 13)

Model 3 $(M3)$: $f(s) = 5 - \frac{4s}{(1+s)^{\frac{1}{100}}}$ (Equation 14)

Where $0 \leq s \leq 1$ and $1 \leq f(s) \leq 5$. Implying that the models designed are monotonically decreasing, hence non-increasing.

As derived from the example in chapter 4, we make some assumptions:

- That the relationship between $f(s) or\ AENS$ and s is either quadratic or exponential or fractional. Together, they could be regarded as polynomial functions.
- Additionally, that the mathematical operations involved (addition, multiplication, and exponentiation) are valid for the domain of interest.
- s represents a value within a certain range (between 0 and 1)

Derivation:

- We combine terms involving s, including a linear term (s) and a quadratic term ($1 - s^2$).
- The coefficients (5 in this case) have been chosen because 5 can be the worst score, which could indicate the failure to detect certain features for use in computation.

The process is repeated for equations 13 and 14.

The models designed in chapter 4 are similar to the models designed in this chapter because the structural similarity index measure component (s) is the foundation parameter (also potentially referred to as independent variable) for conversion into a

number between 1 and 5. The models in chapter 4 have not been used because of the need to understand and reflect upon the robustness of mathematical modelling, evaluate their performance using different mathematical models, and discover if any benchmark models would be achieved in the process (Banerjee, 2021).

Therefore, s is computed for individual treated images in each of the three subsets of the main dataset (PS, G1, GT2) and then automatically converted into their respective numeric scores (PS_AENS, GT1_AENS, GT2_AENS), from which correlation coefficients were calculated against HNS. The higher the coefficient, the more accurate the automatically estimated appearance numeric score of the CL treatment outcome.

For every proposed model, AENS is computed for every possible dataset and their correlation coefficients computed about HNS. For example, M1, M2 and M3 are used to compute the AENSs for PS (that is columns 'PS_AENS (M1)', 'PS_AENS (M2)' and 'PS_AENS (M3)'), respectively, under the consideration of S1, whose results are shown in Table 5.1.

Table 5. 1: Results for PS_AENS over Scenario 1 and the three models

Set_Code	Set_ID	SSIM	SSIM_NORM	M1	M2	M3
AOBA	1	0.98451053	1.000	1.0	1.0	1.0
AOFA	2	0.89743804	0.395	4.6	2.6	3.4
APAA	3	0.924125194	0.580	3.9	2.0	2.7
AQCA	4	0.972945528	0.920	1.7	1.1	1.3
CLDA	5	0.906739756	0.459	4.4	2.4	3.2
CNCA	6	0.945251625	0.727	3.1	1.6	2.1
CPCA	7	0.930614227	0.625	3.7	1.8	2.5
DMAA	8	0.944182274	0.720	3.1	1.6	2.1
DMCA	9	0.93949915	0.687	3.3	1.7	2.3
DNBA	10	0.94906182	0.754	2.9	1.5	2.0
DPAA	11	0.949288457	0.755	2.9	1.5	2.0
DPBA	12	0.911902762	0.495	4.3	2.3	3.0
EOAA	13	0.888051776	0.329	4.8	2.9	3.7
EPFA	14	0.925933149	0.593	3.8	1.9	2.6
FMBA	15	0.885880415	0.314	4.8	3.0	3.7
FNEA	16	0.914207339	0.511	4.2	2.2	3.0
FNFA	17	0.87926085	0.268	4.9	3.2	3.9
FNGA	18	0.860962537	0.141	5.0	4.0	4.4
FOEA	19	0.846640729	0.042	5.0	4.7	4.8
FPAA	20	0.840660856	0.000	5.0	5.0	5.0
FPBA	21	0.922898051	0.572	3.9	2.0	2.7
FPHA	22	0.958874709	0.822	2.4	1.3	1.7

FPIA	23	0.917825731	0.536	4.1	2.1	2.9
GLAA	24	0.931627171	0.632	3.6	1.8	2.5
GLCA	25	0.90167591	0.424	4.5	2.5	3.3

For each visual element in the dataset, SSIM is computed and linearly normalised using the maximum and minimum approach before acting as the input parameter for the different mathematical models. Normalisation is helpful because choosing a common scale and representation strategy is a recommendable practice for comparison fairness, uniform data distribution and decisions modelling (Vafaei, Ribeiro and Camarinha-Matos, 2018).

Following Table 5.1, two other tabular sets of numeric data are generated, with PS as the input: scenario 2 (S2) – Table 5.2, and scenario 3 (S3) – Table 5.3, respectively.

Table 5. 2: Results for PS_AENS over Scenario 2 and the three models

Set_Code	SSIM	SSIM_NORM	PS_AENS (M1)	PS_AENS (M2)	PS_AENS (M3)	HNS
AOBA	0.794286932	0.880	2.0	1.2	1.5	2
AOFA	0.785675688	0.865	2.1	1.2	1.6	4
APAA	0.579278206	0.509	4.2	2.2	3.0	2
AQCA	0.814685762	0.915	1.7	1.1	1.4	1
CLDA	0.284428452	0.000	5.0	5.0	5.0	2
CNCA	0.667275392	0.661	3.5	1.7	2.4	3
CPCA	0.7796272	0.855	2.2	1.3	1.6	3
DMAA	0.715846417	0.745	3.0	1.5	2.0	4
DMCA	0.657024383	0.643	3.6	1.8	2.4	4
DNBA	0.656182293	0.642	3.6	1.8	2.4	3
DPAA	0.717351912	0.747	3.0	1.5	2.0	4
DPBA	0.757496715	0.816	2.5	1.3	1.8	2
EOAA	0.738977307	0.784	2.7	1.4	1.9	5
EPFA	0.414502384	0.224	5.0	3.5	4.1	4
FMBA	0.48136554	0.340	4.8	2.9	3.6	3
FNEA	0.707645437	0.730	3.1	1.5	2.1	1
FNFA	0.826107278	0.935	1.6	1.1	1.3	2
FNGA	0.738191923	0.783	2.7	1.4	1.9	2
FOEA	0.592829848	0.532	4.1	2.1	2.9	5
FPAA	0.299107524	0.025	5.0	4.8	4.9	3
FPBA	0.863847491	1.000	1.0	1.0	1.0	1
FPHA	0.795470587	0.882	2.0	1.2	1.5	2
FPIA	0.430875445	0.253	4.9	3.3	4.0	3
GLAA	0.62022925	0.580	3.9	2.0	2.7	5
GLCA	0.67455779	0.673	3.4	1.7	2.3	4

Table 5. 3: Results for PS_AENS over Scenario 3 and the three models

Set_Code	SSIM	SSIM_NORM	PS_AENS (M1)	PS_AENS (M2)	PS_AENS (M3)	HNS
AOBA	0.99415334	1.000	1.0	1.0	1.0	2
AOFA	0.964996816	0.459	4.4	2.4	3.2	4
APAA	0.969440482	0.541	4.1	2.1	2.8	2
AQCA	0.989788824	0.919	1.7	1.1	1.3	1
CLDA	0.963005764	0.422	4.5	2.5	3.3	2
CNCA	0.979077131	0.720	3.1	1.6	2.1	3
CPCA	0.975327187	0.651	3.5	1.8	2.4	3
DMAA	0.980553343	0.748	3.0	1.5	2.0	4
DMCA	0.977561029	0.692	3.3	1.6	2.2	4
DNBA	0.981373941	0.763	2.9	1.5	2.0	3
DPAA	0.981962076	0.774	2.8	1.4	1.9	4
DPBA	0.968507692	0.524	4.2	2.2	2.9	2
EOAA	0.961668402	0.397	4.6	2.6	3.4	5
EPFA	0.971764181	0.585	3.9	2.0	2.7	4
FMBA	0.958799876	0.344	4.8	2.9	3.6	3
FNEA	0.968365544	0.521	4.2	2.2	2.9	1
FNFA	0.952739379	0.231	5.0	3.4	4.1	2
FNGA	0.951614992	0.211	5.0	3.6	4.2	2
FOEA	0.945609351	0.099	5.0	4.3	4.6	5
FPAA	0.940266349	0.000	5.0	5.0	5.0	3
FPBA	0.971782913	0.585	3.9	2.0	2.7	1
FPHA	0.985045057	0.831	2.4	1.3	1.7	2
FPIA	0.967890825	0.513	4.2	2.2	3.0	3
GLAA	0.974788298	0.641	3.6	1.8	2.5	5
GLCA	0.964537142	0.450	4.4	2.4	3.2	4

Since the PS dataset is determined automatically, its respective numeric scores PS_AENS are also automatically obtained. The correlation coefficient between HNS and PS_AENS is considered the most significant correlation (MSC). This is because the potential agreement between human-generated and computer-generated numeric scores is the most important relationship for use in evaluation of computational methods as seen in Table 5.4.

The outcomes in the three sub-columns (‘S1’, ‘S2’, ‘S3’) of column ‘PS_AENS vs HNS’ in Table 5.4 are generated from computations of correlation coefficients from the columns in Tables 5.1, 5.2 and 5.3, respectively. Specifically, the result of 0.236 is the correlation coefficient of scores in the column PS_AENS (appearance assessment scores of PS computed using model M1) and column HNS (scores of human appearance assessment). Similarly, other results are generated.

The highest MSC is 0.371 resulting from computing the correlation coefficient between PS_AENS and HNS (M1, S2). Overall, scenario 1 also presents the consistent MSC results.

Table 5. 4: Correlation coefficients between the HNS and AENS for PS, GT1, and GT2 for different scenarios (S1, S2 and S3) and models (M1, M2 and M3).

	PS_AENS vs HNS			$GT1\_AENS$ vs HNS			$GT2\_AENS$ vs HNS		
	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
$M1$	0.236	0.371	0.205	0.079	0.062	0.039	0.055	0.181	0.024
$M2$	0.200	0.102	0.151	0.007	0.072	-0.033	-0.056	0.567	-0.102
$M3$	0.219	0.213	0.176	0.041	0.052	-0.003	-0.011	0.457	0.056

To generate the results in columns 'GT1_AENS vs HNS' and 'GT2_AENS vs HNS' of Table 5.4, six sets of results are needed. Three sets of results for GT1_AENS, each for S1, S2 and S3 are computed. Likewise, another three sets of results for GT2_AENS, each for S1, S2 and S3 are calculated. The aim is to obtain different combinations of correlations and evaluate the best possible predictions compared to human experts' generated appearance scores (HNS).

The best result in Table 5.4 is 0.567. This implies that the strongest agreement is between the human expert appearance assessment (HNS) and the second ground truth set appearance assessment numeric score (GT2_AENS) using the second model, M2 and the second scenario 2, S2. This agrees with some assertions that the smaller the region of interest the easier and better it is to perform appearance assessment and extract features. This is attributed to focus on local perceptions in the image (Liu, 2018). The worst result in Table 5.4 is -0.102. Likewise, the least agreement is between the human expert appearance assessment (HNS) and the second ground truth set appearance assessment numeric score (GT2_AENS) using the second model, M2 but with the third scenario, S3. This potentially indicates the fact that assessment should be performed holistically. This is because there is a high possibility of missing out on vital appearance assessment features in a smaller region of interest.

At this stage, nine (9) sets of tabular numeric results are used to compute the correlation coefficients, even though only three tables (Table 5.1, 5.2, 5.3) have been explicitly presented above. Using different numeric scores, interesting combinations are made, and other correlation coefficients are computed, as presented in Table 5.5.

Table 5.5 presents the highest correlation result of 0.940 between GT1_AENS and GT2_AENS, Model 1, scenarios 1 and 3. This is also a significant outcome, implying potential higher similarity in the human experts' datasets (GT1 and GT2).

Table 5. 5: Correlation coefficients between different AENS combinations (PS and GT1; PS and GT2 and GT1 and GT2) for different scenarios ($S1, S2$ and $S3$) and different models ($M1, M2$ and $M3$).

	PS_AENS vs $GT1\_AENS$			PS_AENS vs $GT2\_AENS$			$GT1\_AENS$ vs $GT2\_AENS$		
	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
$M1$	0.809	0.560	0.811	0.854	0.653	0.850	0.903	0.888	0.906
$M2$	0.827	0.356	0.825	0.834	0.560	0.842	0.940	0.910	0.940
$M3$	0.836	0.654	0.833	0.856	0.456	0.856	0.924	0.908	0.928

The predicted dataset (PS) appearance numeric assessment (PS_AENS) 'agrees more' with the second expert generated dataset assessment (GT2_AENS) than with GT1_AENS, returning higher correlation coefficients of 0.856 in two scenarios, S1 and S3. Across Table 5.4 and 5.5, the first model (M1) and the third model (M3) are more robust than the second model (M2).

Figure 5.11 (left, orange) presents a comparative study between the method in this chapter (referred to as approach $2, A2$) and an existing method (referred to as approach $1, A1$) (Bakaki et al., 2021). It shows that $A1$ usually generates lower structural similarity across the dataset than the former. This is because the former uses more feature points and thus generates more accurate symmetric axes.

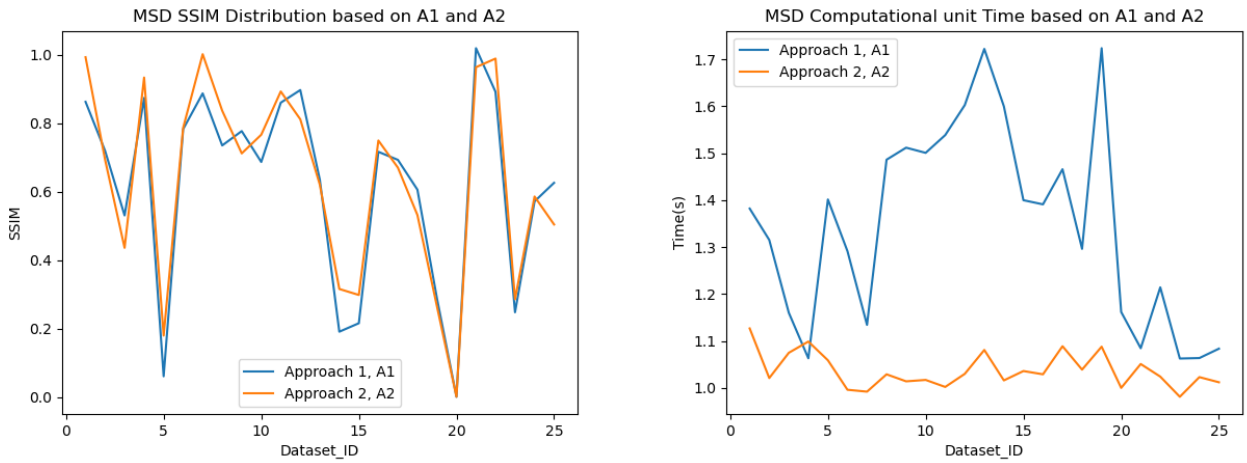

Figure 5. 11: Left: Computed SSIM for each of the 25 images in the test dataset using A1 and A2. Right: Computation time is calculated for each of the 25 images in the test dataset using A1 and A2. Computation time is the duration between input stage and assessment.

Figure 5.11 (right, orange) also shows that $A2$ takes shorter time than $A1$. It is an indicator that whole face feature detection is faster than partial feature detection. It is easier to perceive and assess a whole face than its portion.

Figure 5.12 compares feature detection using Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 2004) and $A2$, by showing how $A1$ and $A2$ map symmetric axes.

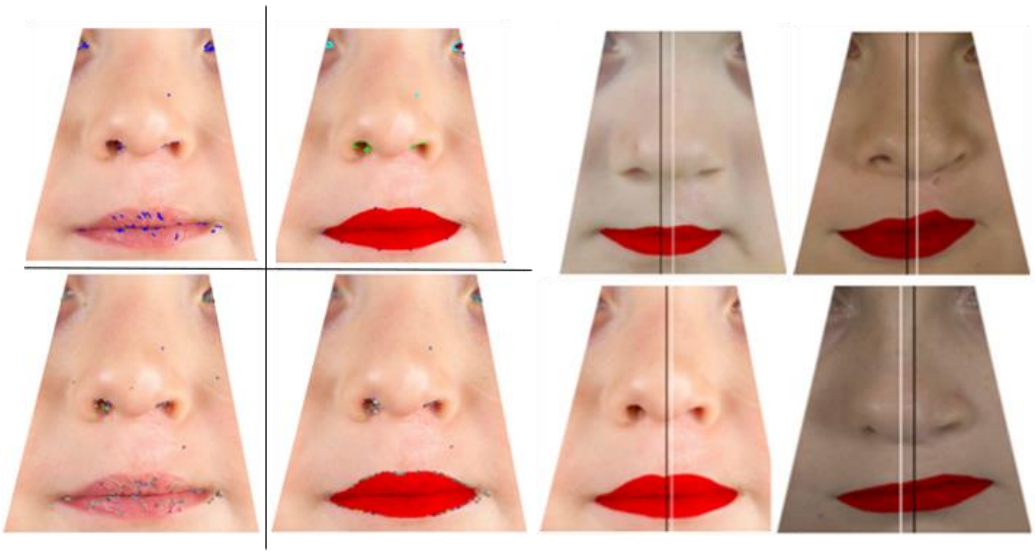

Figure 5. 12: Features (left two columns) detected using SIFT (Left column) and proposed approach (Second column). Then, symmetrical axis detection from some examples of cleft images (right two columns) using approach 1 (A1 - Black axis by (Bakaki et al. 2021)) and approach 2 (A2 - White axis by the proposed method)

The Figure 5.12 shows that features are detected using SIFT (1st column from left) and the approach presented in this chapter (2nd column from left). Additionally, Figure 5.12 presents symmetric axis detection from some examples of cleft images, seen in last two columns. In the 3rd and 4th columns, the black vertical axis is generated by approach A1 (Bakaki et al., 2021) while the white vertical axis is generated by the approach presented in this chapter (approach 2 – A2).

More visual outcomes of interest are presented in Figures 5.13 and 5.14.

Selecting the predicted set (PS) or expert generated images in GT1 and GT2 from the dataset maintains the assertion that the bottom third of the outcome image contains the region of interest, given the number of features depicted in the graphs in Figure 5.13. Additionally, Figure 5.13 also shows some features at pixel level. This is further reflected in Figure 5.14.

In Table 5.6, the higher the MED and AVG the better while the lower the SDD the better. Therefore, the best scenario across all datasets is S3, followed by S1 and S2. This is deduced from the median results (MED), mean results (AVG) and standard deviation results (SDD), presented in Table 5.6. However, shorter computational time is observed in scenario 1 and 2 compared to the computational time of scenario 3. This is because of the larger region of interest, hence potentially more features are identified and takes longer for the models to process the subsequent parameters. More details are in Table 5.7 and the graphs of Figure 5.14.

Table 5. 6: Some SSIM statistical measures across the three datasets and three scenarios

	PS			GT1			GT2		
	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
MED	0.923	0.708	0.969	0.922	0.706	0.968	0.907	0.671	0.965
AVG	0.917	0.656	0.970	0.917	0.667	0.969	0.915	0.659	0.968
SDD	0.036	0.157	0.013	0.038	0.147	0.014	0.036	0.139	0.014

Table 5. 7: Some Time spent statistical measures across the three datasets and three scenarios.

	PS			GT1			GT2		
	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$	$S1$	$S2$	$S3$
MED	1.029	1.028	1.217	1.087	1.175	1.271	1.093	1.044	1.201
AVG	1.037	1.042	1.241	1.085	1.119	1.266	1.120	1.132	1.215
SDD	0.037	0.057	0.109	0.040	0.158	0.082	0.103	0.201	0.089

Table 5.7 reveals that the first scenario, S1, is the most effective, followed by the second scenario, S2 and the third scenario, S3. In this case, all the considered statistical parameters (MED, AVG and SDD) should be lowest for optimality consideration.

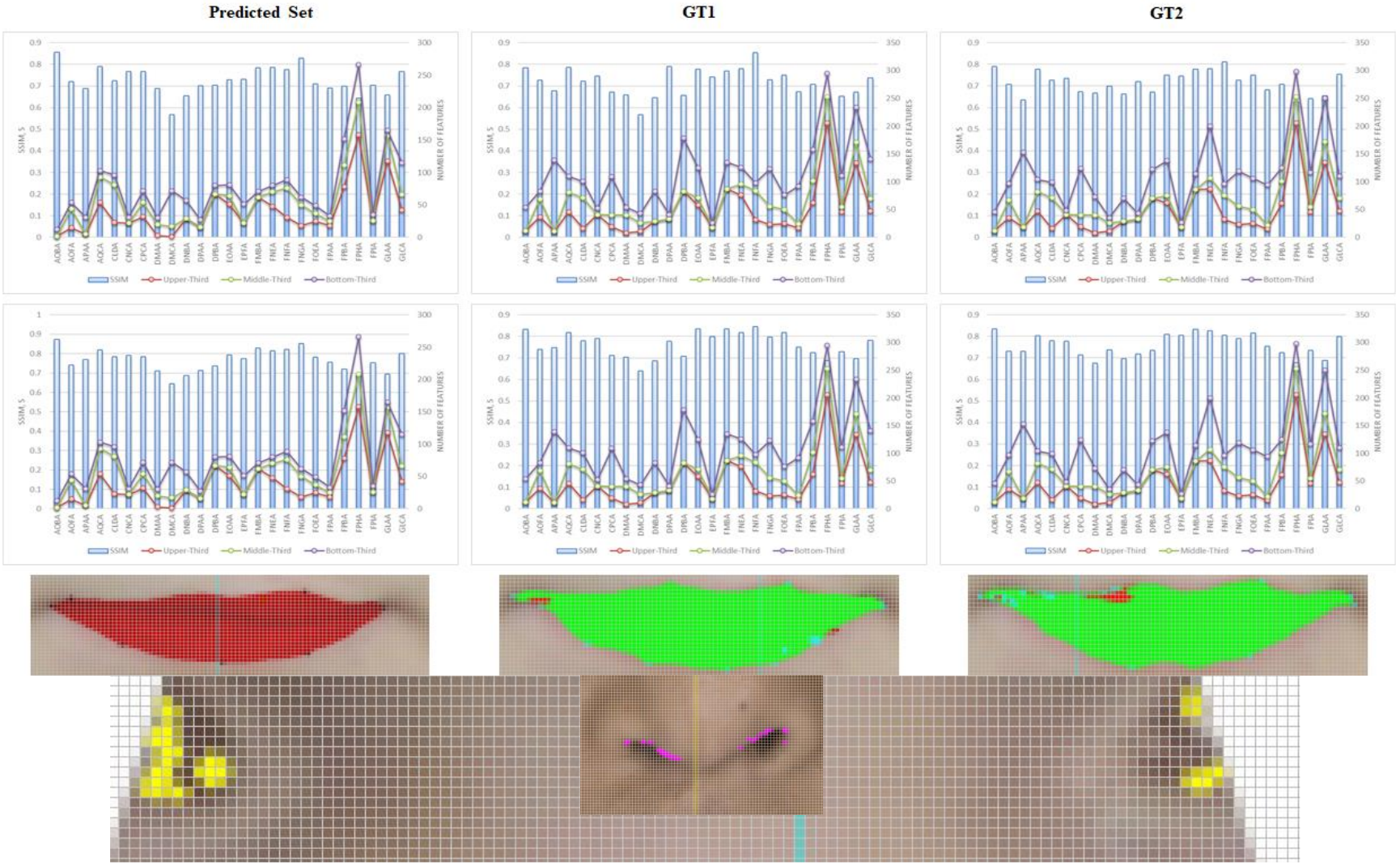

Figure 5. 13: Features Map comparison at Pixel Level between PS, $GT1$ and $GT2$. Pixels of Features are compared in the different datasets, PS, GT1 and GT2. The indicated features are detailed for those generated in the different three thirds (3-line graphs) of the individual images in the dataset. SSIM is also plotted as a bar graph.

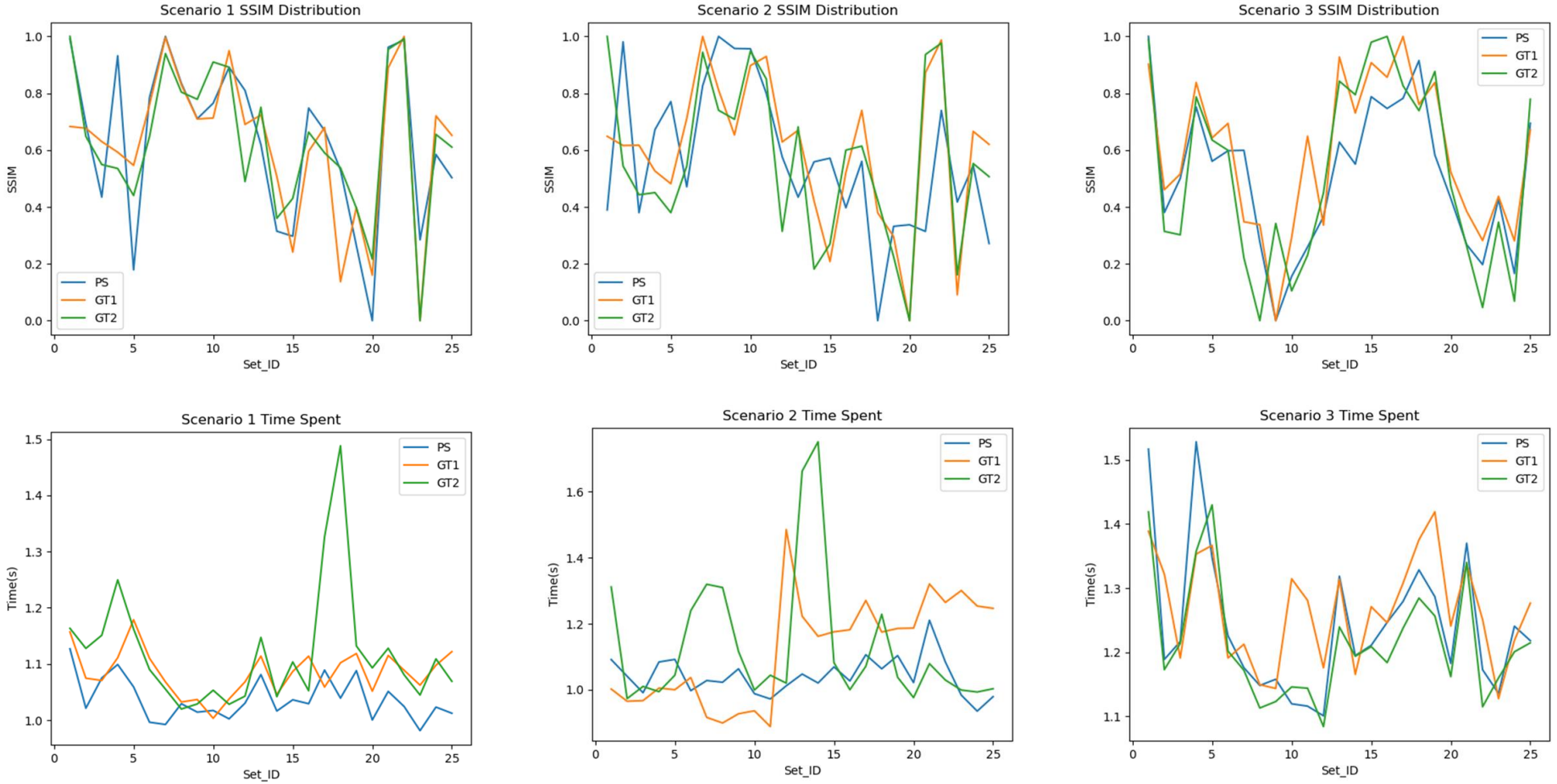

Figure 5. 14: Selected SSIM distribution ($top\ row$) and computation time ($bottom\ row$) for the 3 scenarios. SSIM and computational time are as defined in Figure 5.11, but for the different scenarios. Additionally, SSIM and computational time are further computed on the different datasets as detailed in Figure 5.13.

5.5 Summary

Detecting key feature positions requires a hybrid approach that combines deep learning and traditional approaches. For instance, a deep learning approach for segmentation combined with traditional edge detection, led to detection of more features (nose region especially) and feature points within the various regions of interest. Automatic structural comparison and analysis of colour appearances outcomes is more in harmony with human visual perception and judgement due to inclusion of luminosity and contrast features. This is represented by the consistent MSC results across the scenarios for the three models, M1, M2 and M3. Finally, anonymised and occluded facial images in our dataset have more features in the upper and lower horizontal third segments, implying that they may provide more potential for the estimation of the appearances from the cleft images. Other shape analysis techniques applicable to biomedical images as reviewed in (Loncaric, 1998a) could be tested in future studies. The next step will be investigating purely deep learning techniques to extract, detect and, where necessary predict, facial image features specific to the public CCUK dataset.

Key landmarks detection for anonymised facial images was experimentally performed using a deep learning approach as presented in Appendix A. The You Only Look Once (YOLO) framework was successfully used and tested following an ablation study, also presented in Appendix A.

Chapter 6 Regression Analysis and Assessment of Partial Facial Images Using Deep Learning

6.1 Introduction

One of the most challenging tasks in medical computer vision is to mimic human experts in performing expert roles such as calibration, surgical operations, and medical imaging, among others (Scheirer et al., 2014). Cleft lip (CL) among other craniofacial congenital malformations is surgically treated based on well-defined protocols established by the Plastic and Reconstructive Unit of the Royal College of Surgeons of England (Shaw et al., 1996). This follows a thorough assessment of the patient to determine their suitability for the operation and availability of post-operative care. The latter plays a significant role in the recovery process of the patient. The post-operative care team is diverse and includes specialists such as orthodontists, surgeons, nurses and social workers like phycologists and guardians (Sell et al., 2001).

Assessment of surgical treatment outcome is tracked based on the satisfaction of the care teams and facial anthropometry. Human digital scoring based on a photographic database is one way to measure the level of satisfaction, normally using a continuous discrete number in the range of 1 to 5 where 1 is the best outcome while 5 is the worst outcome (Schwirtz et al., 2018). This can also be done based on any other predefined Likert Scale such as 1 to 10, or even as conservative as 0 or 1. In the event of a bigger dataset of facial images from several cleft treatment centres, human digital scoring may suffer from human weaknesses, hence the need to be supported by a computational approach.

Scoring of photographic datasets has been used in several applications in past research studies. Beauty and attractiveness have been widely researched where datasets have been created and studied in the wild and through social media sites (Gan et al., 2014; Xie et al., 2015; Lebedeva, Guo and Ying, 2022). The common subject in all these studies is to promote a recommender approach for service consumption purposes, for example, in social dating sites, in tourism/hospitality industry, real estate and housing schemes among others. Researchers focus on how

beautiful or attractive the images in the datasets are rated by the participants (Li, Huang and Christianson, 2016; Poursaeed, Matera and Belongie, 2018).

Rating of medical-related datasets to determine a treatment outcome or deduce a diagnostic outcome has been rarely and vaguely conducted, using computer vision. Besides, the above-identified studies did not regulate any computational restrictions with dataset features, especially where face images were used. Appreciation of beauty or attractiveness is considered primary where images are ideal and from the wild. However, this is contrary to the aims of corrective surgery for CL. The facial visuals used in this study are presented as outcomes following corrective CL surgical treatment, not beauty or superficial aesthetic correctness. This is a key uniqueness of the dataset used in this study.

Therefore, using geometric facial features from occluded images can be significant towards development of a computational scoring solution. Using convolutional theory for key features discovery/detection from different images can be fundamental for formulation of a scoring model. Realistically, this means modelling the several key features to regress into a single continuous outcome, the appearance score. A scene of geometry is mathematically defined by a set of detected key features, considered central to following any successful CL corrective surgical treatment. Subsequently, any computer vision solution would be challenged with detection and may be identification of key features to facilitate appearance outcome scoring and rating. The key features, also known as outcome objects, can be used to determine other metrics, both continuous and categorical in nature. In image quality studies such as (Narwaria and Lin, 2010), feature extraction and representations are critical to successful scoring. Committing human visual intuition with scoring of vast datasets of medical images is prone to mistakes because image features memorability is worryingly low over a long time (Jing et al., 2019). Hence, computer vision can be a crucial intervention towards creation of sustainable solutions for scoring of large datasets.

6.1.1 Background and Context

Key feature extraction in medical image datasets is an object detection and recognition computer vision task leading to image understanding and pattern recognition. Object detection is challenged with localization and classification of specific objects in a given image while object recognition is responsible for identification and perception (Singh,

2019). These computer vision tasks are essential and have been widely applied to different scenarios in ground-breaking applications in medical science as reviewed in (Ker et al., 2017) and (Drahansky et al., 2016). Studies such as skin cancer detection/classification, automated pancreas segmentation and brain tumour segmentation among others have demonstrated considerable progress in computer vision applications in medicine. Two decades ago, successful systems such as the Viola Jones detector (Viola and Jones, 2001), which pioneered facial recognition tasks were based on handmade features by using sliding windows to search across the whole image. To mitigate some of the challenges it faced, especially facial features' localisation, the Histogram of Oriented Gradients (HoG) detector method was proposed by using gradients and scale invariant features to detect faces in oriented images (Albiol et al., 2008). Oriented images are a common site even in a controlled environment. This research uses cropped partial facial images, sometimes lacking rightful luminosity and orientation.

Features' extraction, whether in controlled environments or in the wild, is facilitated by presence of some datasets to allow researchers to conduct experiments. Most datasets that enable facial features exploitation research studies such as facial recognition, beauty and security related tasks include full images/visuals. This facilitates researchers to exploit as many features as possible. This research utilises a particular anonymised dataset of images from the Cleft Care UK (CCUK). To reemphasise the integrity of this research study, for ethical reasons, patient identity is protected and anonymised, leaving mainly the inner eye corners section and the nasolabial region to aid this study. Table 6.1 summarises the related studies.

Table 6. 1: Previous studies categories that inspired the study of appearances assessment with regression analysis assessment using deep learning models.

Approach	Research Team	Category
Artificial Intelligence Tools review for	(Rokhshad, Keyhan and Yousefi, 2023)	Semi-automatic and automatic methods reviewed for feasibility
Objective hypernasality measure using deep learning	(Mathad et al., 2021)	Automatic method for acoustics of patient with CL condition
CNN model for automatic detection and measurement of facial landmarks to assign severity grades	(McCullough et al., 2021)	Automatic method that assigns severity grades with no assessment of the outcome

Shape Analysis and Similarity Measure	(Bakaki et al., 2021)	Automatic assessment though based on essential features
Hybrid deep learning-based approach for detection of landmarks in the different segments of the facial image	(Bakaki et al., 2022)	Automatic assessment method but feature detection presents weaknesses
Deeper understanding and location of key surgical incision markers	(Li, Cheng, et al., 2019)	Automatic approach for facilitation of a robust solution for surgical treatments advisory of cleft lip and palate to maximise chances of a better appearance outcome following surgery
Human expert manipulation of SymNose	(Pigott and Pigott, 2010)	Semi-automatic approach that helps with annotation of outcomes for identification of regions of interest and the resulting symmetry for assessment preparation studies
Analyse It Doc (A.I.D) to facilitate quantitative assessment based on digital semiautomatic photogrammetry	(Pietruski, Majak, Debski, et al., 2017, Pietruski, Majak, Pawlowska, et al., 2017)	Semi-automatic method that generates a catalogue of anthropometric numeric values for an individual appearance outcome that may be subject to expert misinterpretation
Human qualitative assessment, sometimes based on different Likert scales	(Schwirtz et al., 2018)(Mosmuller, Mennes, et al., 2017a)(Lee et al., 2019)(Al-Ghatam, Jones, Ireland, Atack, Chawla, Deacon, Albery, Cobb, Cadogan, Leary, Waylen, et al., 2015)	Traditional approaches that are inconsistent, verbose, mostly give irreproducible results and prone to human fatigue.

6.1.2 Context of the Challenge

Most computer vision tasks aim at reproduction of human visual recognition capability for several tasks because human experts were previously considered better than machines for specific tasks (Scheirer et al., 2014). With large medical datasets, the trend is steadily changing. Utilisation of artificial intelligence (AI) methods is more prevalent and preferred for visual data analysis. Hence, AI methods have slowly become a trend of extensive datasets analysis with exciting techniques (Greenspan, Van Ginneken and Summers, 2016, Litjens et al., 2017, Fourcade and Khonsari, 2019). To assess the surgical outcome from facial images, it would require sustained

substantial awareness, brain time, closer features recognition and visual analysis expertise by human subjects to judge the repair. Scoring is not only aimed at improving surgical repair practices but also ensuring that different patients get the desired social acceptance following the treatment (Sharma et al., 2012, Schindler et al., 2017). To aid with advanced and robust objective assessment and scoring, the cropped partial facial images can be studied and analysed for deeper features mainly in the mouth and nose regions (the nasolabial section).

This study proposes an end-to-end deep convolutional transfer learning-based assessment pipeline. Transfer learning (TL) can be understood as the adaptation of a model (potentially the weights and/or architectural settings) trained for one domain into a similar domain (Bengio, 2011). Several open-source state-of-the-art models have been trained on diverse and rich visual datasets and are available for use by researchers and scientists. TL is a proven and feasible technique for good results especially in the event the destination domain has a limited dataset. Four deep transfer learning categories and strategies are described in (Tan et al., 2018). A summary of the different deep transfer learning categories is presented in Table 6.2.

Table 6. 2: A brief of some categories of deep transfer learning (DTL)

Category of DTL	Category Brief
Adversarial DTL	There are transferable representations that is relevant to both the source domain and the target domain. the source domain and target domain.
Network-based DTL	This denotes the reuse of the partial network that pre-trained in the source domain, including its network structure and connection parameters, transfer it to be a part of DNN which is used in target domain
Instances-based DTL	Utilize instances in source domain by appropriate weight. This is to the strategy is to select partial instances from the source domain as supplements to the training set in the target domain by assigning suitable weight values to these selected instances.
Mapping-based DTL	This refers to representing instances from the source domain and target domain into a new data space. That is to say that instances from two domains are related and suitable for a union deep neural network

This chapter presents the adversarial-based and network-based deep transfer learning strategies. This is because the visuals used in some classes in ImageNet (the parent domain) are suitable to the research problem (child domain). Additionally, the base network structure is as well used to fine tune the new network structure.

6.2 Approach to Regression Analysis

6.2.1 Overview

Following surgical treatment, it is important to score appearance outcomes. The goal is to determine a continuous/real number score of the partial facial image, ranging from 1 to 5. A score of 1 represents excellent outcome while 5 represents very poor outcome. It is conventional to read or input an image into the model before pre-processing, feature extraction and finally prediction of its numeric score.

Deep Learning modelling, a subset of machine learning, is used. In deep learning modelling, deep neural networks consist of three general categories of layers: input layer, hidden layer(s), and output layer (Fourcade and Khonsari, 2019, Roberts and Yaida, 2021). The flow chart in Figure 6.1 is an abstract representation of the working framework for the proposed (regression) model in this study. In Figure 6.1, there is a black box at a high level which simply accepts input facials and their respective raters' scores to aid generation of the deep features knowledge base. The box should output a score ranging from 1 to 5. This study explores the design and development of the black box using deep transfer learning techniques.

In our supervised learning framework, appearance outcomes and raters' scores are inputs to the model. Conservatively, one may regard all the scores as weighing equally. It is vital to systematically eliminate appearance outcome assessors who generate the ground truth scores. Facial visuals are read into a convolutional neural network (CNN) for deep features extraction, hence feature pattern learning in relation to ground truth scores. This is implemented by gradually increasing the CNN complexity by appending more dense layers after the feature extractor until a desired outcome is achieved.

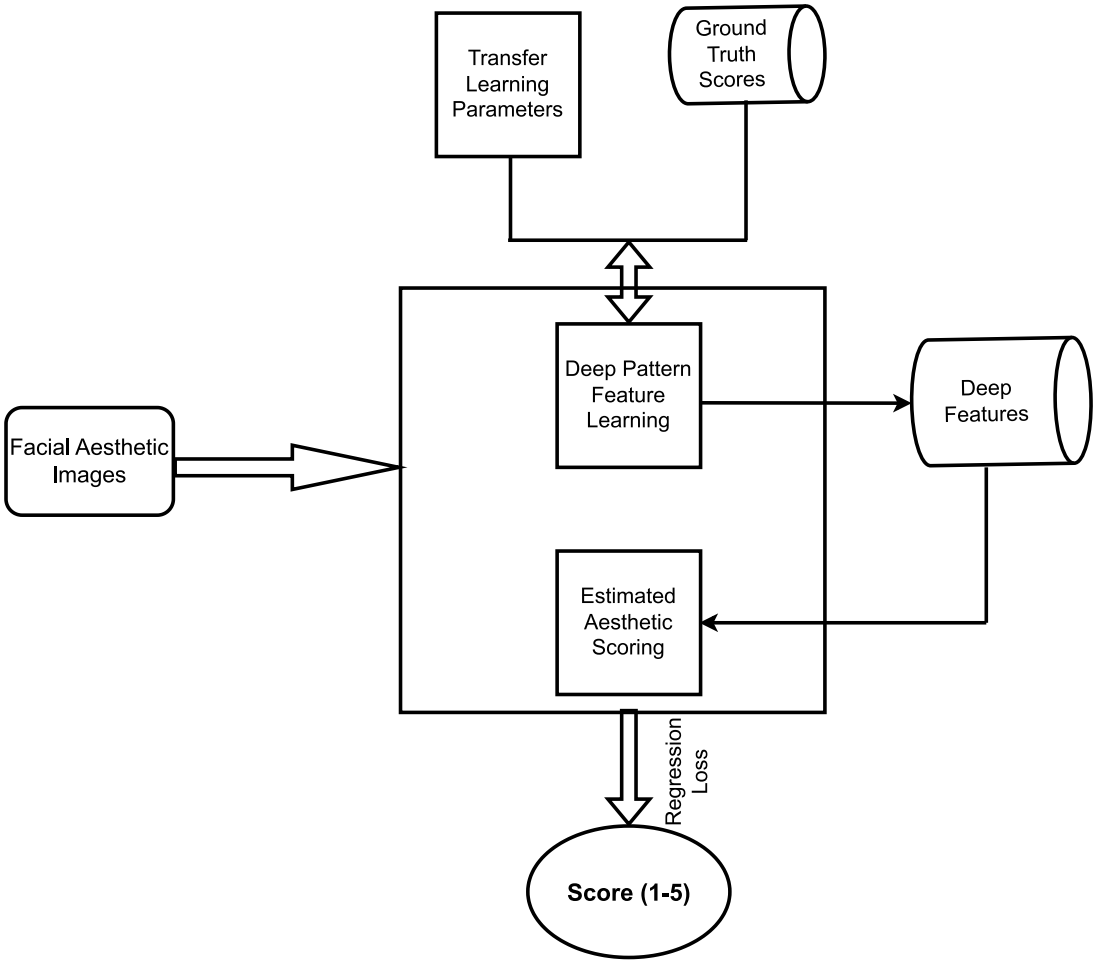

Figure 6. 1: Abstract design of RCNN. The black box at high level simply accepts input aesthetics and their respective raters' scores to aid generation of the deep features knowledge base. The box should output a score ranging from 1 to 5. This study explores the design and development of the black box using deep transfer learning techniques.

6.2.2 Deep Features Extraction and Pattern Learning

Deep features are consistent responses of a node or layer within a hierarchical model such as a CNN. Deep features in our context are numeric descriptors obtained from a conceptual and insightful CNN presented in Figure 6.2.

Figure 6.2 shows the potential inner workings of the three adapted models. One of the models, the VGG16 adapted model is visually represented in Figure 6.3.

Whereas these features are often used for classification, object recognition and localisation (Russakovsky et al., 2015), our research uses deep features as a source of 'hidden or obscured parameters' for regression analysis and modelling. Deep features such as landmark/RoI intensity and mouth shape, nose proportionality, among others, are represented by some variable x. Exponential growth of x could bias the model outcome (in Figure 6.4). Therefore, features (represented by x) should be

suppressed through early and frequent monitoring of the outcome through proper fitting.

The chosen frameworks of VGG16, ResNet50 and MobileNetv1 have been previously trained using ImageNet dataset. Human visual capacity apparently outcompetes VGG16 but is outcompeted by ResNet50 on ImageNet dataset (Alzubaidi et al., 2021).

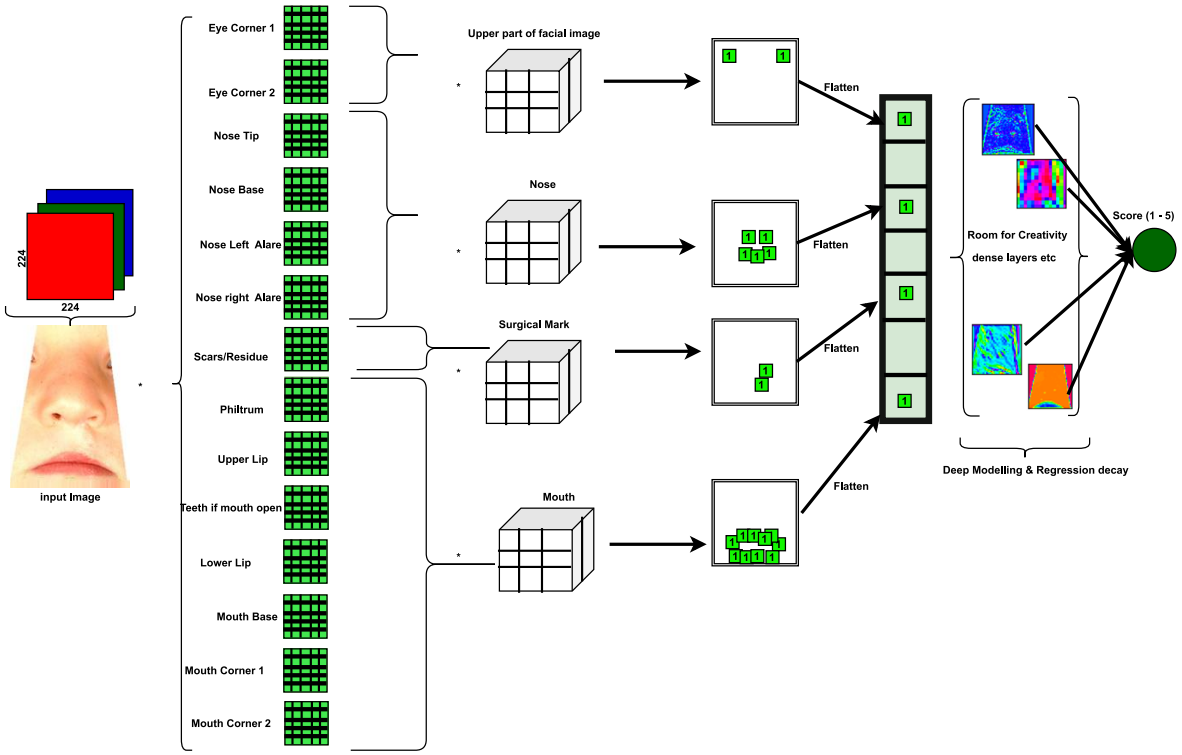

Figure 6. 2: An illustration of the deep learning model for this study. This diagram shows the potential inner workings of the three adapted models. A visual is read, and split into smaller visuals, potentially containing the required features that are eventually aggregated. One of the models, the VGG16 adapted model is visually represented in Figures 6.3 and 6.5.

MobileNetv1 on the other hand is a portable lightweight model with a streamlined architecture that uses deep-wise separable convolutions (Howard et al., 2017). It has performed well in non-mobile applications such as breast cancer mammography studies and skin lesion classification, using transfer learning (Falconi, Perez and Aguilar, 2019, Sae-Lim, Wettayaprasit and Aiyarak, 2019).

In Figure 6.3, the flattening framework and dense layers have managed and handled the different deep features. This visual representation can be generated for the other adapted frameworks, ResNet50 and MobileNetv1.

Figure 6.5 is a Layer-wise and Block-wise representation of the model architecture presented in Figure 6.3.

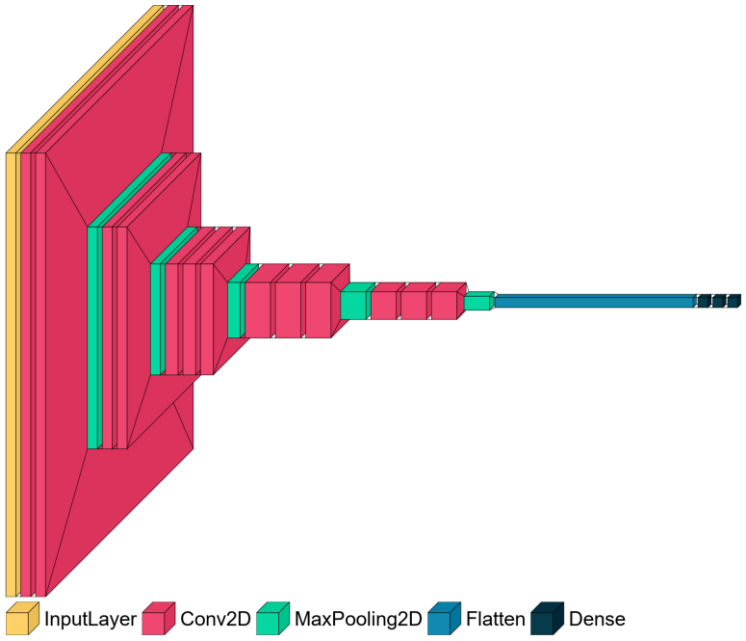

Figure 6. 3: A VGG16 transfer learning framework used in this study. Several pooling layers and convolutional layes used on the input visuals before flatenning into several dense layers. The numbers are empirically determined in a fine tuning exercise. This visual representation can be generated for the other adapted frameworks, ResNet50 and MobileNetv1. Figure 6.5 is a Layer-wise and Block-wise representation of the VGG16 model architecture as presented in this visual.

Our design leverages on the principle of transfer learning to initialise the model weights, W, and partially suppress the potential need for large datasets. Figure 6.4 represents the relationship between input features, weights, biases, and outputs.

Figure 6. 4: Model Weight intuition. Every feature has a weight (level of importance) attached to it, numerically and a bias to regulate (normalise or 'balance up') the outcome.

Where: x is the input/features, W is the weight, b is the bias, and y is the output. Different model weights and parameters are often used with varying combinations of input to get the right outcome. Therefore, W and b are used during hyperparameter tuning.

6.2.3 Regression Model Adaption and Design

Our model employs transfer learning to aid extraction of features based on existing weights. We merge the features using different combinations of pooling layers. The head of the base models is for classification purposes. Figure 6.5 shows the base of the VGG-16 model and the adapted design for the regression task as explained below.

Modifying classification heads in a deep learning model is a conventional practice when adapting a pre-trained model for a different classification or regression task. This process is often part of transfer learning. The following steps have guided the adaptation of the new model from existing frameworks.

1. **Choose and Load Pre-trained Model(s)**

Carefully selecting a pre-trained deep learning model suitable for the regression problem. The models chosen for this task are VGG-16, ResNet50 and MobileNet. This is because these models are of different extremes of computational resources requirements.

Next is loading the pre-trained model weights and architecture. Pre-trained models and their weights are typically resident in deep learning libraries like TensorFlow or PyTorch. TensorFlow was used.

2. **Remove and/or Freeze Existing Layers**

The study requirements dictate that some layers are first removed, that is remove the existing classification head (the output layer). Empirically, some layers were sequentially frozen to extract intermediate features for comparison and some layers, to prevent them from being updated during training. The intention was to fine-tune part of the base network model and eventually replace the classification head with a regression node.

3. **Add and Connect a Regression Head:**

After the feature extractor, the classification head has been replaced with two dense Layers using rectified linear unit (RELU) activations. This facilitates a robust decay for a regression result. The Functional API of the Keras Library is used to build the model.

Next step is to connect the output of the last layer of the regression head to the existing/base model's architecture. This step is fundamental for creating a single, unified model as illustrated in Figure 6.5.

A loss function minimization evaluates a regression model’s reliability. We compile the built model using Loss functions L1 and L2 and an optimiser. Adam optimiser was used with different values. Finally, the model is fitted on the training and validation dataset for executable model generation (i.e., training process occurs). This marks the

first steps towards building a stable executable model because setting the right parameters is a top-heavy iterative process.

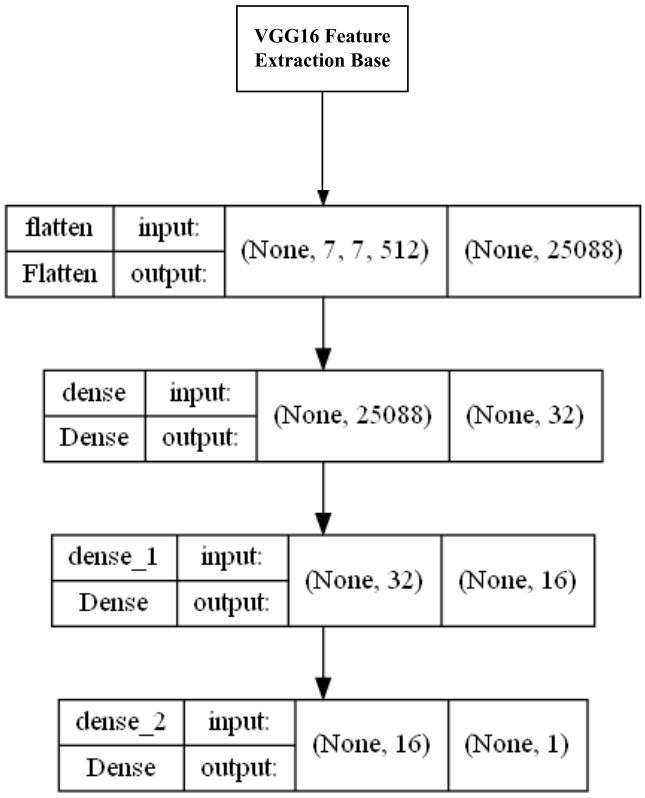

Figure 6. 5: An alternative representation of the VGG-16 adapted model, represented as layers and blocks. Basically, many can layers constitute a given block. The representation has a base-layer (chosen as VGG16), and other additional layers as can be experimentally fine-tuned.

6.2.4 Definition of Parameters and Hyperparameter Tuning

The different parameters needed for the training are defined in Table 6.3 while Table 6.4 defines the loss functions and evaluation metrics.

Table 6. 3: Definition of parameters and variables.

C	CCUK Dataset
n	Size of C
C_{Aug}	Augmented C
m	Size of C_{Aug} where $m \gg n$
R_j	Rater number j to assess C where $1 \leq j \leq 3$
X_i	Random image i in C_{Aug} where $i \leq 0.95m$
$Y_{i,j}$	Ground truth score of X_i by R_j
$Y'_{i,j}$	Average of $Y_{i,j}$ over i
$\hat{y}_{i,j}$	Predicted Score of X_i based on $Y_{i,j}$ – The main output of the model
$\hat{y}'_{i,j}$	Average of $\hat{y}_{i,j}$ over i
α	Bias introduced through hyperparameters adjustments
$f_\Omega(.)$	Learning algorithm - predefined in general terms in Section 2.
$x_{i,j}$	Feature extracted from X_i and $Y_{i,j}$ are inputs for $f_\Omega(.)$

Table 6. 4: Definition of Loss functions and evaluation metrics

$L1$	Least absolute deviation
$L2$	Least squared error
P	Pearson's correlation coefficient, where -1 ≤ P ≤ 1
MAE	Mean Absolute Error – computed from $L1$
$RMSE$	Root Mean Squared Error – computed from $L2$
R^2_Score	Coefficient of determination regression score, where $\infty < R^2\_Score < \infty$
Ω_t	Adjusted optimiser settings t when defaults are not favourable. Implying a different learning rate for $f_\Omega(.)$

Because hyperparameter tuning is an iteratively resource intensive engineering process, we experimented with a random search algorithm and a grid search algorithm for execution of $f_\Omega(.)$ in a resource constrained environment. Nonetheless, the results section presents outcomes where computing resources are readily available.

Models were trained based on the different ground truth scores for RaterA, RaterB and RaterC, satisfying the following condition based on definitions in Table 6.3: R_j: Rater number j to assess C where $1 \leq j \leq 3$.

Therefore, the definition of $L1$ and $L2$ are respectively influenced by these conditions.

$$L1_j = \sum_{i=1}^{m} |Y_{i,j} - \hat{y}_{i,j}| \quad \text{(Equation 15)}$$

$$L2_j = \sum_{i=1}^{m} (Y_{i,j} - \hat{y}_{i,j})^2 \quad \text{(Equation 16)}$$

The lower the values of $L1$ and $L2$ the better the value of $\hat{y}_{i,j}$. Because the CCUK dataset doesn't have any outliers, our model's regularisation function is $L2$, instead of $L1$. $L1$ and $L2$ are therefore computed for the validation phase dataset to mitigate potential overfitting challenges.

Additional metrics used in this study are intuitively presented and mathematically defined below.

Pearson Correlation Coefficient (P): Is the most common way of measuring a linear correlation by giving an indication of the strength and direction of a relationship between two variables. For good results, the value of P should be closer to 1. P is defined as:

$$P_j = \frac{\sum_{i=1}^{m}(f_\Omega(X_i) - f_\Omega(X_i'))(Y_{i,j} - Y_{i,j}')}{\sqrt{\sum_{i=1}^{m}(f_\Omega(X_i) - f_\Omega(X_i'))^2 \sum_{i=1}^{m}(Y_{i,j} - Y_{i,j}')^2}} \quad \text{(Equation 17)}$$

Root Mean Squared Error $(RMSE)$ is computed from the Mean Squared Error (MSE). MSE is the average of the squared difference between the ground truth and predicted values for a given dataset. Essentially, it estimates the variation/variance of any residuals. $RMSE$ closer to zero is the desired outcome although any non-negative outcome is possible.

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{m}(f_\Omega(X_i) - \hat{y}_{i,j})^2}{m}} \quad \text{(Equation 18)}$$

Mean Absolute Error (MAE) is expressed as the average of the absolute difference between ground truth and predicted image scores in the dataset. Unlike $RMSE, MAE$ measures the average of any residuals. MAE closer to zero is the desired outcome although any non-negative outcome is a practical possibility.

$$MAE_j = \frac{\sum_{i=1}^{m}|f_\Omega(X_i) - \hat{y}_{i,j}|}{m} \quad \text{(Equation 19)}$$

R-Squared Score $(R^2 Score)$ is a statistical measure helping with presentation of the proportion of variance for the dependent variable as explained by the independent variable in each regression model. Best possible score is potentially 1.0. If the regression model is subjectively terrible, then $R\_Squared$ can be negative.

$$R\text{^}2\_Score_j = 1 - \frac{Sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

$$= 1 - \frac{\sum_{i=1}^{m}(Y_{i,j} - \hat{y}_{i,j})^2}{\sum_{i=1}^{m}(Y_{i,j} - Y_{i,j}')^2} \quad \text{(Equation 20)}$$

6.2.5 Dataset Distribution

In the previous two chapters, a small dataset of images was used for the experiments. However, this chapter will use the full dataset for the experiments instead. The dataset used in this study was provided by CCUK. It consists of 250 anonymised facial images of 5-year-old children who underwent surgical cleft lip repair. Based on $Table\ 6.3$, $n = 250 = |C|$. Facial anonymity is a mandatory ethical requirement to respect patients' privacy. Further, the dataset contains outcome assessment scores from human experts and carers. Orthodontists, plastic surgeons, language/speech therapists,

psychologists, and individual patients made assessments based on the facial appearance outcomes from surgical treatment.

All facial images in c were included for the different experiments. The dataset was statistically analysed to gain insights into the reliability of the assessors prior to conducting more deep learning experiments. For the five raters (RaterA, RaterB, RaterC, RaterD and RaterE), Figure 6.6 shows the respective scores' distributions.

The median score distribution of all the raters' scores has been included because scores for 4 of the 5 raters are visibly asymmetrically distributed. Otherwise, using mean would be a natural choice. Empirically, RaterB has a score distribution closer to normal compared to the rest. Empirical observations, in addition to statistical estimation of scores spread, are used to eliminate some raters before model building and training (Jordan and Mitchell, 2015).

Summary statistics are used to provide more visual cues about how well the score summaries represent the raters' performance in relation to the median. The default dataset standard deviation and mean values were used to visualise error bars in Figure 6.7, where the divergence is evident. Summary statistics are usually more appropriate with the addition of error bars, which provide a visual cue about how well the summary represents the underlying data points.

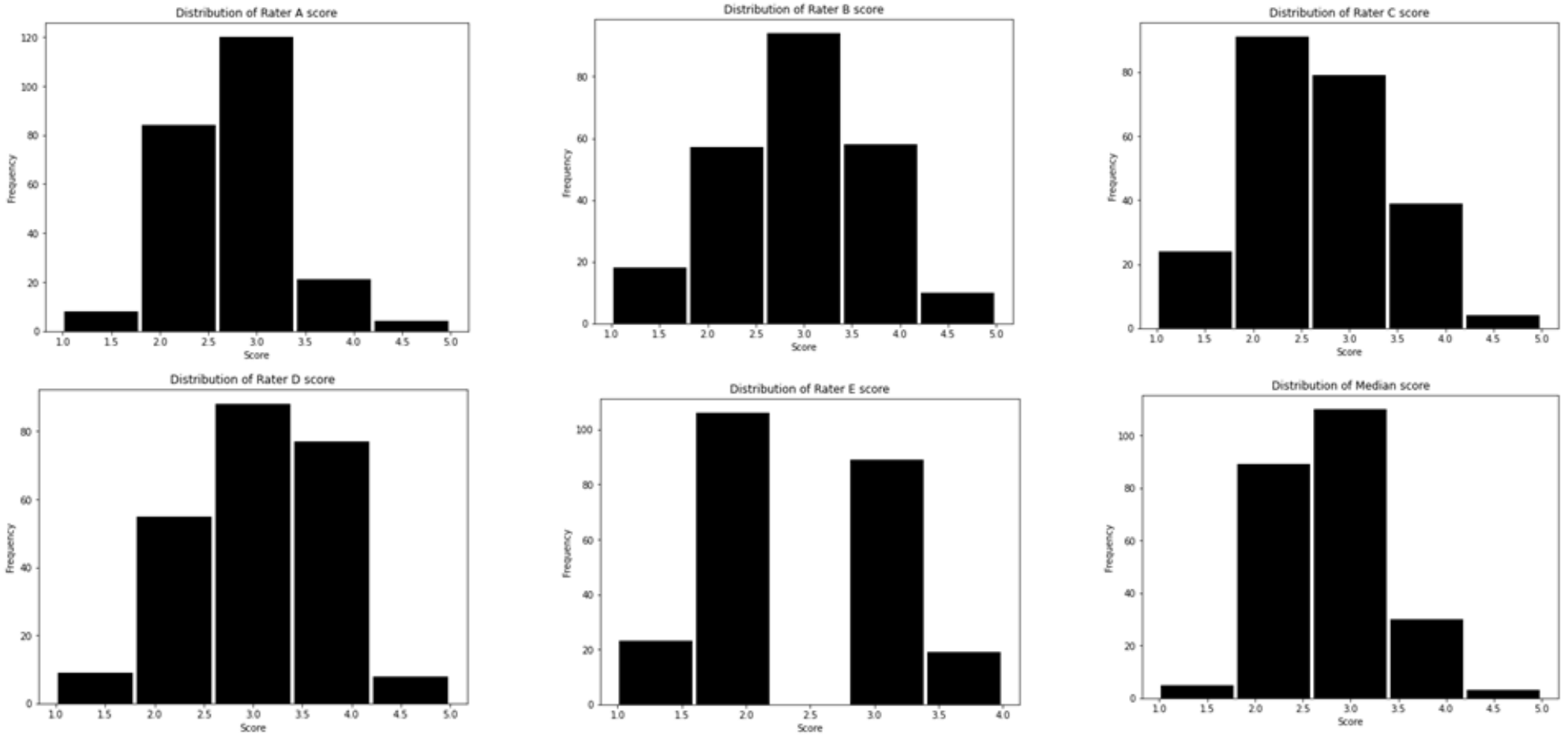

Figure 6. 6: Distribution of Scores based on each Rater, bottom right is the median distribution for all the five raters. A normal distribution is not expected. Uniform rater scores would be the best result though it is not a practical possibility.

Several libraries have sub-modules, notably from the Seaborn module, that can be configured to aid with the automatic calculation of both summary statistics and the error bars from a given dataset. Seaborn is a lending library for making statistical visuals in Python programming language and contains several submodules (Waskom, 2021). We use error bars in bar plots for the additional visualisation to aid with Raters' elimination. In statistics and mathematics, a bar plot represents an estimation of central tendency for a numeric variable with the height of each rectangle and provides some indication of the score ambiguity/uncertainty (Unpingco, 2019).

Additionally, an error bar around a given score relationship estimates central tendency capable of showing either the range of parameter uncertainty or the spread of the underlying dataset around the parameter. Some of the raters may present scores whose uncertainty and accuracy may be biased based upon indecisive vector feature representation in deep learning (Sra, 2016).

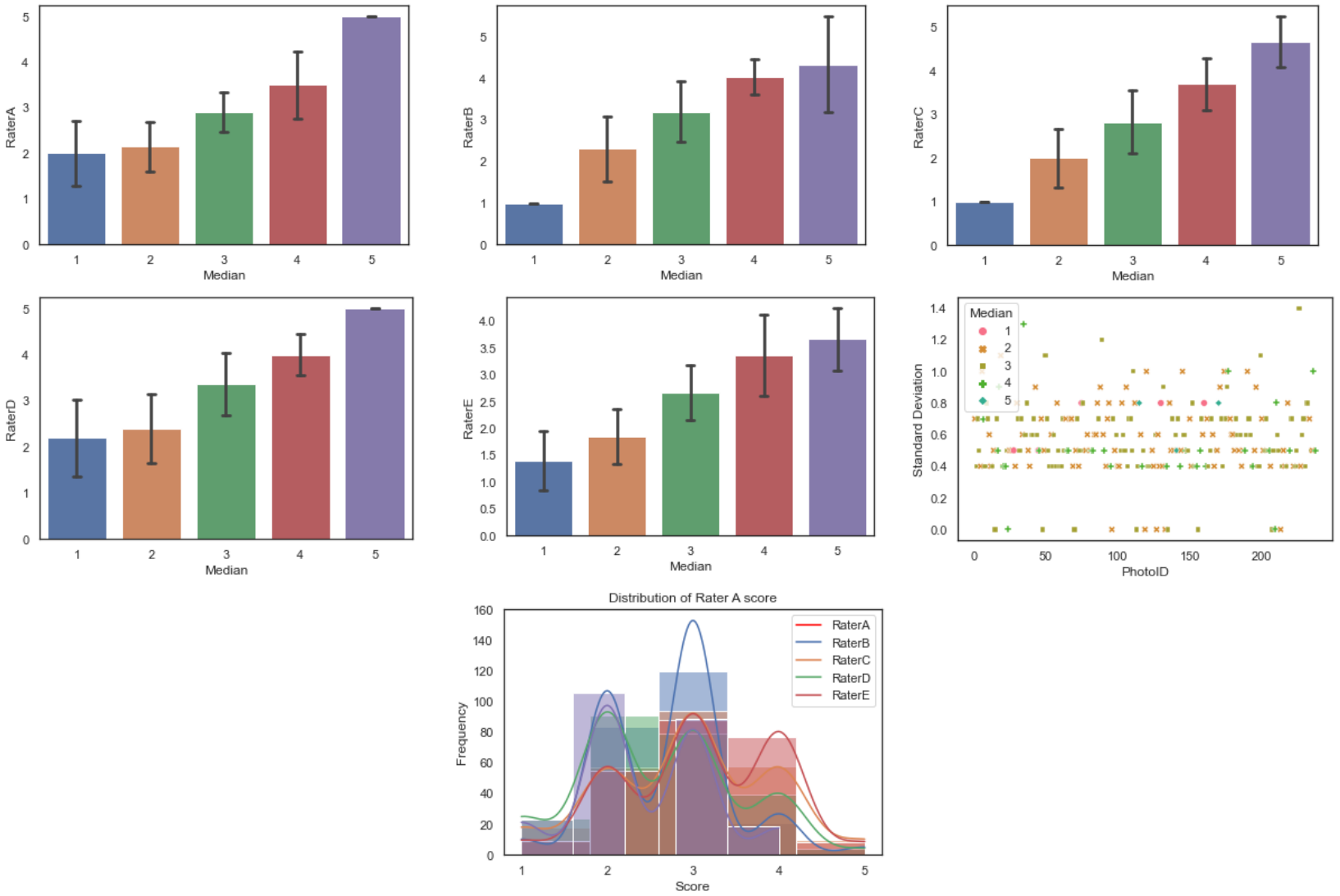

Figure 6. 7: Certainty/Uncertainty visualisation of error cues for raters' scores against the median. Bottom most image is aggregated distribution with Kernel Density Estimation (KDE) for each rater. Raters B and C had difficulty rating visuals as '1 = excellent' and Raters A and D could not rate many visuals as '5 = very poor'.

Having dealt with the entire dataset, the data spread is more significant than the uncertainty. Therefore, Rater E was eliminated due to the largest variance from the median score and having no images with scores in one predefined range (see Figure 6.6 and Figure 6.7, 2nd along Row 2). Rater D scores are preserved for test purposes, implying Rater A, Rater B and Rater C scores were used in building and validating the deep CNN model whose set up is described in this chapter.

6.2.6 Implicit Preprocessing

1. Image Augmentation

This involves the generation of new training images by applying a set of transformations such as rotation, translation, scaling, flipping, or adding noise to the original images. Augmentation helps increase the diversity and size of the training dataset, leading to improved model performance, generalisation, and robustness. This technique is often also used in machine learning and deep learning approaches where the dataset is small.

The ImageDataGenerator module in Keras aided the image augmentation process using the selected transformation techniques in Table 6.5. The intention is to create more examples for the representation of various imaging conditions from the training set to represent a real-world setting.

Table 6. 5: Image Augmentation properties to aid the Transfer Learning-based approach.

Rotation range (degrees)	10
Horizontal flip	True
Zoom_range	0.1
Brightness_range	(0.5 – 1.5)

An outcome of such augmentation properties is the visual combination in Figure 6.8. The image in Figure 6.8 is the same image but under different augmentation/ transformation conditions to represent a real-world setting. Reading from left to right, the first 2 images are horizontally flipped under slightly different lighting conditions and so are the 3rd and 4th. The 5th image is zoomed in while the last image is an illuminated representation of the original outcome.

Consequently, the dataset increased from 250 images to 4735 images with the corresponding labels. The new dataset is used to aid better feature extraction through

transfer learning using deep convolutional neural networks (DCNN).

Figure 6. 8: This sample aesthetic outcome combination represents an image following treatment. This is the same image but under different conditions as per a real-world setting. First 2 images are horizontally flipped under slightly different lighting conditions and so are the 3rd and 4th. The 5th image is zoomed in while the last image is an illuminated representation of the original outcome.

2. **Image Filtering**

This is the convolution of an image with a filter or kernel to perform operations such as blurring, sharpening, or enhancing specific features. Common types of filters include Gaussian filters, median filters, and high-pass filters. Filtering can be used for noise reduction, smoothing, feature extraction, or image enhancement. Machine learning and deep learning techniques can employ several filters.

6.2.7 Deep Learning and Regression Modelling

Regression analysis is the process of estimating the relationship between a minimum of two variables, usually independent and dependent variables. Several regression techniques and functions can be used to fit independent variables to get the dependent variable. Using this approach in machine learning, a prediction model can be designed (Chatterjee and Simonoff, 2013). Normally, regression analysis research is embedded in a supervised machine learning context. In this study, the independent parameters are the different features of the facial appearances' outcomes, following the surgical treatment while the dependent attribute is the outcome assessment estimation. The assessment estimation should be a continuous number, from 1 to 5.

Normally, with regression analysis, the extent of parameters disintegrations and limited interaction is celebrated more than the opposite. Hence, the metrics to verify successful regression prediction are measured using loss functions. The following mathematical expression clarifies this analogy:

A regression estimation/ prediction takes a function $f_\Omega(.)$ $(parameterized\ with\ \Omega)$ given some feature points $(x_i, y_i)\ \forall\ i\ \in \{0, 1, \ldots, m-1\}$ under a loss function l: $L = \sum_i l(f(x_i), y_i)$. The aim of a regression is to estimate the function f while minimizing the total loss L of all the data items.

In this research, given the nature of our dataset, we categorize the regression model based on the function $f_\Omega(.)$ and loss function l. The complicated and subjective nature of independent variables (x_i) makes it harder to discover if their true relationship to the dependent variable (y_i) is linear, polynomial, logarithmic or logistic. Given this fact, physically learning, analysing, and grading of the complicated facial features is harder not only for human beings but also for most categories of regression analyses (Godec et al., 2019). After the different transformations (stated in Table 6.5 and Figure 6.8), some visualisations of extracted features can be generated.

6.2.8 Visualisation of Feature Maps

Deep learning models are traditionally very hard to explain, that's why they are usually treated as black boxes. But CNN models are actually the opposite. This section helps with the visualisation of various components.

With an example of a facial visual image the internal works are decoded and visualised as below. The visualisation of the feature maps aids us to see how the input is transformed passing through the convolution layers. The feature maps are also called intermediate activations since the output of a layer is called the activation.

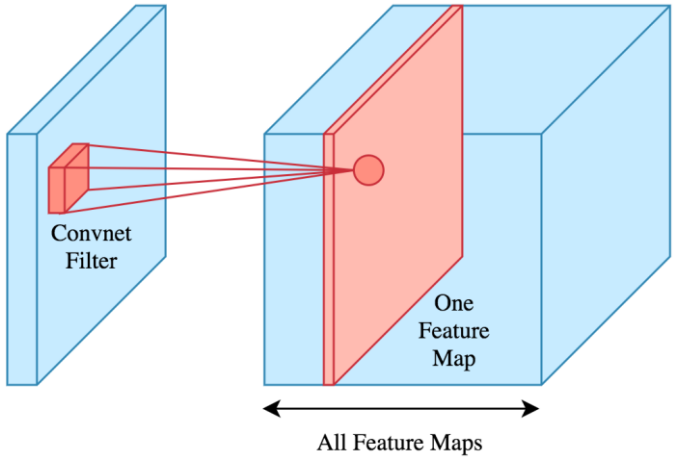

Figure 6. 9: Conceptual visualisation of feature maps using 3D convolution filter. A collection of pixels is a potential source of features (feature map space) with in a given image as demonstrated.

The output of a convolution layer is a 3D volume as seen in Figure 6.9 (adopted from (Yosinski et al., 2015)). The height and width correspond to the dimensions of the feature map, and each depth channel is a distinct feature map encoding independent features. Therefore, individual feature maps are visualised by plotting each channel as a 2D image.

Actual visualisation works by passing an input visual image into a CNN and recording the intermediate activation. Random feature map selection is done to plot the visualised output.

Since the first component of the deep learning model is features extraction, VGG convolutional layers are named as 'BlockX_ConvY'. For instance, the third filter in the first block would be coded as 'Block1_Conv3'. The illustration in Figure 6.10 shows VGG-16 architecture with 16 layers, minus the SoftMax and any other pooling layers.

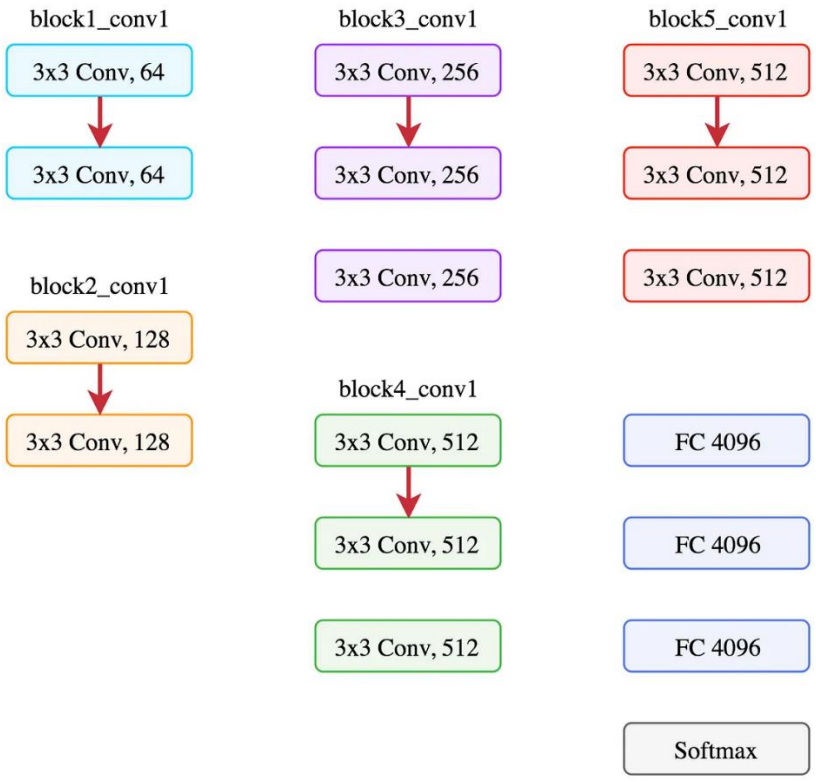

Figure 6. 10: Block-wise and Filter-wise visualisation of VGG16 model. There are 5 blocks with different sizes (number of convolutional layers). These extract high level features before feeding into the dense layers for either a classification or a regression task.

There are several convolution filters in each block as indicated in Figure 6.10 (adopted from online Dertat, 2017), however, we cannot visualise all of the feature maps. For example, block 3 outputs 256 distinct feature maps. Therefore, we sequentially access any desired number of feature maps. In this study, the first sixteen (16) were chosen,

stacked as 4 by 4 (the last row in Figure 6.11) across the input facial image or four (4) 2D feature maps stacked as 2 by 2 (the first 2 rows in Figure 6.11).

Therefore, given the size of the different filters in each layer, we get the first 16 or 4 feature maps from the first convolution layer (conv1) of each block (as per the column headings of Figure 6.11). However, other feature map groups may be specified, as and when necessary.

Figure 6.11 visualises feature maps extracted from a layer-wise and block-wise approach. The first column is for feature maps from Block1, second column, Block2 and continues to fifth column, Block5. The first row in Figure 6.11 uses default RGB colour visualisation while the second and third rows use HSV for visualisation. Feature maps are clear in the latter colour space.

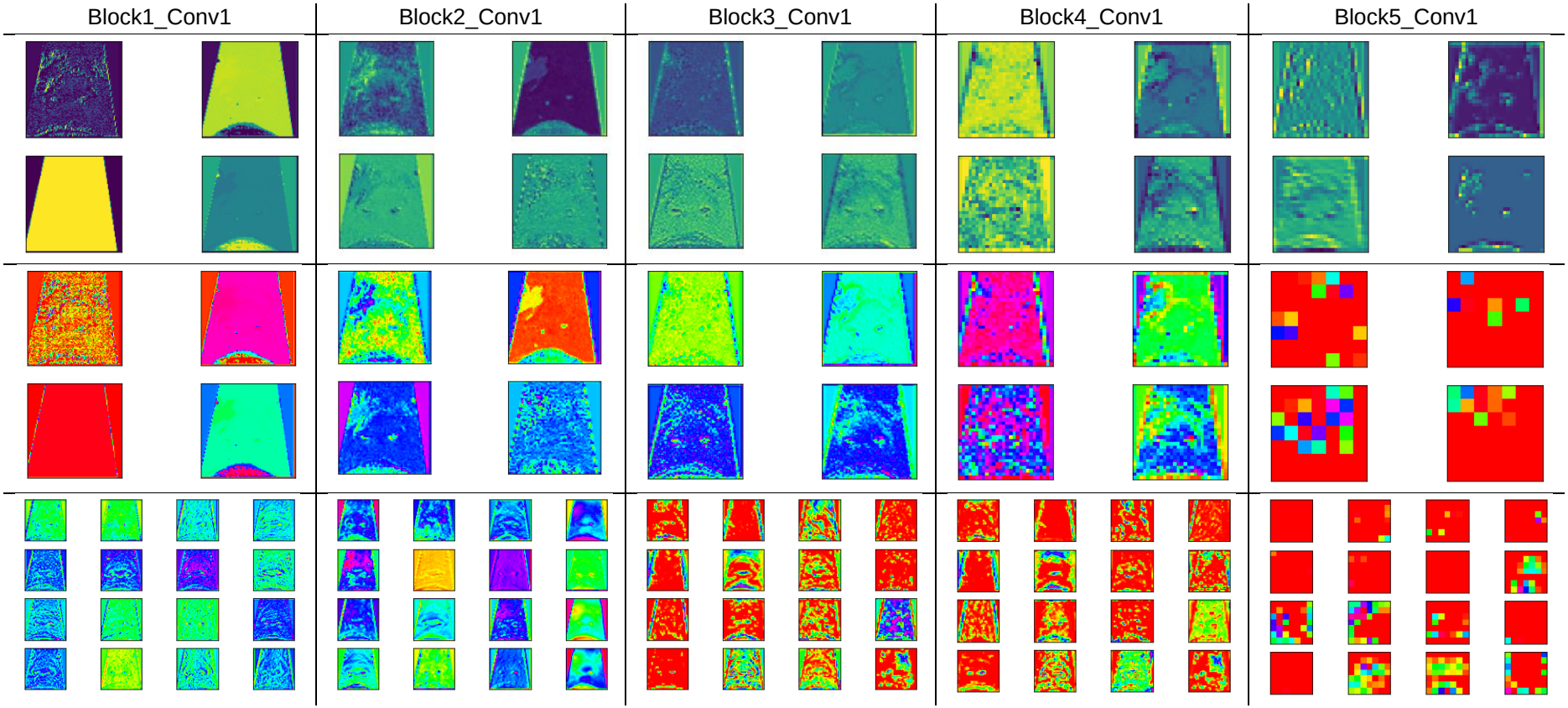

Figure 6. 11: Block-wise feature maps extraction from VGG16-based architecture. The three rows (top to bottom) represent the level of detail of extracted features. It could have been more. For each block, a visualisation is made for the extracted features. As expected, block 5 represents a concrete feature map at level 3 (lowest level).

The regression framework partly exploits such robust features using both random and sequential features selection. Hybrid features selection is not feasible because the output formats of the different model blocks cannot be harmonized using a computational approach. A feasible manual selection approach would be unreliable and inept. Figure 6.12 also demonstrates that the model learns from the same facial image that has been presented differently, following successful augmentation.

Following features extraction as partly visualised, a deep learning model processes thousands or millions of parameters using neurons of the RELU activation function combinations. A given layer of the model can be updated based on the state of the layer before or after, courtesy of the computational flexibility and efficiency introduced by backpropagation (Lecun, Bengio and Hinton, 2015). We can naively assume that the different feature points are processed using a polynomial function to generate an assessment outcome in the range of 1 and 5, inclusive. The visualisation in Figure 6.8 facilitates this intuition, hence eliminating any linear regression function for the calculation of the outcome assessment score. The implementation is however different for the model computational steps of the last blocks/layers of our predictive model.

To compute an assessment score $\hat{y}$, we use a polynomial expression below to fit x:

$$\hat{y} = f_{\Omega}(x) = \Omega_0 + \sum_{j=1}^{m} \Omega_j x^j \qquad \text{(Equation 21)}$$

Assuming a preference and predefined number of features or feature points, F, (x_i, y_i). Where $(x_i, y_i) \in \mathbb{R} \, \forall \, i \in \{0, 1, \ldots, m-1\}$, we can fit the polynomial function, with degree, v through a minimization expression:

$$\min_{\Omega} \sum_{i} ||y_i - f(x_i)||^2$$

The idea is to calculate $(v + 1)$ variables denoted by $\Omega_0, \ldots, \Omega_v$. The intuition is that for linear regression, the polynomial function has a degree, $v = 1$.

Calculation of $\Omega_0, \ldots, \Omega_v$ is feasible for a known dataset, from which parameters/features or independent variables x_i have been extracted. The CCUK dataset distribution has been discussed in the previous sub-section 6.2.5.

6.3 Experimental Configuration and Results

6.3.1 Overview

Presented in this section are the different experimental configurations and results. Before the results presentation, an overview of selected terminologies is given to facilitate understanding the models' architectural visual representations. These concepts have been detailed in (Goodfellow, Bengio and Courville, 2016, Zhang et al., 2021).

1. Activation function: One that facilitates a neural network to learn a non-linear or complex relationship. In our supervised learning context, we have features and labels (scores). Functions (such as the RELU used in this study), facilitate neural networks to map features to labels for eventual inference. Other activation functions include sigmoid, tanh, ELU (exponential linear unit) and many others.
2. Batch Normalisation: Inputs and outputs into activation functions should be normalised in each hidden layer. This helps stabilise neural networks by dropping outlier weights and potentially enable networks learn faster.
3. Batch size: The sample images from the dataset that the model processes and determines how frequently the network parameters are updated per iteration.
4. Convolution: It is a machine learning idiom for convolutional layer or convolutional operation. A convolution is a mathematical computation (hence combination) of two functions, one holding the filter and another holding the input image or intermediate feature matrix, hence, normally invoked as 'Conv2D'.
5. Convolutional Neural Network (CNN): Is a neural network that where one of the layers is a convolutional layer. Other layers could be a combination of pooling layers and/or dense layers. It is only limited by the creativity of the network engineer. CNNs are popular because key features are extracted without human intervention, and they have voracious capability to map features with minimal corresponding labels.
6. Pooling: The process of reducing the size of matrices generated by an earlier convolutional layer into a smaller one for context capture. Global Maximum Pooling or Global Average Pooling are two often used functions, implemented

at the pooling layer, to take the maximum value or average value respectively, across the pooled section.

7. Dense Layer: Also referred to as fully connected layer, is a hidden or output layer from which every node is connected to every other node of the subsequent hidden layer.
8. Flatten: Is a function used for input vectorisation, like pooling, but changes all the resultant 2D arrays from feature maps into a long linear vector. Therefore, flattening always generates a 1D vector, unlike pooling.

6.3.2 Experiments and Parameter Settings

To fit the dataset to the model, a random split for model training, evaluation and testing was necessary of 4735 images. A random distribution of 75% for training purpose and 25% for model testing (evaluation). Additionally, another random split for the training images was conducted, 85% of the training phase images were used for training while 15% were used for validation purposes. Details are found in Table 6.6.

Table 6. 6: Initial Dataset distribution

Training Phase: 75% of C_{Aug}		Evaluation Phase: 25% of C_{Aug}
# images for Training	# images for Validation	# images for Testing
3018 (85% of C_{Aug})	533 (15% of C_{Aug})	1184

However, if this challenge was a classification problem, then stratified dataset splitting would have been a better approach (Kahloot and Ekler, 2021).

Therefore, 25% of the dataset used for evaluation is not read into the network layers during the training phase and is used for model evaluation only. The dataset size may however dictate the distribution.

As indicated, we chose the Adaptive Moment Estimation (Adam) algorithm as the model optimiser with default settings. More settings of Adam were experimented after a grid and random search with outcome detailed in Table 6.7. A description of Adam, Stochastic Gradient Descent (SGD) and other optimisation algorithms can be found in (Curtis and Nocedal, 2018, Alzubaidi et al., 2021).

Table 6. 7: Random Search and Grid Search hyperparameter outcomes for VGG16-based model

Default Optimization Settings	Learning rate: 0.001 Algorithm name: Adam	
Other Experimental Settings	Grid Search	Random Search
Learning rate	0.0005	0.0006
Dropout_rate	0.2	0.4
# filters	64	64
# Units	256	64
# Trainable parameters	3,250,497	840,513

The platform settings have been defined in Table 6.8. NVIDIA GeForce 940MX GPU (now GeForce MX350) is known for accelerating computational power on laptops by 250%, which is considerably faster than many modern CPUs.

Table 6. 8: Hardware and software requirements

Hardware	Software
NVIDIA GeForce 940MX (GeForce MX350), pci bus id: 0000:01:00.0, compute capability: 5.0	64-bit Operating System, Microsoft Windows 10, version 22H2
Computing Capability 5.0	Python Programming Language v 3.9.0
Random-access storage is 16GB	DL Frameworks: Tensorflow, Keras (2.8.0). CuDA, and CuDNN
GPU memory is 8GB	Libraries: Pandas, Matplotlib, Seaborn, Sklearn, Scipy, Pillow, and numpy.

6.3.3 Implementation Summary

The pseudocode below represents the abstract deep learning project involving regression analysis. The goal is to predict a continuous numerical target variable (the appearance assessment scores based on input images. The pseudo code indicates a framework for feature extraction using any suitable selected pre-trained model (such as VGG16, ResNet50 or MobileNetv1). A custom regression node is built on top of the feature extraction base. This is accomplished through the steps below:

1. Importing Libraries.
1.1 Import necessary libraries for deep learning (TensorFlow, Keras), data handling (Pandas), model visualisation, and other utilities.
2. Definition of Constants and Global variables.
2.1 Set various constants and global variables, such as image dimensions, data paths, and working directory.
3. Loading and Preprocessing Data.
3.1 Load a dataset that of images and associated scores. Split the dataset into training and testing subsets.
4. Creation and Initialisation of Data Generators.

4.1 Create data generators for training, validation, and testing. These generators preprocess the images and prepare batches for training and evaluation.
5. Loading any Pre-trained Model of choice.
5.1 Load a pre-trained model with weights initialised from scratch or using ImageNet pre-trained weights. Empirically, the latter was applied to this research.
5.2 Freeze all layers of the loaded pre-trained model to prevent them from being updated during training.
6. Building a Custom Regression Model.
6.1 Create a custom regression model by adding new layers to the pre-trained base model.
6.2 An appropriate features aggregation/ batch framework such as Flattening, Global Average Pooling 2D (GAP), Global Max Pooling 2D (GMP) is applied to the output of the loaded model to reduce the spatial dimensions.
6.3 A Dense layer with a single neuron and a linear activation function is added to produce the regression output.
7. Model Compilation.
7.1 Compile the regression model, specifying the optimizer (Adam), loss function (such as mean squared error - mse), and any other metrics.
8. Model Training.
8.1 Train the regression model using the training and validation datasets.
8.2 Early stopping is implemented to monitor the validation loss and stop training when it stops improving.
9. Model Evaluation on Test Set.
9.1 Evaluate the trained model on the test dataset, calculating various performance metrics such as RMSE (Root Mean Square Error), R-squared (R2) value, Pearson correlation coefficient (PCC), and mean absolute error (MAE).
10. Visualisation
10.1 Visualise the model predictions for a subset of test images alongside their ground truth scores.
10.2 Export the results to a suitable storage site such a file.
11. Plot Loss History.
11.1 Visualise the training and validation loss history to assess the model's convergence and generalisation.

6.3.4 Primary Findings

Transfer learning techniques shaped the model configurations with prior variables about three raters and three architectures. Presented below are results from TL settings and combinations using the VGG16 framework. The training graphs and predictions presented in Figures 6.12 and 6.13 respectively are based on the following condition:

$$If\ j\ =\ 1, then\ R_1\ = RaterA$$

Where the features $x_{1,1}$ extracted are aggregated using the following vectorisation arrangements:
i: GlobalMaxPooling (GMP), vectorisation of features without any dense layers. Otherwise, GMPv2 denotes that 2 additional dense layers were applied; ii: GlobalAveragePooling (GAP), vectorisation of features without any dense layers. Otherwise, GAPv2 denotes that 2 additional dense layers were applied; iii: Flattening (FLT), vectorisation of features without any dense layers. Otherwise, iv: Flattening with additional 2 dense layers (FLTv2).

Features extraction and aggregation for deep propagation is challenging, but some system functions exist to aid users to create models faster. In Figure 6.12, the labelled visuals were generated from the feature maps aggregation frameworks defined as i: GlobalMaxPooling; ii: GlobalAveragePooling; iii: Flattening; iv: Flattening with additional 2 dense layers. Between i and ii, model validation ranges between 0.6 and 0.4 over 50 epochs, leading to early overfitting. In iii, the size of the feature vector potentially leads to early convergence and non-uniformity though the validation range is wider (0.15 and 0.4). With the additional 2 dense layers in iv, the model trains and shows signs of learning beyond the 50 epochs, presenting the best outcome.

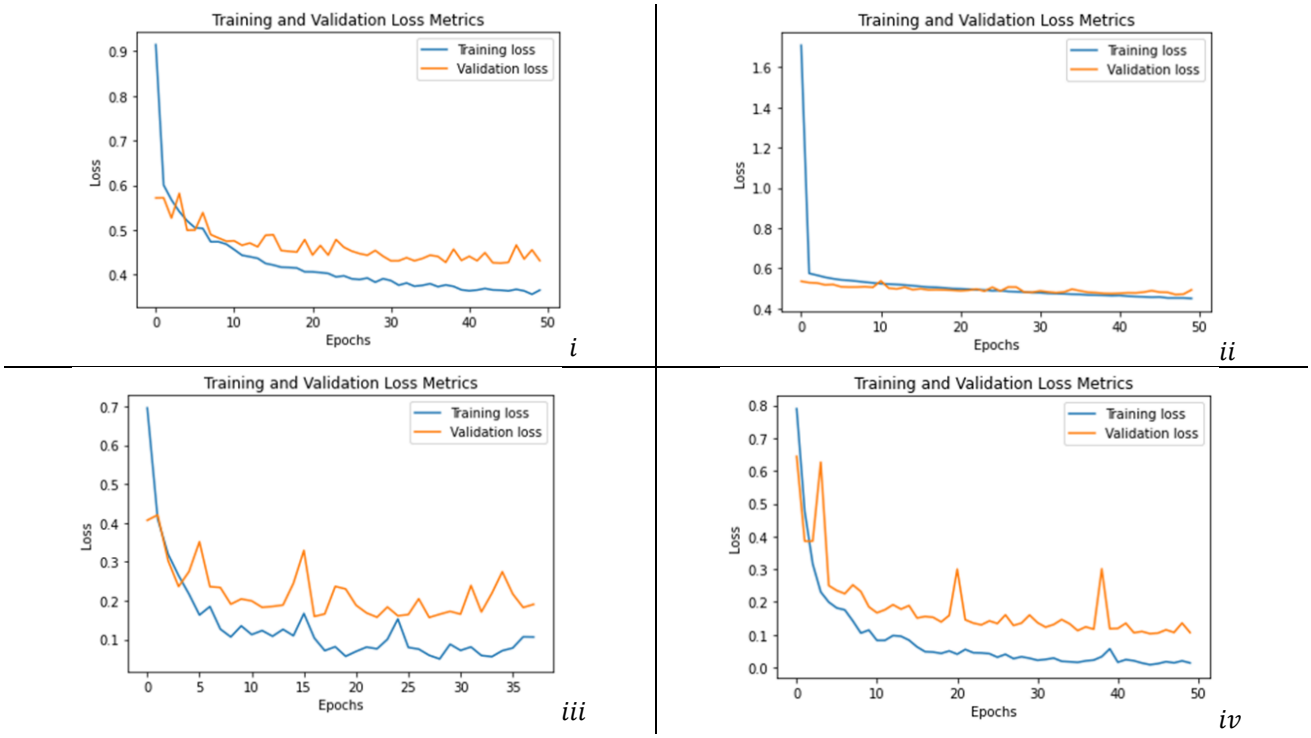

Figure 6. 12: VGG-based Model training and validation visualisation for the first rater under different settings. Features extraction and aggregation for deep propagation is challenging, but some system functions exist to aid users to create models faster. From i: GlobalMaxPooling; ii: GlobalAveragePooling; iii: Flattening; iv: Flattening with additional2 dense layers. Between i and ii, model validation ranges between 0.6 and 0.4 over 50 epochs, leading to early overfitting. In iii, the size of the feature vector potentially leads to early convergence and non-uniformity though the validation range is wider (0.15 and 0.4). With the additional 2 dense layers in iv, the model trains and shows signs of learning beyond the 50 epochs, presenting the best outcome.

The sample scoring outcomes for arrangements i and iv are labelled with PS for predicted score and GTS for ground truth score. There was limited learning in the settings of i and ii. In both cases, there was early overfitting, ii being worse, where the overfitting manifests as while the training loss decreases, the validation loss does not. The settings with options iii and iv reveal a better outcome for model fitting (see the validation ranges as captioned in Figure 6.12). Following successful training is model validation with sample outcomes in Figure 6.13.

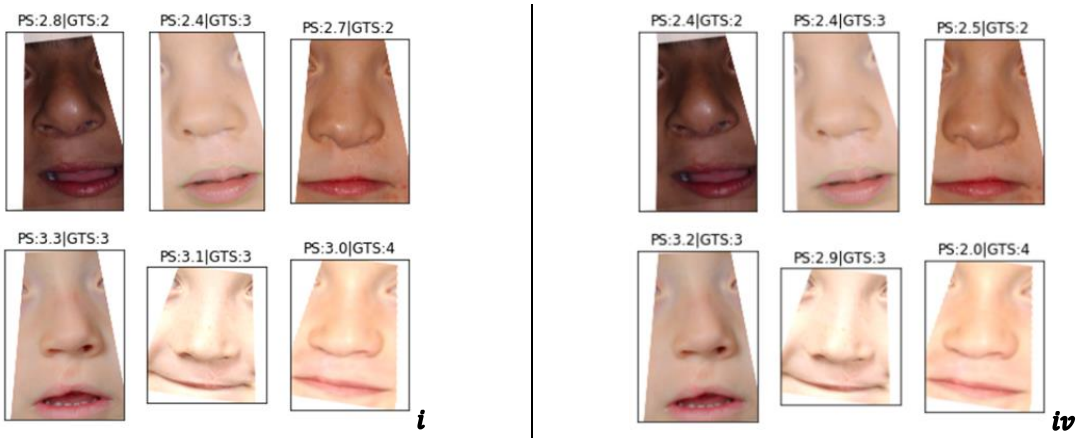

Figure 6. 13: Predicted scores of some images based on the trained VGG16 architecture under conditions i and iv with apparent better predictions under condition i for the showed results.

The sample outcomes in Figure 6.13 are few and superficially indicate i outperforming iv. However, Figure 6.13 and Table 6.9 together indicate that the overall settings in iv offer the best outcome for scoring of the test dataset images with the lowest $RMSE$, MAE and the highest P and R^2_Score, for the first rater, R_1.

Using the VGG16 framework, more quantitative outcomes generated for the following settings summarised in Table 6.3 and Table 6.4 are presented in Table 6.9.

If $j = 2$, then $R_2 = RaterB$,

If $j = 3$, then $R_3 = RaterC$

To generate Tabel 6.9, different models based on the VGG16 framework were fitted under different settings. For instance, the visuals in the dataset were used to fit and evaluate the model using the first rater (R_1) labels. To aggregate the resulting features during model building, different vectorisation frameworks have been applied. These include global maximum pooling (GMP), global average pooling (GAP) and the regular flattening arrangement (FLT). The resulting model is evaluated different metrics. The choice of these metrics is root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient (p) and r-squared score. Once a model is built, an evaluation is conducted to ascertain its accuracy on the unseen/unused subset of the main dataset. This aids in the decisions about rater's and parameter's reliability. The process is repeated using different rater labels for the same dataset.

Table 6. 9: TL with VGG16 architecture. Different raters' ground truth scores are used in training the model with different vectorisation arrangements giving different metric outcomes.

Assessor Ground Truth	Vectorisation Framework	$RMSE$	MAE	P	R^2_Score
R_1	$i\ (GMP)$	0.653	0.514	0.507	0.246
	$ii(GAP)$	0.699	0.536	0.419	0.536
	$ii(GAPv2)$	0.751	0.619	-0.024	-0.004
	$iii(FLT)$	0.448	0.338	0.811	0.645
	$iv(FLTv2)$	**0.370**	**0.256**	**0.876**	**0.758**
R_2	$i\ (GMP)$	0.980	0.743	-0.034	-0.012
	$ii(GAP)$	0.764	0.640	-0.045	-0.033
	$ii(GAPv2)$	0.976	0.752	-0.009	-0.003
	$iii(FLT)$	0.536	0.409	0.841	0.700
	$iv(FLTv2)$	**0.470**	**0.354**	**0.900**	**0.769**
R_3	$i\ (GMP)$	0.934	0.798	-0.002	-0.005
	$ii(GAP)$	0.934	0.798	-0.001	-0.005
	$iii(FLT)$	**0.572**	**0.440**	**0.801**	**0.624**
	$iv(FLTv2)$	0.869	0.698	0.446	0.131
R_4	$i\ (GMP)$	0.935	0.740	0.094	0.005
	$ii(GAP)$	0.935	0.735	0.099	0.007
	$iii(FLT)$	0.943	0.768	0.153	-0.010
	$iv(FLTv2)$	**0.913**	**0.732**	**0.243**	**0.053**
R_5	$i\ (GMP)$	0.717	0.626	-0.059	-0.014
	$ii(GAP)$	0.715	0.624	0.008	-0.007
	$iii(FLT)$	0.673	0.568	0.360	0.106
	$iv(FLTv2)$	**0.631**	**0.510**	**0.515**	**0.214**
R_{Median}	$i\ (GMP)$	0.768	0.630	0.013	-0.005
	$ii(GAP)$	0.767	0.627	0.025	-0.003
	$iii(FLT)$	**0.462**	**0.350**	**0.803**	**0.627**
	$iv(FLTv2)$	0.725	0.589	0.332	0.105
$R_{Average}$	$i\ (GMP)$	0.661	0.543	0.008	-0.005
	$ii(GAP)$	0.661	0.543	0.006	-0.004
	$ii(GAPv2)$	0.659	0.546	0.040	0.001
	$iii(FLT)$	0.676	**0.346**	**0.582**	**0.099**
	$iv(FLTv2)$	**0.639**	0.531	0.287	0.061

It is consistently true from Table 6.9 that vectorisation by flattening and flattening with additional dense layers outputs optimal metrics. Different ground truth values produce mixed results under different conditions. All seven ground truth values produce optimal results when the feature aggregation framework is flattening. However, while flattening with additional dense layers is optimal for Rater A (R_1), Rater B (R_2), Rater D (R_4) and Rater E (R_5) and it is not the case for Rater C (R_3), and the median(R_{Median}) and mean (R_{Mean}) ground truth sets. This indicates the possibility of R_3 heavily skewing the Median and Mean computations. This is potentially because the R_3 has the least appearance assessment knowledge or makes severely biased assessments. Therefore, exploring how the weighted mean dataset performs would be interesting.

Other features' aggregation frameworks are poor because they are not robust and are non-suitable with processing hybrid datasets. Also, VGG16 as a framework potentially favours a few frameworks for aggregation in non-classification deep learning problems. The best results from the optimal metrics were realised using ground truth from Rater A (R_1) and Rater B (R_2). Based on these outcomes, the subsequent experimental results for ResNet50 and MobileNetv1 architectures were explored for only R_1 and R_2, as seen in Tables 6.10 and 6.11, respectively.

Table 6. 10: ResNet50-based Results

Assessor Ground Truth	Vectorisation Framework	$RMSE$	MAE	P	R^2_Score
R_1	$i\ (GMP)$	0.749	0.603	0.060	0.002
	$i\ (GMPv2)$	0.758	0.621	-0.041	-0.022
	$ii(GAP)$	**0.738**	0.608	0.218	**0.037**
	$ii(GAPv2)$	0.749	0.625	0.171	0.008
	$iii(FLT)$	0.750	**0.579**	**0.413**	0.006
	$iv(FLTv2)$	0.745	0.622	0.239	0.018
R_2	$i\ (GMP)$	0.978	0.749	0.023	-0.007
	$i\ (GMPv2)$	0.978	0.738	0.049	-0.007
	$ii(GAP)$	0.985	0.762	0.009	-0.022
	$ii(GAPv2)$	0.978	0.751	0.016	0.000
	$iii(FLT)$	**0.916**	**0.714**	**0.358**	**0.122**
	$iv(FLTv2)$	0.926	0.719	0.346	0.104

Because vectorisation by global max pooling did not result in optimal outcomes with ResNet50 experiments (please see Table 6.10), the subsequent experiments present results where vectorisation was only by global average pooling and flattening. In Table 6.10, it is reasonable to conclude that there is no single stable features vectorisation framework for R_1dataset. ResNet50 is a generally resource intensive framework which might have underperformed over time on a limited dataset (Xu, Fu and Zhu, 2023). However, R_2 is suitable with only GAPv2 across the evaluation metrics. There is considerable consistency in the results across Table 6.11.

Table 6. 11: MobileNetv1-based Results

Assessor Ground Truth	Vectorisation Framework	$RMSE$	MAE	P	R^2_Score
R_1	$ii(GAP)$	0.522	0.405	0.726	0.518
	$ii(GAPv2)$	**0.331**	**0.246**	**0.900**	**0.806**
	$iii(FLT)$	1.840	1.331	0.302	-4.986
	$iv(FLTv2)$	0.338	0.249	0.899	0.798
R_2	$ii(GAP)$	0.641	0.504	0.763	0.571
	$ii(GAPv2)$	**0.334**	**0.243**	**0.945**	**0.884**
	$iii(FLT)$	3.003	1.967	0.315	-8.426
	$iv(FLTv2)$	0.425	0.320	0.904	0.811

Table 6.12 shows an adjusted dataset distribution to mitigate potential early convergence and overfitting.

Table 6. 12: Adjusted dataset distribution.

Training Phase: 95% of C_{Aug}		Evaluation Phase: 5% of C_{Aug}
# images for Training	# images for Validation	# images for Testing
4048	450	237

The aggregated outcomes following Table 6.12 are presented in Table 6.13 below, for MobileNetv1-based transfer learning framework only. Other model frameworks were eliminated because TL based on MobileNetv1 yielded the most optimal outcomes. The results in Table 6.13 are better than those in Table 6.11 for both ground truth datasets. This is attributed to a bigger training and validation dataset; hence elimination of early convergence is significant.

Table 6. 13: MobileNetv1 results after adjusting the dataset distribution.

Assessor Ground Truth	Vectorisation Framework	$RMSE$	MAE	P	R^2Score
R_1	$iii(FLTv3)$	**0.310**	**0.230**	**0.913**	**0.829**
	$ii(GAPv3)$	0.750	0.614	NaN	-0.001
R_2	$iii(FLTv3)$	0.376	0.283	0.932	0.851
	$ii(GAPv3)$	**0.293**	**0.216**	**0.955**	**0.909**

The training curves are flatter with less overfitting as seen in Figure 6.14. The most optimal result in Table 6.13 is where the second dataset for Rater B (R_2) with global average pooling using a few dense layers as per the previous architecture.

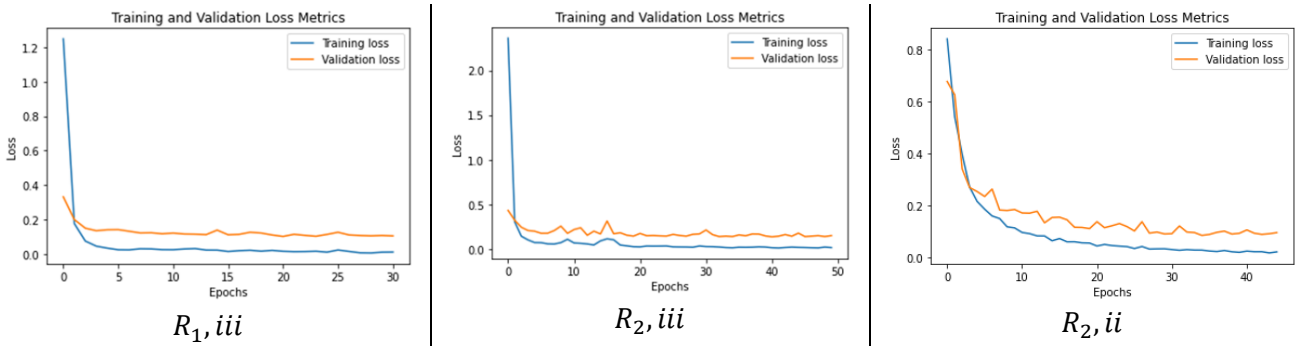

Figure 6. 14: MobileNetv1-based model training and validation visualisation for first and second raters under different settings. R_2, ii shows a better learning outcome implying that a better model was fitted/ built.

6.3.5 Reducing the Regression Task into a Classification Case

The regression model can be naturally specialised as a classification model. Assessment of facial image appearance following cleft lip treatment can be considered a classification task. The scores of different facials are classified into only five (5) categories: 1 – 'Excellent', 2– 'Good', 3– 'Fair', 4– 'Poor' and 5– 'Very Poor'. Therefore, a classification model is created in the following two aspects as a predictive modelling technique for classification or categorisation on facial images into the predefined classes (Kotsiantis et al., 2007, Lu and Weng, 2007): (i) through replacing the regression head in 1D output with the classification head with 5D output, and (ii) while the scores are used as real number for regression, they are treated as categories for classification. The goal is to automatically make decisions that map features of different facial images appearance into their corresponding classes and use the outcomes to serve as a validation technique for the regression model. Note that the classification and regression models use the same datasets for training and learning with different outputs.

The classification model was built based on already class labelled data instances. The facial images and their features are the independent variables while the labels are the dependent variables. Preprocessing was simplified because the class labels are either categorical or numerical. Since the study analyses facial images, a CNN-based architecture was conveniently used for feature extraction and mapping to the different brands and train a model (Caldeira et al., 2020). The Adam optimiser was used with MobileNetv1- based CNN classification model, with scores/labels from Rater B as the basis. Some metrics are summarised in Table 6.14. The three potential optimisers (in Table 6.14) can be used for training our classifier. However, Adam is most accurate

across the entire classifier development cycle, therefore, it is the preferred optimiser for the classifier in this study.

Table 6. 14: Accuracy Metrics for the classification model during training, validation, and testing phases for the three selected optimisers with Rater B labels.

Phase	Adam	SGD	RMSprop
Training	0.993	0.989	0.988
Validation	0.997	0.978	0.983
Testing	1.000	0.963	0.974

Adam optimiser produced better classification outcomes than the other optimisers for the MobileNetv1 baseline architecture. This may not be generalised for the rest of the architectures explored in this study. More experimental work may be conducted. Table 6.9 can be generated using different optimisers for different raters' labels and CNN model architectures.

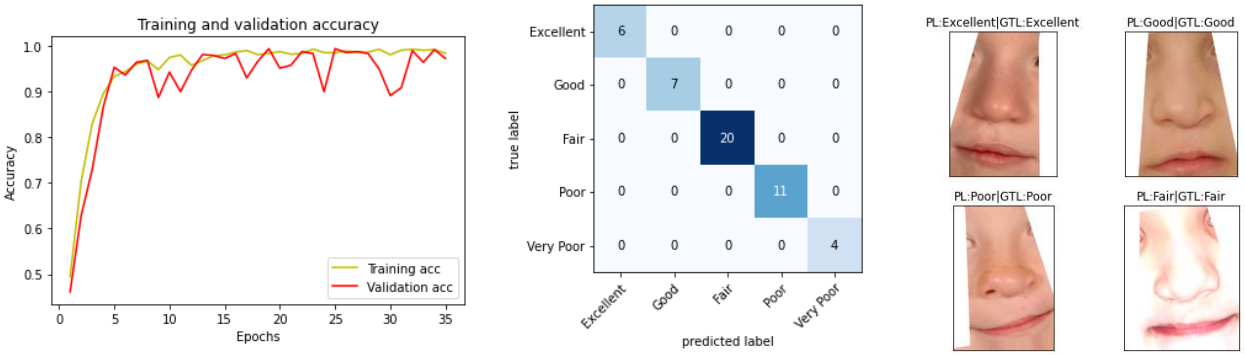

Figure 6. 15: Some results from the CNN classification model. Left – Model Training and Validation Accuracy graph shows that excellent learning took place. Middle- Confusion, matrix indicates that testing classification accuracy was perfect. With early stopping and checkpoint call backs, it is observed that at epoch 25, the validation accuracy was nearly 100. A model was saved at that point, which was eventually loaded and used for classification testing. Right – Prediction results (PL) against the ground truth (GTL).

The graph in Figure 6.15 (Left) indicates how the classification model training and validation was nearly perfect. The model training and validation accuracy graph shows that excellent learning took place. Figure 6.15 (middle) is the confusion matrix with perfect label classification. With early stopping and checkpoint call backs, it is observed that at epoch 25, the validation accuracy was nearly 100%. A model was saved at that point, which was eventually loaded and used for classification testing. The original 25 evaluation images and 23 randomly selected other images were used for testing, resulting in 48 visuals. In Figure 6.15 (Right), predicted label (PL) is made

against the ground truth label (GTL) for selected visuals and the assessment labels are rightly expected.

6.4 Evaluation of Deep Regression Analysis and Assessment Approach

6.4.1 Model Performance

The R^2Score is a good and reliable alternative to Pearson's correlation, P. All key results above, bar one, indicate a direct proportion between the two metrics. Table 6.15 summarises Tables 6.9, 6.10, 6.11 and 6.13 with optimality applied to Table 6.9.

Table 6. 15: Best results aggregated from the three architectures following several experiments.

Assessor Ground Truth	Model	Vectorisation Framework	$RMSE$	MAE	P	R^2Score
R_1	VGG16	$iv(FLTv2)$	0.370	0.256	0.876	0.758
	MobileNet	$ii(GAPv2)$	0.331	0.246	0.900	0.806
		$iii(FLTv3)$	0.310	0.230	0.913	0.829
R_2	VGG16	$iv(FLTv2)$	0.470	0.354	0.900	0.769
	ResNet50	$iii(FLT)$	0.916	0.714	0.358	0.122
	MobileNet	$\boldsymbol{ii(GAPv2)}$	**0.334**	**0.243**	**0.945**	**0.884**
		$\boldsymbol{ii(GAPv3)}$	**0.293**	**0.216**	**0.955**	**0.909**

Table 6.15 was generated using an elimination method based on the optimality of the four metrics. All the four metrics were considered at their optimal best across the entire row/record for inclusion.

The fact we used several metrics, this section attempts to explain the significance of the quantitative outcomes as shown in the tables above. For over a decade, machine learning solutions interoperability has outpaced its 'explainability'. Several practitioners have questioned their validity and societal impact (Carvalho, Pereira and Cardoso, 2019).

Training models with scores provided by the first rater (R_1) performed well across two architectures. In contrast, scores provided by the second rater (R_2) scores were more robust across all the three backbone models, chosen for experimental setup. As earlier stated, the higher the values of P and/or R^2Score and the lower the values of $RMSE$ and/or MAE, the better the regression model. The MAE value is the extent to which an

estimation of a score can go off the ground truth. Therefore, minimization of MAE and $RMSE$ is an optimality target.

Any intuitive choice for the best possible outcome would be either of the two approaches below:

a) Approach 1 with initial dataset distribution:

- Using ground truth for rater R_2
- Transfer learning with MobileNetv1
- Random or Sequential feature aggregation through global average pooling

b) Approach 2 with a revised dataset distribution

- Use ground truth for rater R_2
- Transfer learning for MobileNetv1
- Random or Sequential feature aggregation through global average pooling

Rater R_2 has proven a more reliable assessor for facial appearances outcomes following CL treatment. In general practice, surgeons are best placed to train, monitor, and objectively appraise each other's practices. Despite their multifaceted curriculum during training, surgeons are often assessed based on their clinical ethics, passionate scientists and researchers to improve their practices (Soh, 1998). This has not changed, even in recent times. Therefore, while assessing other surgeons' outcomes, a set of surgeons should be considered objective enough, if a dataset of images presented to them is limited, also referred to as assessment of reduced 'surgical volumes' (Mayer et al., 2009).

In their work on assessment of quality of care in surgery, (Mayer et al., 2009) proposed that clinical pathway measures such as the structure of care, process of care and outcome of care should be emphasized if patient satisfaction is to be improved. Additionally, (Mayer et al., 2009) present a detailed framework for their proposed quality of surgical care framework, which was statistically proved using a dataset for coronary artery bypass surgical activities.

Using a transfer learning-based regression CNN framework, it is possible to assess partial appearance outcomes with errors in the range of $0.293 < RMSE < 0.334$ and $0.216 < MAE < 0.243$. Global average pooling aggregation is preferred for better

results and assures a Pearson correlation of between 94.5% and 95.5%. Segment-wise acceptance of the model network layers facilitated getting the desired features. There is room to potentially improve upon these metrics by having diverse datasets. It was noted that in many medical research studies, “data on outcomes after surgical interventions are often of poor quality. The lack of consistent reporting is well highlighted in the medical literature” (Domenghino et al., 2023:1). Some open datasets potentially include images available to facilitate research. This would open doors for a more comprehensive research activity for an unsupervised learning study for the future. Frameworks should be proposed for potential generation of reliable and expert-evaluated CL-based datasets.

6.4.2 Comparison of Regression Analysis to Other Assessment Methods

Previous attempts to assess appearance outcomes following CL surgical treatment have undertaken quantitative and qualitative approaches. SymNose, in some of the studies, applied the thin lip correction algorithm to improve lip appearances before and after human raters made the assessment (Kornmann et al., 2019). Inter-rater and intra-rater reliability improved between 0.80 and 0.78 to 0.81 and 0.83, respectively. However, this method does not support generation of precise continuous score of a facial appearance outcome following treatment. Another approach involved 3D analysis of soft tissue for symmetry computation on postoperative images. The symmetry is compared between the treatment outcome appearances and the regular facials. If the normal and the postoperative images have convergent balances, then surgical treatment can be individualised otherwise, generalisation of treatment protocols take precedence (Schwenzer-Zimmerer et al., 2008).

In shape analysis framework. Low level features of the mouth are used to compute the facial symmetry. Similarity of the mouth region is calculated using the symmetric axis and eventually converted into a score assessment. The automatically computed assessment score is compared to the human score using Pearson’s correlation coefficient (Bakaki et al., 2021). Additionally, (Bakaki et al., 2022) further divided the different facial image into three third components and detected features separately. Symmetry was detected using the shortest Manhattan distance between the center of the three thirds and the other features. Three scenarios were used to determine similarities of structural components from which assessment score is calculated and a correlation is determined.

Finally, classification model was developed to simplify cleft lip treatment assessment into known categories of 1, 2, 3, 4 and 5. This approach is less restrictive and requires less computational resources than regression analysis. Table 6.16 aggregates the different studies' Pearson's correlation coefficient, where applicable. It shows that CNN based classification has the most potential for simulating the human behaviours in assigning scores to cleft images after cleft repair. However, CNN based regression model provides more detailed and fine-grained appearance numeric assessment.

Table 6. 16: Aggregated correlation metrics from different studies.

Study	PCC
CNN based Regression	0.955
CNN based Classification	1.000
Hybrid Key Landmarks Detection-based Method	0.567
Traditional-based Shape Analysis	0.396
Direct Comparison between AI-based scoring and Conventional Raters	N/A
Quantitative 3D Tissue analysis of Symmetry	N/A
SymNose and Thin Lip Correction algorithm	N/A

6.4.3 Summary

This study investigated the effectiveness of using CNN models for transfer learning to perform appearance assessment through regression analysis. The objective was to leverage pre-trained models and transfer their learned features to improve the performance of our particular regression scoring task. Through different experimentation and analysis settings, several key findings have emerged. Results demonstrate that transfer learning can significantly enhance appearance assessments using regression techniques from the flexible CNN frameworks of VGG16, ResNet50, or MobileNetv1. Using their pre-trained model weights, these networks can effectively learn high-level features from the CCUK dataset, improving scoring accuracy. This finding confirms the potential of transfer learning as a valuable technique for regression tasks, particularly with limited datasets. MobileNetv1 was found as most viable network. Additionally, the choice of the pre-trained models and the layers for transfer play a crucial role in achieving optimal performance. Different pre-trained models have diverse architectures and represent different levels of abstraction in visual features. Therefore, careful consideration should be given to selecting the appropriate pre-trained model. Furthermore, fine-tuning the transferred layers in the CNN can help refine the learned representations and appropriately adapt them to appearance outcome scoring. Overall, the findings of this study emphasise the

potential of transfer learning in regression scoring/assessment with CNNs. The approach offers a powerful means to leverage the great potential of pre-trained models, enabling improved regression performance, and overcoming dataset limitations. However, further research is needed to explore the impact of different transfer learning strategies, parameters/features selection criteria and investigate the optimal choice of hyperparameters. This study contributes to the growing body of research on transfer learning in regression analysis, providing insights into its effectiveness. The demonstrated benefits open new opportunities for applying CNNs with transfer learning in supervised regression appearance scoring challenges. Table 6.16 summarises some results from different studies.

Chapter 7 Discussion and Conclusion

7.1 Introduction

This section involves an interpretation and analysis of research findings and a critical discussion of their implications. A summary of the thesis and the future research direction is indicated. Last but not least, the research strengths and limitations are also discussed.

7.2 Discussion

This PhD study aimed to develop and evaluate computational appearance assessment techniques for cleft lip treatment using partial facial images. Leveraging advanced image processing algorithms and machine learning models aimed to provide an empirical and quantitative means of assessing appearance outcomes in cleft lip patients following surgical treatment. Further, the minor objectives include comparison of accuracy and effectiveness of the different assessment techniques developed in the research study.

The approach used in this research study was based on analysis of the different post-treatment facial images from children living with cleft lip condition. The images were processed using state-of-the-art image processing pipelines and computer vision algorithms to extract relevant facial features and quantify appearance parameters. Machine learning models were then trained to predict appearance ratings based on expert assessments using the extracted features. The latter assessments served as a validation parameter sets for the developed computational techniques. Shape analysis employed basic features from the mouth as the region of interest. However, because the mouth is composed of flexible tissue (Colston et al., 1998), its features may not be as reliable and those of other regions like the nose and the eye corners. Accordingly, key landmarks detection for different regions of interest was investigated using a hybrid mechanism. In both approaches, appearance assessment is a quantification of the structural similarity index measure (SSIM) (Wang et al., 2004) using carefully devised mathematical models (Brunet, Vrscay and Wang, 2012). An approach that uses deep learning is also investigated to facilitate better extraction of features. The extracted features are mapped to the human assessors' scores using a transfer

learning regression framework to create a predictive fine-grained assessment model. Transfer learning increases the capability of traditional machine learning models (Pan and Yang, 2009, Zhuang et al., 2021).

7.2.1 General Discussion of Results

The results obtained from the computational appearance assessment techniques demonstrated promising accuracy and reliability in predicting appearance outcomes after cleft lip treatment. In shape analysis, the test dataset consists of human expert generated visuals (GT1, GT2, and GT3). The shape analysis computational approach automatically generates the appearance numeric scores for each test dataset (GT1_AENS, GT2_AENS, and GT3_AENS). There is also the automatically predicted set (PS) of visuals, which is also assessed numerically to generate PS_AENS. The highest computed correlation coefficient is 0.959, between the PS_AENS and GT3_AENS. However, for the most significant correlation coefficient, the highest is 0.399. The most significant correlation is computed between PS_AENS and human expert generated scores (HNS). The three-region key landmarks assessment approach has a correlation coefficient as high 0.940 and as low as -0.102. However, there is noted balanced distribution of structural similarity index measure across all datasets where colour images have been used. The deep learning regression model approach produces the most significant correlation coefficient in the range of 0.945 and 0.955. This is the best generated range of correlation coefficient across the test dataset.

7.2.2 General Comparison and Relevance

Compared to traditional subjective assessments conducted by human experts, the automatic computational appearance assessment techniques showcased several advantages. Firstly, they provided a standardised and consistent evaluation process, eliminating inter-observer variability. Secondly, the computational methods enabled rapid and efficient assessment of appearance outcomes, potentially saving time for clinicians, researchers, and policy makers. Therefore, the tools developed in this research can be considered a good audit resource. However, it is important to acknowledge that these computational techniques cannot fully replace trained professionals' expertise and clinical judgment. Thirdly, the development and evaluation of computational techniques are based on ground truth. While they have

shown great potential, if the ground truth includes bias and the small datasets, these techniques may produce inaccurate results or even fail. Thus, collecting and generating a large enough ground truth dataset is essential for developing the computational and learning based methods.

The clinical relevance of computational appearance assessment techniques for cleft lip treatment is significant. By providing objective measurements of appearance outcomes, these techniques can aid clinicians/surgeons in treatment planning and decision-making processes. Computational techniques offer quantitative tools to evaluate treatment effectiveness and could be used to track patient progress over time through incremental imagery processing adaptation and analysis. Additionally, integrating patient preferences and perceptions can be facilitated by incorporating patient feedback into the computational models. Further, integrating these technics into regular clinical practice can potentially improve treatment outcomes, patient satisfaction, and long-term monitoring and audit.

7.2.3 Discoveries from Research Questions

In the event of a facial visual outcome, computational techniques can be designed and developed to automatically assess its appearance. Additionally, automatic appearance assessment methods have capability to yield both quantitative or numeric and semi-quantitative scores. Additionally, the developed automatic assessment methods can provide the same outcomes when desired over the provided datasets. Hence automatic assessment methods are easily reproducible.

Visual assessment is associated with visual understanding using features from the principal components. Partial facial images have few principal components which constitute regions of interest, from which features are extracted and analysed for appearance assessment. The mouth, and nose and eye corners contain detailed features, from which we define the different parameters. For example, the mouth has the mouth corners, philtrum ridges and boundary. Likewise, the nose features and parameters are well documented. In non-cleft patients, the features may have known or conventional defined shapes and geometric properties, potentially unlike in visuals from people with cleft conditions.

The main challenge affecting automatic assessment of cleft lip treatment outcomes is poor features recognition, stemming from poor dataset preparation. When it is difficult to recognise mouth features or nose features, parameters for modelling contain incorrect values. This in turn generates the wrong assessment from the model. This is overcome by using a supervised deep modelling approach.

Traditional modelling methods are less scalable in case the visual dataset is inconsistently prepared. This affects their accuracy as seen by results in Chapter 4. Hybrid methods and deep learning methods exhibit a self-mitigating mechanism with inconsistent data. This results in consistent generation of parameters from correctly identified features. Consequently, the assessment is more accurate as demonstrated in Chapters 5 and 6.

Progressively, it was discovered that minimal human involvement leads to better assessment results. This is an indicator that unsupervised deep learning is a potential significant step towards achieving the best appearance assessment results on facial outcomes following cleft lip treatment.

Therefore, automatic assessment methods are more robust, accurate, and consistent than human assessment methods as demonstrated in Chapters 4 – 6.

7.3 Conclusion

This research designed and evaluated some computational methods using partial facial images to analyse and assess cleft lip repair. Traditional, hybrid and deep learning computational methods were developed and validated to achieve the research objectives.

Chapter 2 presented the review of relevant literature for this study. The epidemiology of cleft lip was covered and highlighted the underlying causes and potential clinical management or mitigation strategies. However, this Chapter also revealed that cleft lip is a disease of high psycho-social significance, especially among non-black population. Consequently, the need to intervene using contemporary computational techniques was based on the summarised knowledge gaps especially the lack of standardised treatment outcome assessment procedures, limited validation of treatment outcomes, and crucially, to minimise human involvement in cleft lip treatment outcome evaluation.

In Chapter 3, the general and specific methods of realising the research objectives have been presented. Notably, the visual data processing pipeline is discussed in detail, breaking down the significance of the steps, while discussing their implementation and outcomes. Several preprocessing machine learning and deep learning techniques have been experimented as a springboard for Chapters 4, 5 and 6. This Chapter additionally discussed the dataset, its ethical considerations, and its limitations.

In Chapter 4, the shape analysis approach presented significant findings. The computation of the structural similarity index measure allowed for a quantitative evaluation of the cleft lip treatment outcomes. The correlation coefficients indicated a moderate to solid agreement between human assessments and the computationally based assessment, suggesting the potential of this approach in objective appearance assessment.

In Chapter 5, the region-based key landmarks detection and assessment approach showed promising results. Identifying key landmarks in different regions of the partial facial image contributed to an accurate evaluation of treatment outcomes. The correlation coefficients revealed a strong agreement between human expert scores and automatically generated scores, indicating the effectiveness of the models in capturing appearance improvements.

Finally in Chapter 6, the deep learning regression analysis demonstrated notable performance. The model successfully generated an assessment predictive model by mapping deep extracted features to human expert scores. The correlation coefficients between human scores and automatically generated scores were consistently high across the test dataset, indicating the reliability of the deep learning approach in appearance assessment.

Overall, the research findings indicate the potential of these approaches in providing objective and quantitative assessments of cleft lip treatment outcomes. However, further refinement, validation, and collaboration with stakeholders are needed to fully realize their clinical applicability and realistic impact on patient care.

7.4 Future Research, Recommendations, and Limitations

7.4.1 Introduction

Recommendations for this study should shape the future research while focusing on refining the computational models in an iterative manner, involving continuous evaluation, feedback, and improvement. It aims to enhance the models' accuracy, reliability, and practical applicability in assessing appearance outcomes of cleft lip treatment.

Future research should focus on refining the computational models, expanding the dataset to include diverse populations, and incorporating additional features or modalities for improved accuracy. Integrating patient perspectives and preferences into the assessment process is crucial to ensure personalised treatment outcomes. Collaboration with clinicians and patients can provide valuable insights and guide further developments in this multidisciplinary research study. Additionally, the need to fully understand a face in a personal context should be left to a non-supervised machine learning approach. This potentially eliminates some dataset bias, especially introduced by human experts' assessment scores and ratings. The following discussion is a hybrid presentation for future research and recommendations to mitigate some implicit limitations in chapter 4, 5 and 6.

7.4.2 Algorithmic Enhancements

Future studies should explore and implement advancements to the proposed algorithms used in the computational models. For example, one can investigate more sophisticated techniques for shape analysis, key landmarks detection, or other deep learning architectures. This may include considering hybrid state-of-the-art methods in computer vision, pattern recognition, or machine learning to enhance the accuracy and robustness of the proposed models.

7.4.3 Feature Detection, Selection and Extraction

The refinement process of the features may involve analysing the components used in the models and determining which ones are most relevant and informative for appearance assessment. This can include exploring different feature extraction techniques or incorporating additional features that may capture essential aspects of

facial images. Feature engineering and dimensionality reduction methods can also be employed to improve the efficiency and performance of the models. The potential use of generative adversarial networks (GANs) for features engineering is encouraged and recommended if ethical clearance is granted.

7.4.4 Dataset Expansion and Diversity

Refining the models often requires collecting and incorporating more extensive and diverse datasets. Expanding the dataset can help capture a broader range of facial variations, including different ethnicities, ages, and gender. This increased dataset diversity can improve the generalisability and robustness of the models, making them applicable to a broader population.

Numeric or quantitative assessments by different raters (doctors, nurses, surgeons) could be replaced with their respective correlation ranks from the different experimental settings in future studies.

Dataset validation is vital to successful collection of a reliable dataset.

7.4.5 Bias and Error Analysis

It is essential to thoroughly analyse potential biases or errors in the computational models. This includes assessing the performance of the models across different subgroups of the dataset and identifying any discrepancies or limitations. By understanding and addressing potential biases, one can refine the models to ensure they provide more accurate and fair assessments for individuals with different characteristics.

7.4.6 Validation and Evaluation

Rigorous validation and evaluation are crucial in refining the computational models. This involves comparing the computational assessments with expert or subjective assessments by clinicians and patients. Iterative refinement can be performed based on the feedback and insights gained during the validation process. This helps improve the models' performance and ensures their reliability in real-world clinical settings. Given that the duration of the research study was limited for academic purposes, extensive and practical validation in clinical settings is highly recommended.

7.4.7 User Interface and Integration

Refining the computational models may also involve developing a user-friendly interface or integrating existing clinical systems. This would potentially facilitate seamless adoption and utilisation of the models by clinicians in their routine practice. Considering the practical aspects, such as ease of use, interpretability of results, and integration into existing workflows, can enhance the usability and acceptance of the computational approaches.

Among the future research dimensions would be exploring feasible computational avenues for integrating CL treatment assessment into routine clinical practice. Research is needed because some assessment methods may be clinically impractical or unfriendly.

7.4.8 Ethics Clearance

In this light, more research is needed to map any ethical considerations associated with using computational techniques to assess CL treatment outcomes, from patients' and clinicians' viewpoints.

References

ABRÀMOFF, M.D., MAGALHÃES, P.J., and RAM, S.J., 2004. Image processing with imageJ. *Biophotonics International*. 11 (7), pp. 36–41.

ADOBE INC, 2021. Unit 6 : Adobe Photoshop Professional Lesson 1 : Installation and Components of Adobe Photoshop. *Preuzeto, ebookbou.edu.bd,cited by 11*. pp. 87–116.

AEFFNER, F., WILSON, K., MARTIN, N.T., BLACK, J.C., HENDRIKS, C.L.L., BOLON, B., RUDMANN, D.G., GIANANI, R., KOEGLER, S.R., KRUEGER, J., and YOUNG, G.D., 2017. The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth. *Archives of Pathology and Laboratory Medicine*. 141 (9), pp. 1267–1275.

AHMED, M.K., BUI, A.H., and TAIOLI, E., 2017. Epidemiology of Cleft Lip and Palate. In: M.A. ALMASRI, ed. *Designing Strategies for Cleft Lip and Palate Care* [online]. Rijeka: IntechOpen. Available from: https://doi.org/10.5772/67165.

AL-GHATAM, R., JONES, T.E.M., IRELAND, A.J., ATACK, N.E., CHAWLA, O., DEACON, S., ALBERY, L., COBB, A.R.M., CADOGAN, J., LEARY, S., and OTHERS, 2015. Structural outcomes in the Cleft Care UK study. Part 2: dento-facial outcomes. *Orthodontics \& Craniofacial Research*. 18, pp. 14–24.

AL-GHATAM, R., JONES, T.E.M., IRELAND, A.J., ATACK, N.E., CHAWLA, O., DEACON, S., ALBERY, L., COBB, A.R.M., CADOGAN, J., LEARY, S., WAYLEN, A., WILLS, A.K., RICHARD, B., BELLA, H., NESS, A.R., and SANDY, J.R., 2015. Structural outcomes in the Cleft Care UK study. Part 2: Dento-facial outcomes. *Orthodontics and Craniofacial Research*. 18, pp. 14–24.

AL-OMARI, I., MILLETT, D.T., AYOUB, A., BOCK, M., RAY, A., DUNAWAY, D., and CRAMPIN, L., 2003. An appraisal of three methods of rating facial deformity in patients with repaired complete unilateral cleft lip and palate. *Cleft Palate-Craniofacial Journal*. 40 (5), pp. 530–537.

ALBIOL, A., MONZO, D., MARTIN, A., SASTRE, J., and ALBIOL, A., 2008. Face recognition using HOG-EBGM. *Pattern Recognition Letters*. 29 (10), pp. 1537–1543.

ALI, M., PENA, R.M.G., RUIZ, G.O., and ALI, S., 2022. A comprehensive survey on recent deep learning-based methods applied to surgical data. [online]. Available from: http://arxiv.org/abs/2209.01435.

ALIGHIERI, C., BETTENS, K., BRUNEEL, L., D'HAESELEER, E., VAN GAEVER, E., and VAN LIERDE, K., 2021. Reliability of outcome measures to assess consonant proficiency following cleft palate speech intervention: The percentage of consonants correct metric and the probe scoring system. *Journal of Speech, Language, and Hearing Research*. 64 (6), pp. 1811–1828.

ALLOGHANI, M., AL-JUMEILY, D., MUSTAFINA, J., HUSSAIN, A., and ALJAAF, A.J., 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. pp. 3–21.

ALZUBAIDI, L., ZHANG, J., HUMAIDI, A.J., AL-DUJAILI, A., DUAN, Y., AL-

SHAMMA, O., SANTAMARÍA, J., FADHEL, M.A., AL-AMIDIE, M., and FARHAN, L., 2021. *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions* [online]. Journal of Big Data. Springer International Publishing. Available from: https://doi.org/10.1186/s40537-021-00444-8.

ALZUBI, J., NAYYAR, A., and KUMAR, A., 2018. Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*. 1142 (1), pp. 0–15.

ARHIN, K., BALDINI, I., WEI, D., RAMAMURTHY, K.N., and SINGH, M., 2021. Ground-Truth, Whose Truth? -- Examining the Challenges with Annotating Toxic Text Datasets. [online]. 1 (1), pp. 1–15. Available from: http://arxiv.org/abs/2112.03529.

ARNAUDON, A., HOLM, D.D., and SOMMER, S., 2019. A Geometric Framework for Stochastic Shape Analysis. *Foundations of Computational Mathematics* [online]. 19 (3), pp. 653–701. Available from: https://doi.org/10.1007/s10208-018-9394-z.

AVIDAN, S. and SHAMIR, A., 2007. Seam carving for content-aware image resizing. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*. 26 (3).

AYOUB, A.F., BELL, A., SIMMONS, D., BOWMAN, A., BROWN, D., LO, T.W., and XIAO, Y., 2011. 3D assessment of lip scarring and residual dysmorphology following surgical repair of cleft lip and palate: A preliminary study. *Cleft Palate-Craniofacial Journal*. 48 (4), pp. 379–387.

BAIGORRI, M., CROWLEY, C.J., SOMMER, C.L., and MOYA-GALÉ, G., 2021. Barriers and Resources to Cleft Lip and Palate Speech Services Globally: A Descriptive Study. *Journal of Craniofacial Surgery*. 32 (8), pp. 2802–2807.

BAKAKI, P., RICHARD, B., PEREIRA, E., TAGALAKIS, A., NESS, A., BEHERA, A., and LIU, Y., 2022. Key Landmarks Detection of Cleft Lip-Repaired Partially Occluded Facial Images for Aesthetics Outcome Assessment. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 718–729.

BAKAKI, P., RICHARD, B., PEREIRA, E., TAGALAKIS, A., NESS, A., and LIU, Y., 2021. Shape Analysis Approach Towards Assessment of Cleft Lip Repair Outcome. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 165–174.

BALLABENI, A., APOLLONIO, F.I., GAIANI, M., and REMONDINO, F., 2015. Advances in image pre-processing to improve automated 3d reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. 40 (5W4), pp. 315–323.

BANERJEE, S., 2021. *Mathematical modeling: models, analysis and applications*. CRC Press.

BANKHEAD, P., LOUGHREY, M.B., FERNÁNDEZ, J.A., DOMBROWSKI, Y., MCART, D.G., DUNNE, P.D., MCQUAID, S., GRAY, R.T., MURRAY, L.J., COLEMAN, H.G., JAMES, J.A., SALTO-TELLEZ, M., and HAMILTON, P.W., 2017. QuPath: Open source software for digital pathology image analysis.

Scientific Reports. 7 (1), pp. 1–7.

BARR, B., XU, K., SILVA, C., BERTINI, E., REILLY, R., BRUSS, C.B., and WITTENBACH, J.D., 2020. Towards Ground Truth Explainability on Tabular Data. [online]. (Whi). Available from: http://arxiv.org/abs/2007.10532.

BÄUMLER, M., FAURE, J.-M., BIGORRE, M., BÄUMLER-PATRIS, C., BOULOT, P., DEMATTEI, C., and CAPTIER, G., 2011. Accuracy of prenatal three-dimensional ultrasound in the diagnosis of cleft hard palate when cleft lip is present. *Ultrasound in obstetrics \& gynecology*. 38 (4), pp. 440–444.

BEHERA, A., WHARTON, Z., KEIDEL, A., and DEBNATH, B., 2020. Deep CNN, Body Pose and Body-Object Interaction Features for Drivers' Activity Monitoring. *IEEE Transactions on Intelligent Transportation Systems*. pp. 1–8.

BEKELE, K.K., EKANEM, P.E., and MEBERATE, B., 2019. Anatomical patterns of cleft lip and palate deformities among neonates in Mekelle, Tigray, Ethiopia; implication of environmental impact. *BMC pediatrics*. 19 (1), p. 254.

BELLA, H., KORNMANN, N.S.S., HARDWICKE, J.T., WALLIS, K.L., WEARN, C., SU, T.L., and RICHARD, B.M., 2016. Facial aesthetic outcome analysis in unilateral cleft lip and palate surgery using web-based extended panel assessment. *Journal of Plastic, Reconstructive and Aesthetic Surgery*. 69 (11), pp. 1537–1543.

BENALI AMJOUD, A. and AMROUCH, M., 2020. *Convolutional neural networks backbones for object detection* [online]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing. Available from: http://dx.doi.org/10.1007/978-3-030-51935-3_30.

BENGIO, Y., 2011. Deep Learning of Representations for Unsupervised and Transfer Learning. *JMLR: Workshop and Conference Proceedings*. 7, pp. 1–20.

BENNETT, C.R., BEX, P.J., BAUER, C.M., and MERABET, L.B., 2019. The Assessment of Visual Function and Functional Vision. *Seminars in Pediatric Neurology*. 31 (617), pp. 30–40.

BENNETT, K.G., RANGANATHAN, K., PATTERSON, A.K., BAKER, M.K., VERCLER, C.J., KASTEN, S.J., BUCHMAN, S.R., and WALJEE, J.F., 2018. Caregiver-Reported Outcomes and Barriers to Care among Patients with Cleft Lip and Palate. *Plastic and reconstructive surgery*. 142 (6), pp. 884e-891e.

BERHE, H.W. and MAKINDE, O.D., 2020. Computational modelling and optimal control of measles epidemic in human population. *BioSystems* [online]. 190 (July 2018), p. 104102. Available from: https://doi.org/10.1016/j.biosystems.2020.104102.

BERLIN, N.F., BERSSENBRÜGGE, P., RUNTE, C., WERMKER, K., JUNG, S., KLEINHEINZ, J., and DIRKSEN, D., 2014. Quantification of facial asymmetry by 2D analysis - A comparison of recent approaches. *Journal of Cranio-Maxillofacial Surgery* [online]. 42 (3), pp. 265–271. Available from: http://dx.doi.org/10.1016/j.jcms.2013.07.033.

BICHLER, M., 2017. Algorithms and Complexity. *Market Design*. pp. 256–267.

BLOOMFIELD, V. and LIAO, C., 2015. GLOBAL TRENDS IN THE RATE OF CLEFT LIP AND PALATE: BRIDGING THE GAP. *Paediatrics & Child Health* [online]. 20 (5), pp. E75–E104. Available from: http://ezproxy.lib.ucalgary.ca/login?url=http://search.proquest.com/docview/1691300981?accountid=9838%5Cnhttp://dc8qa4cy3n.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQ%3Acbcacomplete&rft_val_fmt=info:ofi.

BONIDIA, R.P., SAMPAIO, L.D.H., DOMINGUES, D.S., PASCHOAL, A.R., LOPES, F.M., DE CARVALHO, A.C.P.L.F., and SANCHES, D.S., 2021. Feature extraction approaches for biological sequences: A comparative study of mathematical features. *Briefings in Bioinformatics*. 22 (5), pp. 1–20.

BOZKURT, A.P. and ARAS, I., 2021. Cleft Lip and Palate YouTube Videos: Content Usefulness and Sentiment Analysis. *Cleft Palate-Craniofacial Journal*. 58 (3), pp. 362–368.

BRODLAND, G.W., 2015. How computational models can help unlock biological systems. *Seminars in Cell and Developmental Biology*. 47–48, pp. 62–73.

BRONSHTEIN, M., BLUMENFELD, I., and BLUMENFELD, Z., 1996. Early prenatal diagnosis of cleft lip and the potential impact on the number of babies with cleft lip. *British Journal of Oral and Maxillofacial Surgery*. 35 (4), p. 296.

BRUNET, D., VRSCAY, E.R., and WANG, Z., 2012. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*. 21 (4), pp. 1488–1495.

BUCH, N., VELASTIN, S.A., and ORWELL, J., 2011. A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on Intelligent Transportation Systems*. 12 (3), pp. 920–939.

BURG, M.L., CHAI, Y., YAO, C.A., MAGEE, W., and FIGUEIREDO, J.C., 2016. Epidemiology, etiology, and treatment of isolated cleft palate. *Frontiers in Physiology*. 7 (MAR), pp. 1–16.

BURGOS-ARTIZZU, X.P., PERONA, P., and DOLLAR, P., 2013. Robust face landmark estimation under occlusion. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 1513–1520.

CAI, Y., DU, W., LIN, F., YE, S., and YE, Y., 2018. Agreement of young adults and orthodontists on dental aesthetics & influencing factors of self-perceived aesthetics. *BMC Oral Health*. 18 (1), pp. 1–5.

CALDEIRA, M., MARTINS, P., COSTA, R.L.C., and FURTADO, P., 2020. Image Classification Benchmark (ICB). *Expert Systems with Applications*. 142.

CANNY, J., 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-8 (6), pp. 679–698.

CAO, X., WEI, Y., WEN, F., and SUN, J., 2014. Face alignment by explicit shape regression. *International Journal of Computer Vision*. 107 (2), pp. 177–190.

CARDOSO, J.R., PEREIRA, L.M., IVERSEN, M.D., and RAMOS, A.L., 2014. What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics*. 19 (5), pp. 27–30.

CAREY, J.C., COHEN, M.M., CURRY, C.J.R., DEVRIENDT, K., HOLMES, L.B., and VERLOES, A., 2009. Elements of morphology: Standard terminology for the lips, mouth, and oral region. *American Journal of Medical Genetics, Part A*. 149 (1), pp. 77–92.

CARVALHO, D. V., PEREIRA, E.M., and CARDOSO, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*. 8 (8), pp. 1–34.

CASH, A.C., 2012. Orthodontic treatment in the management of cleft lip and palate. *Frontiers of Oral Biology*. 16, pp. 111–123.

CHADA, A., n.d. 3D Photography of Cleft Lip: Applying Imaging Biomarkers Pre-and Post-operatively to Facilitate a Precision Medicine Approach. *Royal College of Surgeons of England* [online]. pp. 1–20. Available from: https://www.rcseng.ac.uk/-/media/files/rcs/standards-and-research/future-of-surgery/written-contributions/imaging/chada-a-3d-photography-of-cleft-lip-applying-imaging-biomarkers-pre-and-postoperatively-to-facilitat.pdf.

CHAN, H.P., SAMALA, R.K., HADJIISKI, L.M., and ZHOU, C., 2020. Deep Learning in Medical Image Analysis. *Advances in Experimental Medicine and Biology*. 1213, pp. 3–21.

CHATTERJEE, S. and SIMONOFF, J.S., 2013. *Handbook of Regression Analysis*. Handbook of Regression Analysis. Hoboken: John Wiley & Sons, Inc.

CHEN, C., SURETTE, R., and SHAH, M., 2020. Automated monitoring for security camera networks: promise from computer vision labs. *Security Journal* [online]. (0123456789). Available from: https://doi.org/10.1057/s41284-020-00230-w.

CHEN, T., 2005. Computer Vision Workload Analysis: Case Study of Video Surveillance Systems. *Understanding the Platform Requirements of Emerging Enterprise Solutions*. 9 (2), pp. 109–119.

CHEN, Z., QI, Z., MENG, F., CUI, L., and SHI, Y., 2015. Image segmentation via improving clustering algorithms with density and distance. *Procedia Computer Science* [online]. 55 (Itqm), pp. 1015–1022. Available from: http://dx.doi.org/10.1016/j.procs.2015.07.096.

CHOWDHURY, D.P., KUMARI, R., BAKSHI, S., SAHOO, M.N., and DAS, A., 2022. *Lip as biometric and beyond: a survey*. Multimedia Tools and Applications.

CLAPA, 2022. *The CRANE Database* [online]. [online]. Available from: https://www.clapa.com/treatment/research/the-crane-database/ [Accessed 7 Jul 2023].

CLEFT REGISTRY AND AUDIT NETWORK, 2020. Cleft Registry and Audit NEtwork Database 2020 Annual Report. (January 2000), pp. 1–76.

COLSTON, B.W., EVERETT, M.J., DA SILVA, L.B., OTIS, L.L., STROEVE, P., and NATHEL, H., 1998. Imaging of hard- and soft-tissue structure in the oral cavity by optical coherence tomography. *Applied Optics*. 37 (16), p. 3582.

COMANICIU, D. and MEER, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24 (5), pp. 603–619.

COSTA, B., MCWILLIAMS, D., BLIGHE, S., HUDSON, N., HOTTON, M., SWAN, M.C., and STOCK, N.M., 2023. Isolation, Uncertainty and Treatment Delays: Parents' Experiences of Having a Baby with Cleft Lip/Palate During the Covid-19 Pandemic. *Cleft Palate-Craniofacial Journal*. 60 (1), pp. 82–92.

CRANE, 2021. CRANE 2021 Annual Report: Summary of findings for patients and parents/carers. On children born with a cleft in England, Wales and Northern Ireland between January 2000 and December 2020. [online]. (January 2000). Available from: www.crane-database.org.uk.

CRERAND, C.E., RUMSEY, N., KAZAK, A., CLARKE, A., RAUSCH, J., and SARWER, D.B., 2020. Sex differences in perceived stigmatization, body image disturbance, and satisfaction with facial appearance and speech among adolescents with craniofacial conditions. *Body image*. 32, pp. 190–198.

CRESWELL, J.W., 2003. *RESEARCH DESIGN*. Awkward Dominion.

CURTIS, F.E. and NOCEDAL, J., 2018. Optimization Methods for Large-Scale Machine Learning ∗. 60 (2), pp. 223–311.

DEALL, C.E., KORNMANN, N.S.S., BELLA, H., WALLIS, K.L., HARDWICKE, J.T., SU, T.-L., and RICHARD, B.M., 2016a. Facial aesthetic outcomes of cleft surgery: assessment of discrete lip and nose images compared with digital symmetry analysis. *Plastic and Reconstructive Surgery*. 138 (4), pp. 855–862.

DEALL, C.E., KORNMANN, N.S.S., BELLA, H., WALLIS, K.L., HARDWICKE, J.T., SU, T.L., and RICHARD, B.M., 2016b. Facial Aesthetic Outcomes of Cleft Surgery: Assessment of Discrete Lip and Nose Images Compared with Digital Symmetry Analysis. *Plastic and Reconstructive Surgery*. 138 (4), pp. 855–862.

DENG, Y., LOY, C.C., and TANG, X., 2017. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine*. 34 (4), pp. 80–106.

DHARAVATH, K., TALUKDAR, F.A., and LASKAR, R.H., 2014. Improving face recognition rate with image preprocessing. *Indian Journal of Science and Technology*. 7 (8), pp. 1170–1175.

DIXON, M.J., MARAZITA, M.L., BEATY, T.H., and MURRAY, J.C., 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nature Reviews Genetics*. 12 (3), pp. 167–178.

DOLLÁR, P., WELINDER, P., and PERONA, P., 2010. Cascaded pose regression. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 1078–1085.

DOMENGHINO, A., WALBERT, C., BIRRER, D.L., PUHAN, M.A., CLAVIEN, P., and OUTCOMEMEDICINE, T., 2023. Consensus recommendations on how to assess the quality of surgical interventions. 29 (April), pp. 811–822.

DRAHANSKY, M., PARIDAH, M.., MORADBAK, A., MOHAMED, A.., OWOLABI, F. abdulwahab taiwo, ASNIZA, M., and ABDUL KHALID, S.H.., 2016. Research in Medical Imaging Using Image Processing Techniques. *Intech* [online]. i (tourism), p. 13. Available from: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics.

EL-BAZ, A., GIMEL'FARB, G., and SUZUKI, K., 2017. Machine Learning

Applications in Medical Image Analysis. *Computational and Mathematical Methods in Medicine*. 2017.

EMEKA, C.I., ADEYEMO, W.L., LADEINDE, A.L., and BUTALI, A., 2017. A comparative study of quality of life of families with children born with cleft lip and/or palate before and after surgical treatment. *Journal of the Korean Association of Oral and Maxillofacial Surgeons*. 43 (4), pp. 247–255.

ERCIYES, K., 2014. Algorithms and Complexity. *Complex Networks*. (January 2010), pp. 27–61.

ERIAN, A. and SHIFFMAN, M.A., 2011. *Advanced surgical facial rejuvenation: Art and clinical practice*. Springer Science \& Business Media.

ESHETE, M., BUTALI, A., DERESSA, W., PAGAN-RIVERA, K., HAILU, T., ABATE, F., MOHAMMED, I., DEMISSIE, Y., HAILU, A., DAWSON, D. V, DERIBEW, M., GESSESE, M., GRAVEM, P.E., and MOSSEY, P., 2017. Descriptive Epidemiology of Orofacial Clefts in Ethiopia. *The Journal of craniofacial surgery*. 28 (2), pp. 334–337.

FALCONI, L.G., PEREZ, M., and AGUILAR, W.G., 2019. Transfer Learning in Breast Mammogram Abnormalities Classification with Mobilenet and Nasnet. *International Conference on Systems, Signals, and Image Processing*. 2019-June (August 2019), pp. 109–114.

FINK, B. and NEAVE, N., 2005. The biology of facial beauty. *International Journal of Cosmetic Science*. 27 (6), pp. 317–325.

FONCUBIERTA-RODRÍGUEZ, A. and MÜLLER, H., 2012. Ground truth generation in medical imaging: A crowdsourcing-based iterative approach. *CrowdMM 2012 - Proceedings of the 2012 ACM Workshop on Crowdsourcing for Multimedia, Co-located with ACM Multimedia 2012*. pp. 9–14.

FOURCADE, A. and KHONSARI, R.H., 2019. Deep learning in medical image analysis: A third eye for doctors. *Journal of Stomatology, Oral and Maxillofacial Surgery* [online]. 120 (4), pp. 279–288. Available from: https://doi.org/10.1016/j.jormas.2019.06.002.

FRANK-ITO, D.O., CARPENTER, D.J., CHENG, T., AVASHIA, Y.J., BROWN, D.A., GLENER, A., ALLORI, A., and MARCUS, J.R., 2019. Computational Analysis of the Mature Unilateral Cleft Lip Nasal Deformity on Nasal Patency. *Plastic and Reconstructive Surgery - Global Open*. 7 (5), p. E2244.

FREDERICK, R., HOGAN, A.C., SEABOLT, N., and STOCKS, R.M.S., 2022. An Ideal Multidisciplinary Cleft Lip and Cleft Palate Care Team. *Oral Diseases*. 28 (5), pp. 1412–1417.

FREEMAN, A.K., MERCER, N.S.G., and ROBERTS, L.M., 2013. Nasal asymmetry in unilateral cleft lip and palate. *Journal of Plastic, Reconstructive and Aesthetic Surgery* [online]. 66 (4), pp. 506–512. Available from: http://dx.doi.org/10.1016/j.bjps.2012.12.001.

FRERY, A.C., 2013. Image Filtering. In: C.A.B. DE MELLO, ed. *Digital Document Analysis and Processing*. New York: Nova Science Pub Inc. pp. 55–70.

GALLOWAY, J., DAVIES, G., and MOSSEY, P., 2017. International Knowledge of

Direct Costs of Cleft Lip and Palate Treatment. *Archives of Pediatric Surgery*. 1 (1), pp. 10–25.

GALVÁN, E. and MOONEY, P., 2021. Neuroevolution in Deep Neural Networks: Current Trends and Future Challenges. *IEEE Transactions on Artificial Intelligence*. 2 (6), pp. 476–493.

GARCIA-GARCIA, A., ORTS-ESCOLANO, S., OPREA, S., VILLENA-MARTINEZ, V., and GARCIA-RODRIGUEZ, J., 2017. A Review on Deep Learning Techniques Applied to Semantic Segmentation. [online]. pp. 1–23. Available from: http://arxiv.org/abs/1704.06857.

GARCIA-MARIN, F., 2021. Access to oral & maxillofacial surgery in Sub-Saharan African countries. *Journal of Oral Biology and Craniofacial Research* [online]. 11 (4), pp. 608–611. Available from: https://doi.org/10.1016/j.jobcr.2021.09.001.

GAUR, J., GOEL, A.K., ROSE, A., and BHUSHAN, B., 2019. Emerging Trends in Machine Learning. *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2019*. (2), pp. 881–885.

GIROD, B., 2015. Image Processing and Related Fields. [online]. (22). Available from: http://fourier.eng.hmc.edu/e161/lectures/e161ch1.pdf.

GODEC, P., PANČUR, M., ILENIČ, N., ČOPAR, A., STRAŽAR, M., ERJAVEC, A., PRETNAR, A., DEMŠAR, J., STARIČ, A., TOPLAK, M., ŽAGAR, L., HARTMAN, J., WANG, H., BELLAZZI, R., PETROVIČ, U., GARAGNA, S., ZUCCOTTI, M., PARK, D., SHAULSKY, G., and ZUPAN, B., 2019. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nature Communications*. 10 (1), pp. 1–7.

GONG, X. and YU, Q., 2012. Correction of maxillary deformity in infants with bilateral cleft lip and palate using computer-assisted design. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 114 (SUPPL. 5), pp. 74–78.

GONZALEZ, R.C. and WOODS, R.E., 2002. Digital image processing. upper saddle River. *J.: Prentice Hall*.

GONZALEZ, R.C. and WOODS, R.E., 2008. *Digital Image Processing*. Third. Image and Vision Computing. New Jersey: Pearson Education, Inc.

GOODFELLOW, I., BENGIO, Y., and COURVILLE, A., 2016. *Deep Learning*. MIT Press.

GOPINATH, V.K. and MUDA, W.A.M.W., 2005. Assessment of growth and feeding practices in children with cleft lip and palate. *Southeast Asian Journal of Tropical Medicine and Public Health*. 36 (1), pp. 254–258.

GREENSPAN, H., VAN GINNEKEN, B., and SUMMERS, R.M., 2016. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*. 35 (5), pp. 1153–1159.

GREWAL, S.S., PONDURI, S., LEARY, S.D., WREN, Y., THOMPSON, J.M.D., IRELAND, A.J., NESS, A.R., and SANDY, J.R., 2021. Educational Attainment of Children Born with Unilateral Cleft Lip and Palate in the United Kingdom. *Cleft Palate-Craniofacial Journal*. 58 (5), pp. 587–596.

GRITZMAN, A.D., RUBIN, D.M., and PANTANOWITZ, A., 2015. Comparison of colour transforms used in lip segmentation algorithms. *Signal, Image and Video Processing*. 9 (4), pp. 947–957.

GU, J., YANG, X., MELLO, S. De, and KAUTZ, J., 2017. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In: *IEEE conference on computer vision and pattern recognition*. pp. 1548–1557.

HACKENBERG, B., RAMOS, M.S., CAMPBELL, A., RESCH, S., FINLAYSON, S.R.G., SARMA, H., HOWALDT, H.P., and CATERSON, E.J., 2015. Measuring and comparing the cost-effectiveness of surgical care delivery in low-resource settings: Cleft lip and palate as a model. *Journal of Craniofacial Surgery*. 26 (4), pp. 1121–1125.

HALL, B.D., GRAHAM, J.M., CASSIDY, S.B., and OPITZ, J.M., 2009. Elements of morphology: Standard terminology for the periorbital region. *American Journal of Medical Genetics, Part A*. 149 (1), pp. 29–39.

HAQUE, S., KHAMIS, M.F., ALAM, M.K., and WAN AHMAD, A.W.M., 2021. The Assessment of 3D Digital Models Using GOSLON Yardstick Index: Exploring Confounding Factors Responsible for Unfavourable Treatment Outcome in Multi-Population Children With UCLP. *Frontiers in Pediatrics*. 9 (June), pp. 1–12.

HARTIGAN, J.A. and WONG, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm Author (s): J . A . Hartigan and M . A . Wong Published by : Blackwell Publishing for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2346830. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 28 (1), pp. 100–108.

HASHIM, P.W., NIA, J.K., TALIERCIO, M., and GOLDENBERG, G., 2017. Ideals of facial beauty. *Cutis*. 100 (4), pp. 222–224.

HASHIMOTO, D.A., ROSMAN, G., RUS, D., and MEIRELES, O.R., 2018. Artificial Intelligence in Surgery: Promises and Perils. *Annals of Surgery* [online]. 268 (1), pp. 70–76. Available from: http://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC5604322&blobtype=pdf.

HASSABALLAH, M., BEKHET, S., RASHED, A.A.M., and ZHANG, G., 2019. Facial features detection and localization. In: *Studies in Computational Intelligence* [online]. Springer International Publishing. pp. 33–59. Available from: http://dx.doi.org/10.1007/978-3-030-03000-1_2.

HAVAEI, M., DAVY, A., WARDE-FARLEY, D., BIARD, A., COURVILLE, A., BENGIO, Y., PAL, C., JODOIN, P.M., and LAROCHELLE, H., 2017. Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*. 35, pp. 18–31.

HENNEKAM, R.C.M., CORMIER-DAIRE, V., HALL, J.G., MÉHES, K., PATTON, M., and STEVENSON, R.E., 2009. Elements of morphology: Standard terminology for the nose and philtrum. *American Journal of Medical Genetics, Part A*. 149 (1), pp. 61–76.

HOIEM, D., EFROS, A.A., and HEBERT, M., 2011. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*. 91 (3), pp. 328–346.

HOWARD, A.G., ZHU, M., CHEN, B., KALENICHENKO, D., WANG, W., WEYAND, T., ANDREETTO, M., and ADAM, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. [online]. Available from: http://arxiv.org/abs/1704.04861.

HU, J., JIANG, Q., CONG, R., GAO, W., and SHAO, F., 2021. Two-Branch Deep Neural Network for Underwater Image Enhancement in HSV Color Space. *IEEE Signal Processing Letters*. 28, pp. 2152–2156.

HUQH, M.Z.U., ABDULLAH, J.Y., WONG, L.S., JAMAYET, N. Bin, ALAM, M.K., RASHID, Q.F., HUSEIN, A., AHMAD, W.M.A.W., EUSUFZAI, S.Z., PRASADH, S., SUBRAMANIYAN, V., FULORIA, N.K., FULORIA, S., SEKAR, M., and SELVARAJ, S., 2022. Clinical Applications of Artificial Intelligence and Machine Learning in Children with Cleft Lip and Palate—A Systematic Review. *International Journal of Environmental Research and Public Health*. 19 (17).

ISIEKWE, G.I. and AIKINS, E.A., 2019. Self-perception of dental appearance and aesthetics in a student population. *International Orthodontics* [online]. 17 (3), pp. 506–512. Available from: https://doi.org/10.1016/j.ortho.2019.06.010.

JAGADISH CHANDRA, H., RAVI, M.S., SHARMA, S.M., and RAJENDRA PRASAD, B., 2012. Standards of Facial Esthetics: An Anthropometric Study. *Journal of Maxillofacial and Oral Surgery*. 11 (4), pp. 384–389.

JAMILIAN, A., SARKARAT, F., JAFARI, M., NESHANDAR, M., AMINI, E., KHOSRAVI, S., and GHASSEMI, A., 2017. Family history and risk factors for cleft lip and palate patients and their associated anomalies. *Stomatologija*. 19 (3), pp. 78–83.

JENNINGS, B.K., 2007. The Scientific Method. *arXiv preprint arXiv:0707.1719*.

JEONG, S.H., WOO, M.W., SHIN, D.S., YEOM, H.G., LIM, H.J., KIM, B.C., and YUN, J.P., 2022. Three-Dimensional Postoperative Results Prediction for Orthognathic Surgery through Deep Learning-Based Alignment Network. *Journal of Personalized Medicine*. 12 (6).

JIANG, R., BUSH, J.O., and LIDRAL, A.C., 2006. Development of the upper lip: Morphogenetic and molecular mechanisms. *Developmental Dynamics*. 235 (5), pp. 1152–1166.

JONKERS, H. and FRANKEN, H.M., 1996. Quantitative modelling and analysis of business processes. *Simulation in Industry* [online]. 1, pp. 175–179. Available from: http://citeseerx.ist.psu.edu.

JORDAN, M.I. and MITCHELL, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*. 349 (6245), pp. 255–260.

JOSKOWICZ, L., 2017. Computer-aided surgery meets predictive, preventive, and personalized medicine. *EPMA Journal*. 8 (1), pp. 1–4.

KAHLOOT, K.M. and EKLER, P., 2021. Algorithmic Splitting: A Method for Dataset Preparation. *IEEE Access*. 9, pp. 125229–125237.

KANCLERZ, K., GRUZA, M., KARANOWSKI, K., BIELANIEWCZ, J., MIŁKOWSKI, P., KOCON, J., and KAZIENKO, P., 2022. What if Ground Truth is Subjective? Personalized Deep Neural Hate Speech Detection. *1st Workshop on*

Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference, LREC 2022 Workshop. (June), pp. 37–45.

KAR, M., MULUK, N.B., BAFAQEEH, S.A., and CINGI, C., 2018. Is it possible to define the ideal lips? *Acta Otorhinolaryngologica Italica*. 38 (1), pp. 67–72.

KASSAM, S.N., PERRY, J.L., AYALA, R., STIEBER, E., DAVIES, G., HUDSON, N., and HAMDAN, U.S., 2020. World Cleft Coalition International Treatment Program Standards. *Cleft Palate-Craniofacial Journal*. 57 (10), pp. 1171–1181.

KATSAROS, C., 2013. Cleft lip and palate--epidemiology, aetiology and treatment (Frontiers of oral biology, Vol.16) (2012). *The European Journal of Orthodontics*. 35 (2), pp. 275–275.

KER, J., WANG, L., RAO, J., and LIM, T., 2017. Deep Learning Applications in Medical Image Analysis. *IEEE Access*. 6, pp. 9375–9379.

KHALID, S., GOLDENBERG, M., GRANTCHAROV, T., TAATI, B., and RUDZICZ, F., 2020. Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA network open*. 3 (3), p. e201664.

KHANCHEZAR, F., MORADI, N., TAHMASEBI FARD, N., LATIFI, S.M., BASSAK NEJAD, S., and HOSSEINI BEIDOKHTI, M., 2019. The Effect of Teamwork on Children With Cleft Lip and Palate and Their Mother's Quality of Life. *Cleft Palate-Craniofacial Journal*. 56 (10), pp. 1353–1358.

KHANDELWAL, K.D., VAN BOKHOVEN, H., ROSCIOLI, T., CARELS, C.E.L., and ZHOU, H., 2013. Genomic approaches for studying craniofacial disorders. In: *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. pp. 218–231.

KIMORI, Y., 2013. Morphological image processing for quantitative shape analysis of biomedical structures: Effective contrast enhancement. *Journal of Synchrotron Radiation*. 20 (6), pp. 848–853.

KLARE, B. and JAIN, A.K., 2010. On a taxonomy of facial features. *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2010*. pp. 1–8.

KLASSEN, A.F., RAE, C., WONG RIFF, K.W., BULSTRODE, N., DENADAI, R., GOLDSTEIN, J., HOL, M.L., MURRAY, D.J., BRACKEN, S., COURTEMANCHE, D.J., O'HARA, J., BUTLER, D., TASSI, A., MALIC, C.C., GANSKE, I.M., PHUA, Y.S., MARUCCI, D.D., JOHNSON, D., SWAN, M.C., BREUNING, E.E., GOODACRE, T.E., PUSIC, A.L., and CANO, S., 2021. FACE-Q Craniofacial Module: Part 1 validation of CLEFT-Q scales for use in children and young adults with facial conditions. *Journal of Plastic, Reconstructive and Aesthetic Surgery* [online]. 74 (9), pp. 2319–2329. Available from: https://doi.org/10.1016/j.bjps.2021.05.040.

KLINGENBERG, C.P., 2015. Analyzing fluctuating asymmetry with geometric morphometrics: Concepts, methods, and applications. *Symmetry*. 7 (2), pp. 843–934.

KNOOPS, P.G.M., PAPAIOANNOU, A., BORGHI, A., BREAKEY, R.W.F., WILSON,

A.T., JEELANI, O., ZAFEIRIOU, S., STEINBACHER, D., PADWA, B.L., DUNAWAY, D.J., and SCHIEVANO, S., 2019. A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific Reports* [online]. 9 (1), pp. 1–12. Available from: http://dx.doi.org/10.1038/s41598-019-49506-1.

KONDERMANN, D., 2013. Ground truth design principles: An overview. *ACM International Conference Proceeding Series*. pp. 1–4.

KORNMANN, N.S.S., TAN, R.A., MULDER, F.J., HARDWICKE, J.T., RICHARD, B.M., PIGOTT, B.B., and PIGOTT, R.W., 2019. Defining the aesthetic range of normal symmetry for lip and nose features in 5-year-old children using the computer-based program symnose. *Cleft Palate-Craniofacial Journal*. 56 (6), pp. 799–805.

KOTHARI, C.R., 2004. *Research methodology: Methods and techniques*. New Age International.

KOTSIANTIS, S.B., ZAHARAKIS, I., PINTELAS, P., and OTHERS, 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*. 160 (1), pp. 3–24.

KREUZBERGER, D., KUHL, N., and HIRSCHL, S., 2023. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*. 11 (February), pp. 31866–31879.

KUEHN, D.P. and HENNE, L.J., 2003. Speech evaluation and treatment for patients with cleft palate. *American Journal of Speech-Language Pathology*. 12 (1), pp. 103–109.

KUKHAREV, G.A. and KAZIYEVA, N., 2020. Digital Facial Anthropometry: Application and Implementation. *Pattern Recognition and Image Analysis*. 30 (3), pp. 496–511.

KUMMER, A.W., 2014. Speech evaluation for patients with cleft palate. *Clinics in Plastic Surgery* [online]. Available from: http://dx.doi.org/10.1016/j.cps.2013.12.004.

KURUVILLA, J., SUKUMARAN, D., SANKAR, A., and JOY, S.P., 2016. A review on image processing and image segmentation. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*. pp. 198–203.

LEAO, A.A.S., TOLEDO, F.M.B., OLIVEIRA, J.F., CARRAVILLA, M.A., and ALVAREZ-VALDÉS, R., 2020. Irregular packing problems: A review of mathematical models. *European Journal of Operational Research* [online]. 282 (3), pp. 803–822. Available from: https://doi.org/10.1016/j.ejor.2019.04.045.

LECUN, Y., BENGIO, Y., and HINTON, G., 2015. Deep learning. *Nature*. 521 (7553), pp. 436–444.

LEE, T.V.N., IRELAND, A.J., ATACK, N.E., DEACON, S.A., JONES, T.E.M., MATHARU, J., WILLS, A., AL-GHATAM, R., RICHARD, B.M., NESS, A.R., and SANDY, J.R., 2019. Is There a Correlation Between Nasolabial Appearance and Dentoalveolar Relationships in Patients With Repaired Unilateral Cleft Lip and

Palate? *Cleft Palate-Craniofacial Journal*. 57 (1), pp. 21–28.

LEOPOLDO-RODADO, M., PANTOJA-PERTEGAL, F., BELMONTE-CARO, R., GARCIA-PERLA, A., GONZALEZ-CARDERO, E., and INFANTE-COSSIO, P., 2021. Quality of life in early age Spanish children treated for cleft lip and/or palate: a case-control study approach. *Clinical Oral Investigations*. 25 (2), pp. 477–485.

LESLIE, E.J. and MARAZITA, M.L., 2013. Genetics of cleft lip and cleft palate. *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*. 163 (4), pp. 246–258.

LI, S., HAO, Q., GAO, G., and KANG, X., 2018. The effect of ground truth on performance evaluation of hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*. 56 (12), pp. 7195–7206.

LI, Y., CHENG, J., MEI, H., MA, H., CHEN, Z., and LI, Y., 2019. CLPNet: Cleft Lip and Palate Surgery Support With Deep Learning. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. (1), pp. 3666–3672.

LI, Y., ZENG, J., SHAN, S., and CHEN, X., 2019. Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. *IEEE Transactions on Image Processing*. 28 (5), pp. 2439–2450.

LIRA, G., KOKKINOGENIS, Z., ROSSETTI, R.J.F., MOURA, D.C., and RÚBIO, T., 2016. A computer-vision approach to traffic analysis over intersections. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. pp. 47–53.

LITJENS, G., KOOI, T., BEJNORDI, B.E., SETIO, A.A.A., CIOMPI, F., GHAFOORIAN, M., VAN DER LAAK, J.A.W.M., VAN GINNEKEN, B., and SÁNCHEZ, C.I., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*. 42 (December 2012), pp. 60–88.

LITTLE, A.C. and JONES, B.C., 2003. Evidence against perceptual bias views for symmetry preferences in human faces. *Proceedings of the Royal Society B: Biological Sciences*. 270 (1526), pp. 1759–1763.

LITTLE, A.C., JONES, B.C., and DEBRUINE, L.M., 2011. Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 366 (1571), pp. 1638–1659.

LIU, Y.H., 2018. Feature Extraction and Image Recognition with Convolutional Neural Networks. *Journal of Physics: Conference Series*. 1087 (6).

LIU, Z., LUO, P., WANG, X., and TANG, X., 2015. Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision*. 2015 Inter, pp. 3730–3738.

LONCARIC, S., 1998a. A survey of shape analysis techniques. *Pattern Recognition*. 31 (8), pp. 983–1001.

LONCARIC, S., 1998b. *A survey of shape analysis techniques*. Pattern Recognition.

LOW, D.M., BENTLEY, K.H., and GHOSH, S.S., 2020. Automated assessment of

psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*. 5 (1), pp. 96–116.

LOWE, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*. 60, pp. 91–110.

LU, D. and WENG, Q., 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*. 28 (5), pp. 823–870.

MACQUEEN, JAMES AND OTHERS, 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* [online]. 1 (14), pp. 281–297. Available from: http://books.google.de/books?hl=de&lr=&id=IC4Ku_7dBFUC&oi=fnd&pg=PA281&dq=MacQueen+some+methods+for+classification&ots=nNTcK1IdoQ&sig=fHzdVcbvmYJ-lTNHu1HncmOFOkM#v=onepage&q=MacQueen some methods for classification&f=false.

MARTÍN-MORATÓ, I. and MESAROS, A., 2021. What is the ground truth? Reliability of multi-annotator data for audio tagging. *European Signal Processing Conference*. 2021-Augus, pp. 76–80.

MASSENBURG, B.B., HOPPER, R.A., CROWE, C.S., MORRISON, S.D., ALONSO, N., CALIS, M., DONKOR, P., KRESHANTI, P., and YUAN, J., 2021. Global Burden of Orofacial Clefts and the World Surgical Workforce. *Plastic and reconstructive surgery*. 148 (4), pp. 568e-580e.

MASSIE, J.P., RUNYAN, C.M., STERN, M.J., ALPEROVICH, M., RICKERT, S.M., SHETYE, P.R., STAFFENBERG, D.A., and FLORES, R.L., 2016. Nasal septal anatomy in skeletally mature patients with cleft lip and palate. *JAMA Facial Plastic Surgery*. 18 (5), pp. 347–353.

MATHAD, V.C., SCHERER, N., CHAPMAN, K., LISS, J.M., and BERISHA, V., 2021. A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. *IEEE Transactions on Biomedical Engineering*. 68 (10), pp. 2986–2996.

MAYER, E.K., CHOW, A., VALE, J.A., and ATHANASIOU, T., 2009. Appraising the quality of care in surgery. *World Journal of Surgery*. 33 (8), pp. 1584–1593.

MCCULLOUGH, M., LY, S., AUSLANDER, A., YAO, C., CAMPBELL, A., SCHERER, S., and MAGEE, W.P., 2021. Convolutional Neural Network Models for Automatic Preoperative Severity Assessment in Unilateral Cleft Lip. *Plastic and Reconstructive Surgery*. 148 (1), pp. 162–169.

MCELROY, H., HABEL, A., MURPHY, G., and TUOHY, W., 2017. Improving early detection cleft palate_Infant2017. 13 (September), pp. 3–7.

MCKEARNEY, R.M., WILLIAMS, J. V., and MERCER, N.S., 2013. Quantitative computer-based assessment of lip symmetry following cleft lip repair. *Cleft Palate-Craniofacial Journal*. 50 (2), pp. 138–143.

MEDINA, J., COPLEY, L., DEACON, S., and VAN DER MEULEN, J., 2017. CRANE Database Annual Report 2017. *Clinical Effectiveness Unit, The Royal College of*

Surgeons of England.

MIIKKULAINEN, R., LIANG, J., MEYERSON, E., RAWAL, A., FINK, D., FRANCON, O., RAJU, B., SHAHRZAD, H., NAVRUZYAN, A., DUFFY, N., and HODJAT, B., 2018. *Evolving deep neural networks* [online]. Artificial Intelligence in the Age of Neural Networks and Brain Computing. Elsevier Inc. Available from: http://dx.doi.org/10.1016/B978-0-12-815480-9.00015-3.

MILJKOVI, O., 2009. IMAGE PRE - PROCESSING. 32.

MOHAMMAD-RAHIMI, H., NADIMI, M., ROHBAN, M.H., SHAMSODDIN, E., LEE, V.Y., and MOTAMEDIAN, S.R., 2021. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. *American Journal of Orthodontics and Dentofacial Orthopedics* [online]. 160 (2), pp. 170-192.e4. Available from: http://dx.doi.org/10.1016/j.ajodo.2021.02.013.

MOSMULLER, D., TAN, R., MULDER, F., BACHOUR, Y., DE VET, H., and DON GRIOT, P., 2016. The use and reliability of SymNose for quantitative measurement of the nose and lip in unilateral cleft lip and palate patients. *Journal of Cranio-Maxillofacial Surgery* [online]. 44 (10), pp. 1515–1521. Available from: http://dx.doi.org/10.1016/j.jcms.2016.07.022.

MOSMULLER, D.G.M., DON GRIOT, J.P.W., BIJNEN, C.L., and NIESSEN, F.B., 2013. Scoring systems of cleft-related facial deformities: A review of literature. *Cleft Palate-Craniofacial Journal*. 50 (3), pp. 286–296.

MOSMULLER, D.G.M., MAAL, T.J., PRAHL, C., TAN, R.A., MULDER, F.J., SCHWIRTZ, R.M.F., DE VET, H.C.W., BERGÉ, S.J., and DON GRIOT, J.P.W., 2017. Comparison of two- and three-dimensional assessment methods of nasolabial appearance in cleft lip and palate patients: Do the assessment methods measure the same outcome? *Journal of Cranio-Maxillofacial Surgery*. 45 (8), pp. 1220–1226.

MOSMULLER, D.G.M., MENNES, L.M., PRAHL, C., KRAMER, G.J.C., DISSE, M.A., VAN COUWELAAR, G.M., NIESSEN, F.B., and DON GRIOT, J.P.W., 2017a. The development of the cleft aesthetic rating scale: A new rating scale for the assessment of nasolabial appearance in complete unilateral cleft lip and palate patients. *Cleft Palate-Craniofacial Journal*. 54 (5), pp. 555–561.

MOSMULLER, D.G.M., MENNES, L.M., PRAHL, C., KRAMER, G.J.C., DISSE, M.A., VAN COUWELAAR, G.M., NIESSEN, F.B., and DON GRIOT, J.P.W., 2017b. The development of the cleft aesthetic rating scale: A new rating scale for the assessment of nasolabial appearance in complete unilateral cleft lip and palate patients. *Cleft Palate-Craniofacial Journal*. 54 (5), pp. 555–561.

MOSSEY, P. and EE, C., 2003. Global Registry and Database on Craniofacial Anomalies.

MOSSEY, P.A. and MODELL, B., 2012. Epidemiology of oral clefts 2012: an international perspective. *Cleft lip and palate*. 16, pp. 1–18.

MULDER, F.J., MOSMULLER, D.G.M., DE VET, R.H.C.W., and DON GRIOT, J.P.W., 2019. Aesthetics Assessment and Patient Reported Outcome of Nasolabial Aesthetics in 18-Year-Old Patients With Unilateral Cleft Lip. *Cleft Palate-Craniofacial Journal*. 56 (8), pp. 1058–1064.

MURTHY, J., 2019. Burden of Care: Management of Cleft Lip and Palate. *Indian journal of plastic surgery : official publication of the Association of Plastic Surgeons of India*. 52 (3), pp. 343–348.

NAHAI, F.R., WILLIAMS, J.K., BURSTEIN, F.D., MARTIN, J., and THOMAS, J., 2005. The Management of Cleft Lip and Palate: Pathways for Treatment and Longitudinal Assessment. *Seminars in Plastic Surgery*. 19 (04), pp. 275–285.

NAIDU, P., YAO, C.A., CHONG, D.K., and MAGEE, W.P., 2022. Cleft Palate Repair: A History of Techniques and Variations. *Plastic and Reconstructive Surgery - Global Open*. 10 (3), pp. 1–9.

NAQVI, S., HOSKENS, H., WILKE, F., WEINBERG, S.M., SHAFFER, J.R., WALSH, S., SHRIVER, M.D., WYSOCKA, J., and CLAES, P., 2022. Decoding the Human Face: Progress and Challenges in Understanding the Genetics of Craniofacial Morphology. *Annual Review of Genomics and Human Genetics*. 23, pp. 383–412.

NESS, A.R., WILLS, A.K., WAYLEN, A., AL-GHATAM, R., JONES, T.E.M., PRESTON, R., IRELAND, A.J., PERSSON, M., SMALLRIDGE, J., HALL, A.J., SELL, D., and SANDY, J.R., 2015. Centralization of cleft care in the UK. Part 6: A tale of two studies. *Orthodontics and Craniofacial Research*. 18, pp. 56–62.

NING, C. and YOU, F., 2019. Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers and Chemical Engineering* [online]. 125, pp. 434–448. Available from: https://doi.org/10.1016/j.compchemeng.2019.03.034.

NIXON, M. and AGUADO, A., 2019. *Feature extraction and image processing for computer vision*. Academic press.

NORTON, K.I., 2019. *Standards for Anthropometry Assessment*. Kinanthropometry and Exercise Physiology.

OAKDEN-RAYNER, L., 2020. Exploring Large-scale Public Medical Image Datasets. *Academic Radiology*. 27 (1), pp. 106–112.

OH, H., YANG, H., and YI, K., 2015. Learning a strategy for adapting a program analysis via bayesian optimisation. *ACM SIGPLAN Notices*. 50 (10), pp. 572–588.

OLIVEIRA, R.B., FILHO, M.E., MA, Z., PAPA, J.P., PEREIRA, A.S., and TAVARES, J.M.R.S., 2016. Computational methods for the image segmentation of pigmented skin lesions: A review. *Computer Methods and Programs in Biomedicine* [online]. 131, pp. 127–141. Available from: http://dx.doi.org/10.1016/j.cmpb.2016.03.032.

OTSU, N., 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*. 20 (1), pp. 62–66.

OWENS, J.R., JONES, J.W., and HARRIS, F., 1985. Epidemiology of facial clefting. *Archives of Disease in Childhood*. 60 (6), pp. 521–524.

PALEYES, A., URMA, R.G., and LAWRENCE, N.D., 2022. Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Computing Surveys*. 55 (6).

PALMER, S.E., SCHLOSS, K.B., and SAMMARTINO, J., 2013. Visual aesthetics and human preference. *Annual Review of Psychology*. 64, pp. 77–107.

PAN, S.J. and YANG, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*. 22 (10), pp. 1345–1359.

PANDIS, N., 2016. Multiple linear regression analysis. *American Journal of Orthodontics and Dentofacial Orthopedics* [online]. 149 (4), p. 581. Available from: http://dx.doi.org/10.1016/j.ajodo.2016.01.012.

PANNEERSELVAM, R., 2014. *RESEARCH METHODOLOGY* [online]. PHI Learning. Available from: https://books.google.co.uk/books?id=-pBeBAAAQBAJ.

PASSALIS, G., PERAKIS, P., THEOHARIS, T., and KAKADIARIS, I.A., 2011. Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 33 (10), pp. 1938–1951.

PATCAS, R., TIMOFTE, R., VOLOKITIN, A., AGUSTSSON, E., ELIADES, T., EICHENBERGER, M., and BORNSTEIN, M.M., 2019. Facial attractiveness of cleft patients: A direct comparison between artificial-intelligence-based scoring and conventional rater groups. *European Journal of Orthodontics*. 41 (4), pp. 428–433.

PENTON-VOAK, I.S., JONES, B.C., LITTLE, A.C., BAKER, S., TIDDEMAN, B., BURT, D.M., and PERRETT, D.I., 2001. Symmetry, sexual dimorphism in facial proportions and male facial attractiveness. *Proceedings of the Royal Society B: Biological Sciences*. 268 (1476), pp. 1617–1623.

PERSSON, M., SANDY, J.R., WAYLEN, A., WILLS, A.K., AL-GHATAM, R., IRELAND, A.J., HALL, A.J., HOLLINGWORTH, W., JONES, T., PETERS, T.J., and OTHERS, 2015. A cross-sectional survey of 5-year-old children with non-syndromic unilateral cleft lip and palate: the Cleft Care UK study. Part 1: background and methodology. *Orthodontics \& craniofacial research*. 18, pp. 1–13.

PIETRUSKI, P., MAJAK, M., and ANTOSZEWSKI, B., 2017. Clinically Oriented Software for Facial Symmetry, Morphology, and Aesthetic Analysis. *Aesthetic surgery journal*. 38 (1), pp. NP19–NP22.

PIETRUSKI, P., MAJAK, M., DEBSKI, T., and ANTOSZEWSKI, B., 2017. A novel computer system for the evaluation of nasolabial morphology, symmetry and aesthetics after cleft lip and palate treatment. Part 1: General concept and validation. *Journal of Cranio-Maxillofacial Surgery* [online]. 45 (4), pp. 491–504. Available from: http://dx.doi.org/10.1016/j.jcms.2017.01.024.

PIETRUSKI, P., MAJAK, M., PAWLOWSKA, E., SKIBA, A., and ANTOSZEWSKI, B., 2017. A novel computer system for the evaluation of nasolabial morphology, symmetry and aesthetics after cleft lip and palate treatment. Part 2: Comparative anthropometric analysis of patients with repaired unilateral complete cleft lip and palate and healthy i. *Journal of Cranio-Maxillofacial Surgery* [online]. 45 (4), pp. 505–514. Available from: http://dx.doi.org/10.1016/j.jcms.2017.01.022.

PIGOTT, R.W. and PIGOTT, B.B., 2010. Quantitative measurement of symmetry

from photographs following surgery for unilateral cleft lip and palate. *Cleft Palate-Craniofacial Journal*. 47 (4), pp. 363–367.

PIGOTT, R.W. and PIGOTT, B.B., 2016. Quantifying asymmetry and scar quality of children with repaired cleft lip and Palate using Symnose 2. *Cleft Palate-Craniofacial Journal*. 53 (3), pp. 298–301.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D., TEAM, R.C., and OTHERS, 2007. Linear and nonlinear mixed effects models. *R package version*. 3 (57), pp. 1–89.

PRATT, W.K., 1994. Digital Image Processing. *European Journal of Engineering Education*. 19 (3), p. 377.

PRENDERGAST, P.., 2011. Facial proportions. *Advanced Surgical Facial Rejuvenation: Art and Clinical Practice*. pp. 15–22.

PROKOPAKIS, E.P., VLASTOS, I.M., PICAVET, V., TRENITÉ, G.N., THOMAS, R., CINGI, C., and HELLINGS, P.W., 2013. The golden ratio in facial symmetry. *Rhinology*. 51 (1), pp. 18–21.

PULLI, K., BAKSHEEV, A., KORNYAKOV, K., and ERUHIMOV, V., 2012. Realtime computer vision with OpenCV. *Queue*. 10 (4), pp. 40–56.

RAGHAVAN, U., VIJAYADEV, V., RAO, D., and ULLAS, G., 2018. Postoperative Management of Cleft Lip and Palate Surgery. *Facial Plastic Surgery*. 34 (6), pp. 605–611.

RAHMANN, S., 2000. Polarization images: A geometric interpretation for shape analysis. *Proceedings - International Conference on Pattern Recognition*. 15 (3), pp. 538–542.

RAJOUB, B., 2020. Supervised and unsupervised learning. *Biomedical Signal Processing and Artificial Intelligence in Healthcare*. (April), pp. 51–89.

RAZZAK, M.I., NAZ, S., and ZAIB, A., 2018. Deep learning for medical image processing: Overview, challenges and the future. *Lecture Notes in Computational Vision and Biomechanics*. 26, pp. 323–350.

REDMON, J. and FARHADI, A., 2018. YOLOv3: An Incremental Improvement. [online]. Available from: http://arxiv.org/abs/1804.02767.

RENNELS, J.L. and CUMMINGS, A.J., 2013. Sex differences in facial scanning: Similarities and dissimilarities between infants and adults. *International Journal of Behavioral Development*. 37 (2), pp. 111–117.

RICHTER, S.R., VINEET, V., ROTH, S., and KOLTUN, V., 2016. Playing for data: Ground truth from computer games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 9906 LNCS, pp. 102–118.

RIEDLE, H., BURKHARDT, A.E., SEITZ, V., PACHALY, B., REID, R.R., LEE, J.C., and FRANKE, J.E., 2019. Design and fabrication of a generic 3D-printed silicone unilateral cleft lip and palate model. *Journal of Plastic, Reconstructive and Aesthetic Surgery* [online]. 72 (10), pp. 1669–1674. Available from: https://doi.org/10.1016/j.bjps.2019.06.030.

RITZ-TIMME, S., GABRIEL, P., TUTKUVIENE, J., POPPA, P., OBERTOVÁ, Z., GIBELLI, D., DE ANGELIS, D., RATNAYAKE, M., RIZGELIENE, R., BARKUS, A., and CATTANEO, C., 2011. Metric and morphological assessment of facial features: A study on three European populations. *Forensic Science International*. 207 (1–3), pp. 239.e1-239.e8.

ROBERTS, D.A. and YAIDA, S., 2021. *The Principles of Deep Learning Theory* [online]. Available from: deeplearningtheory.com.

ROHANI, R., ALIZADEH, S., SOBHANMANESH, F., and BOOSTANI, R., 2008. Lip Segmentation in Color Images. *2008 International Conference on Innovations in Information Technology, IIT 2008*. (April 2016), pp. 747–750.

ROKHSHAD, R., KEYHAN, S.O., and YOUSEFI, P., 2023. Artificial intelligence applications and ethical challenges in oral and maxillo-facial cosmetic surgery: a narrative review. *Maxillofacial Plastic and Reconstructive Surgery* [online]. 45 (1). Available from: https://doi.org/10.1186/s40902-023-00382-w.

ROY, H., YAMASAKI, T., and HASHIMOTO, T., 2018. Predicting image aesthetics using objects in the scene. *MMArt and ACM 2018 - Proceedings of the 2018 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia, Co-located with ICMR 2018*. pp. 14–19.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., BERG, A.C., and FEI-FEI, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* [online]. 115 (3), pp. 211–252. Available from: http://dx.doi.org/10.1007/s11263-015-0816-y.

RUSSELL, J.H.B., KIDDY, H.C., and MERCER, N.S., 2014. The use of SymNose for quantitative assessment of lip symmetry following repair of complete bilateral cleft lip and palate. *Journal of Cranio-Maxillofacial Surgery* [online]. 42 (5), pp. 454–459. Available from: http://dx.doi.org/10.1016/j.jcms.2013.05.041.

RUTANEN, K., GÓMEZ-HERRERO, G., ERIKSSON, S.-L., and EGIAZARIAN, K., 2013. A general definition of the big-oh notation for algorithm analysis. [online]. 1 (1), pp. 1–39. Available from: http://arxiv.org/abs/1309.3210.

SAE-LIM, W., WETTAYAPRASIT, W., and AIYARAK, P., 2019. Convolutional Neural Networks Using MobileNet for Skin Lesion Classification. *JCSSE 2019 - 16th International Joint Conference on Computer Science and Software Engineering: Knowledge Evolution Towards Singularity of Man-Machine Intelligence*. pp. 242–247.

SAEED, U. and DUGELAY, J.L., 2010. Combining edge detection and region segmentation for lip contour extraction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 6169 LNCS, pp. 11–20.

SAGONAS, C., TZIMIROPOULOS, G., ZAFEIRIOU, S., and PANTIC, M., 2013. 300 faces in-the-wild challenge: The first facial landmark Localization Challenge. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 397–403.

SALARI, N., DARVISHI, N., HEYDARI, M., BOKAEE, S., DARVISHI, F., and MOHAMMADI, M., 2022. Global prevalence of cleft palate, cleft lip and cleft palate and lip: A comprehensive systematic review and meta-analysis. *Journal of Stomatology, Oral and Maxillofacial Surgery* [online]. 123 (2), pp. 110–120. Available from: https://doi.org/10.1016/j.jormas.2021.05.008.

SAMEK, W., MONTAVON, G., LAPUSCHKIN, S., ANDERS, C.J., and MÜLLER, K.R., 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*. 109 (3), pp. 247–278.

SAMUEL, A.L., 1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*. 3 (3), pp. 210–229.

SANDY, J., KILPATRICK, N., and IRELAND, A., 2012. Treatment outcome for children born with cleft lip and palate. *Frontiers of Oral Biology*. 16, pp. 91–100.

SCHEIRER, W.J., ANTHONY, S.E., NAKAYAMA, K., and COX, D.D., 2014. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 36 (8), pp. 1679–1686.

SCHINDELIN, J., RUEDEN, C.T., HINER, M.C., and ELICEIRI, K.W., 2015. The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular Reproduction and Development*. 82 (7–8), pp. 518–529.

SCHINDLER, I., HOSOYA, G., MENNINGHAUS, W., BEERMANN, U., WAGNER, V., EID, M., and SCHERER, K.R., 2017. *Measuring aesthetic emotions: A review of the literature and a new assessment tool*.

SCHNEIDER, C.A., RASBAND, W.S., and ELICEIRI, K.W., 2012. NIH Image to ImageJ: 25 years of Image Analysis HHS Public Access. *Nat Methods*. 9 (7), pp. 671–675.

SCHWIRTZ, R.M.F., MULDER, F.J., MOSMULLER, D.G.M., TAN, R.A., MAAL, T.J., PRAHL, C., DE VET, H.C.W., and DON GRIOT, J.P.W., 2018. Rating nasolabial aesthetics in unilateral cleft lip and palate patients: Cropped versus full-face images. *Cleft Palate-Craniofacial Journal*. 55 (5), pp. 747–752.

SELL, D., GRUNWELL, P., MILDINHALL, S., MURPHY, T., CORNISH, T.A.O., BEARN, D., SHAW, W.C., MURRAY, J.J., WILLIAMS, A.C., and SANDY, J.R., 2001. Cleft lip and palate care in the United Kingdom - The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech outcomes. *Cleft Palate-Craniofacial Journal*. 38 (1), pp. 30–37.

SELTMANN, H.J., 2014. *Experimental Design and Analysis*. Evaluation of Human Work, Fourth Edition.

SERIES, W.P. and STERMAN, J.D., 2003. System Dynamics: Systems Thinking and Modeling for a Complex World. *European journal of computer science*. 21 (3), pp. 35–39.

SESCLEIFER, A.M., FRANCOISSE, C.A., WEBBER, J.C., RECTOR, J.D., and LIN, A.Y., 2020. Transforming assessment of speech in children with cleft palate via online crowdsourcing. *PLoS ONE*. 15 (1), pp. 1–11.

SHARMA, V.P., BELLA, H., CADIER, M.M., PIGOTT, R.W., GOODACRE, T.E.E.,

and RICHARD, B.M., 2012. Outcomes in facial aesthetics in cleft lip and palate surgery: A systematic review. *Journal of Plastic, Reconstructive and Aesthetic Surgery* [online]. 65 (9), pp. 1233–1245. Available from: http://dx.doi.org/10.1016/j.bjps.2012.04.001.

SHAYE, D., LIU, C.C., and TOLLEFSON, T.T., 2015. Cleft Lip and Palate. An Evidence-Based Review. *Facial Plastic Surgery Clinics of North America* [online]. 23 (3), pp. 357–372. Available from: http://dx.doi.org/10.1016/j.fsc.2015.04.008.

SHIER, D., BUTLER, J., and LEWIS, R., 2007. *Hole's Human Anatomy & Physiology*. Eleventh. Boston Burr Ridge,.

SHKOUKANI, M.A., CHEN, M., and VONG, A., 2013. Cleft lip - A comprehensive review. *Frontiers in Pediatrics*. 1 (DEC), pp. 1–10.

SHOBA, V.B.T. and SAM, I.S., 2020. A Hybrid Features Extraction on Face for Efficient Face Recognition. *Multimedia Tools and Applications*. 79 (31–32), pp. 22595–22616.

SINGH, H., 2019. *Practical Machine Learning and Image Processing For Facial Recognition, Object Detection, and Pattern Recognition Using Python-Himanshu Singh* [online]. Available from: www.apress.com/978-1-4842-4148-6.

SISCHO, L., PHILLIPS, C., CLOUSTON, S.A.P., and BRODER, H.L., 2016. Caregiver responses to early cleft palate care: A mixed method approach. *Health Psychology*. 35 (5), pp. 474–482.

SITZMAN, T.J. and ALLORI, A.C., 2014. M e a s u r i n g Ou t c o m e s i n C l e f t Lip a nd Palate Treatment Cleft lip Cleft palate Cleft surgery Evidence base Outcomes measurement Outcome data. p. 3.

SOH, K.B.K., 1998. Job Analysis, Appraisal and Performance Assessments of a Surgeon - A Multifaceted Approach. *Singapore Medical Journal*. 39 (4), pp. 180–185.

SOMMER, C.L., CROWLEY, C.J., MOYA-GALÉ, G., ADJASSIN, E., CACERES, E., YU, V., COSETENG-FLAVIANO, K., OBI, N., SHEERAN, P., BUKARI, B., MUSASIZI, D., and BAIGORRI, M., 2023. Global partnerships to create communication resources addressing Sustainable Development Goals 3, 4, 8, 10, and 17. *International Journal of Speech-Language Pathology* [online]. 25 (1), pp. 167–171. Available from: https://doi.org/10.1080/17549507.2022.2130430.

SRA, S., 2016. Directional Statistics in Machine Learning: a Brief Review. [online]. pp. 1–12. Available from: http://arxiv.org/abs/1605.00316.

STEIN, M.J., ZHANG, Z., FELL, M., MERCER, N., and MALIC, C., 2019. Determining postoperative outcomes after cleft palate repair: A systematic review and meta-analysis. *Journal of Plastic, Reconstructive and Aesthetic Surgery* [online]. 72 (1), pp. 85–91. Available from: https://doi.org/10.1016/j.bjps.2018.08.019.

STERMAN, J.D., 2006. Learning from evidence in a complex world. *American Journal of Public Health*. 96 (3), pp. 505–514.

SUNDERLAND, E., 1995. Anthropometry: the Individual and the Population. *Journal*

of Medical Genetics. 32 (7), p. 582.

SWANSON, J.W., YAO, C.A., AUSLANDER, A., WIPFLI, H., NGUYEN, T.H.D., HATCHER, K., VANDERBURG, R., and MAGEE, W.P., 2017. Patient Barriers to Accessing Surgical Cleft Care in Vietnam: A Multi-site, Cross-Sectional Outcomes Study. *World Journal of Surgery*. 41 (6), pp. 1435–1446.

SZELISKI, R., 2011. Computer Vision: Algorithms and Applications. *Choice Reviews Online*. 48 (09), pp. 48-5140-48–5140.

TABIA, H. and LAGA, H., 2017. Learning shape retrieval from different modalities. *Neurocomputing*. 253, pp. 24–33.

TAIB, B.G., TAIB, A.G., SWIFT, A.C., and VAN EEDEN, S., 2015. Cleft lip and palate: Diagnosis and management. *British Journal of Hospital Medicine*. 76 (10), pp. 584–591.

TAJBAKHSH, N., SHIN, J.Y., GURUDU, S.R., HURST, R.T., KENDALL, C.B., GOTWAY, M.B., and LIANG, J., 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*. 35 (5), pp. 1299–1312.

TAKECHI, M., ADACHI, N., HIRAI, T., KURATANI, S., and KURAKU, S., 2013. The Dlx genes as clues to vertebrate genomics and craniofacial evolution. *Seminars in Cell and Developmental Biology* [online]. 24 (2), pp. 110–118. Available from: http://dx.doi.org/10.1016/j.semcdb.2012.12.010.

TALEBI, H. and MILANFAR, P., 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*. 27 (8), pp. 3998–4011.

TAN, C., SUN, F., KONG, T., ZHANG, W., YANG, C., and LIU, C., 2018. A survey on deep transfer learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 11141 LNCS, pp. 270–279.

THOMPSON, J.A., HEATON, P.C., KELTON, C.M.L., and SITZMAN, T.J., 2017. National estimates of and risk factors for inpatient revision surgeries for orofacial clefts. *Cleft Palate-Craniofacial Journal*. 54 (1), pp. 60–69.

TIWARI, V., DEYAL, N., and BISHT, N.S., 2020. Mathematical Modeling Based Study and Prediction of COVID-19 Epidemic Dissemination Under the Impact of Lockdown in India. *Frontiers in Physics*. 8 (November), pp. 1–8.

UNPINGCO, J., 2019. *Python for Probability , Statistics , and Machine Learning* [online]. Second. Gewerbestrasse: Springer Nature. Available from: https://doi.org/10.1007/978-3-030-18545-9.

VAFAEI, N., RIBEIRO, R.A., and CAMARINHA-MATOS, L.M., 2018. Data normalisation techniques in decision making: case study with TOPSIS method. *Int. J. Information and Decision Sciences* [online]. 10 (1), pp. 27–29. Available from: http://www.ca3-uninova.org.

VIDAL, M. and AMIGO, J.M., 2012. Pre-processing of hyperspectral images. Essential steps before image analysis. *Chemometrics and Intelligent Laboratory Systems* [online]. 117, pp. 138–148. Available from: http://dx.doi.org/10.1016/j.chemolab.2012.05.009.

VIOLA, P. and JONES, M., 2001. Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1, pp. 11–18.

VO, D.M. and LE, T.H., 2016. Deep generic features and SVM for facial expression recognition. *NICS 2016 - Proceedings of 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science*. pp. 80–84.

VUJOVIĆ, Ž., 2021. Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*. 12 (6), pp. 599–606.

WANG, B., JIN, S., YAN, Q., XU, H., LUO, C., WEI, L., ZHAO, W., HOU, X., MA, W., XU, Z., ZHENG, Z., SUN, W., LAN, L., ZHANG, W., MU, X., SHI, C., WANG, Z., LEE, J., JIN, Z., LIN, M., JIN, H., ZHANG, L., GUO, J., ZHAO, B., REN, Z., WANG, S., XU, W., WANG, X., WANG, J., YOU, Z., and DONG, J., 2021. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. *Applied Soft Computing*. 98.

WANG, D., ZHANG, B., ZHANG, Q., and WU, Y., 2023. Global, regional and national burden of orofacial clefts from 1990 to 2019: an analysis of the Global Burden of Disease Study 2019. *Annals of medicine* [online]. 55 (1), p. 2215540. Available from: https://doi.org/10.1080/07853890.2023.2215540.

WANG, J., SUN, K., CHENG, T., JIANG, B., DENG, C., ZHAO, Y., LIU, D., MU, Y., TAN, M., WANG, X., LIU, W., and XIAO, B., 2020. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 43 (10), pp. 3349–3364.

WANG, J., THORNTON, J.C., KOLESNIK, S., and PIERSON, R.N., 2000. Anthropometry in body composition. An overview. *Annals of the New York Academy of Sciences*. 904, pp. 317–326.

WANG, N., GAO, X., TAO, D., YANG, H., and LI, X., 2018. Facial feature point detection: A comprehensive survey. *Neurocomputing*. 275, pp. 50–65.

WANG, R., LEI, T., CUI, R., ZHANG, B., MENG, H., and NANDI, A.K., 2022. Medical image segmentation using deep learning: A survey. *IET Image Processing*. 16 (5), pp. 1243–1267.

WANG, Y., LI, J., XU, Y., HUANG, N., SHI, B., and LI, J., 2020. Accuracy of virtual surgical planning-assisted management for maxillary hypoplasia in adult patients with cleft lip and palate. *Journal of Plastic, Reconstructive and Aesthetic Surgery*. 73 (1), pp. 134–140.

WANG, Z., BOVIK, A.C., SHEIKH, H.R., and SIMONCELLI, E.P., 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*. 13 (4), pp. 600–612.

WANG, Z., WANG, E., and ZHU, Y., 2020. *Image segmentation evaluation: a survey of methods* [online]. Artificial Intelligence Review. Springer Netherlands. Available from: https://doi.org/10.1007/s10462-020-09830-9.

WASKOM, M., 2021. Seaborn: Statistical Data Visualization. *Journal of Open Source Software*. 6 (60), p. 3021.

WEI, W., HO, E.S.L., MCCAY, K.D., DAMAŠEVIČIUS, R., MASKELIŪNAS, R., and ESPOSITO, A., 2022. Assessing Facial Symmetry and Attractiveness using Augmented Reality. *Pattern Analysis and Applications* [online]. 25 (3), pp. 635–651. Available from: https://doi.org/10.1007/s10044-021-00975-z.

WEYL, H., 1952. *Symmetry*. Princeton University Press.

WHARTON, Z., BEHERA, A., LIU, Y., and BESSIS, N., 2021. Coarse Temporal Attention Network (CTA-Net) for Driver's Activity Recognition. pp. 1279–1289.

WU, S.-T., SILVA, A.C.G. da, and MÁRQUEZ, M.R.G., 2004. The Douglas-peucker algorithm: sufficiency conditions for non-self-intersections. *Journal of the Brazilian Computer Society*. 9 (3), pp. 67–84.

XIA, Y., NIE, L., ZHANG, L., YANG, Y., HONG, R., and LI, X., 2016. Weakly Supervised Multilabel Clustering and its Applications in Computer Vision. *IEEE Transactions on Cybernetics*. 46 (12), pp. 3220–3232.

XU, M., CHEN, F., LI, L., SHEN, C., LV, P., ZHOU, B., and JI, R., 2021. Bio-Inspired Deep Attribute Learning towards Facial Aesthetic Prediction. *IEEE Transactions on Affective Computing*. 12 (1), pp. 227–238.

XU, W., FU, Y.L., and ZHU, D., 2023. ResNet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine* [online]. 240, p. 107660. Available from: https://doi.org/10.1016/j.cmpb.2023.107660.

XU, Y., MO, T., FENG, Q., ZHONG, P., LAI, M., and CHANG, E.I.C., 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. (1), pp. 1626–1630.

YOSINSKI, J., CLUNE, J., NGUYEN, A., FUCHS, T., and LIPSON, H., 2015. Understanding Neural Networks Through Deep Visualization. [online]. Available from: http://arxiv.org/abs/1506.06579.

YU, C., GAO, C., WANG, J., YU, G., SHEN, C., and SANG, N., 2020. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. *arXiv*.

YU, C., WANG, J., PENG, C., GAO, C., YU, G., and SANG, N., 2018. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 11217 LNCS, pp. 334–349.

ZHANG, A., LIPTON, Z.C., LI, M.U., and ALEXANDER, J., 2021. *Dive into Deep Learning* [online]. Available from: https://arxiv.org/abs/2106.11342.

ZHANG, Q., YUE, Y., SHI, B., and YUAN, Z., 2019. A Bibliometric Analysis of Cleft Lip and Palate-Related Publication Trends From 2000 to 2017. *Cleft Palate-Craniofacial Journal*. 56 (5), pp. 658–669.

ZHANG, X., SHEN, P., LUO, L., ZHANG, L., and SONG, J., 2012. Enhancement and noise reduction of very low light level images. *Proceedings - International Conference on Pattern Recognition*. (Icpr), pp. 2034–2037.

ZHOU, S.K., GREENSPAN, H., DAVATZIKOS, C., DUNCAN, J.S., VAN GINNEKEN, B., MADABHUSHI, A., PRINCE, J.L., RUECKERT, D., and SUMMERS, R.M., 2021. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises. *Proceedings of the IEEE*. 109 (5), pp. 820–838.

ZHOU, S.K., GREENSPAN, H., and SHEN, D., 2017. *Deep Learning for Medical Image Analysis*. First Edit. Deep Learning for Medical Image Analysis. Elsevier Inc.

ZHU, M., SHI, D., ZHENG, M., and SADIQ, M., 2019. Robust facial landmark detection via occlusion-adaptive deep networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2019-June, pp. 3481–3491.

ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H., and HE, Q., 2021. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*. 109 (1), pp. 43–76.

ZORICH, V.A. and PANIAGUA, O., 2016. *Mathematical analysis II*. Springer.

Appendix A: Key Landmarks Detection using Deep Learning in Partially Occluded Images

A1. Introduction

Facial landmark detection is the process of identifying and localizing key points or landmarks on a human face. These landmarks are specific points that serve as reference points to accurately describe the facial geometry and expressions. They are crucial for various computer vision applications, such as face alignment, emotion recognition, facial expression analysis, and face morphing. Face landmark detection can also be used for face features regeneration assessment (Sagonas et al., 2013).

Key facial landmarks have been discussed by (Klare and Jain, 2010) and typically include:

1. Eye landmarks: These include points for the corners of the eyes, the pupil centers, and the eye contours. They help in determining eye shape, gaze direction, and blinking patterns.
2. Nose landmarks: Nose tip, nostrils, and nasal bridge points are essential for assessing nose shape, size, and orientation.
3. Mouth landmarks: Points for the corners of the mouth, upper and lower lips, and the center of the mouth help in detecting mouth opening, smiles, and expressions.
4. Cheek and jawline landmarks: These points represent the contour of the cheek and jawline, providing information about facial structure and symmetry.

In this PhD research, the use anonymised facial images, therefore, features in categories 2 and 3 are detected mainly by most means.

Facial landmark detection can be performed using various techniques such as Traditional feature-based methods, Deep learning-based methods, or 3D facial landmark detection (Sagonas et al., 2013, Cao et al., 2014). DUE TO ITS ROBUST NATURE, the YOLO deep learning framework takes precedence in this work.

Facial landmark detection finds wide applications in computer graphics, augmented reality, medical imaging, and biometrics. It is a fundamental step in many facial

analysis tasks and enables a deeper understanding of facial expressions, emotions, and appearances studies.

A2. Background

Quantitative and aggregated numeric results are desirable and acceptable outcomes for most experimental assessment studies. This writeup aims to create a deep learning pipeline that mitigates the evident biased verbosity of previous attempts. Deep learning-based facial features detection is an alternative method to the one presented in Chapter 5. To appreciate an end-to-end pipeline for assessment outcome, an ablation study and analysis is required, and so presented in this section.

If features are detected from the partial facial images, then a region of interest is highly likely apparent, facilitating the generation of continuous and categorical parameters. Among those parameters include facial symmetry; inclination/alignment of inner eye corners, philtrum ridges, nasal alares etc; Euclidian distances between key features; and elevation of key features from one another. Any scoring approach requires presenting these parameters for an aggregated approach with minimal human intervention/input. The challenge is that many of such primary features might have been distorted or not fully restored during the surgical treatment. Feature localisation, thus quantitative parametrisation is highly dynamic, complicated, radically stochastic, and potentially irreproducible. This may cause mathematical limitations for a linear or non-linear parametric functional model (Pandis, 2016). Therefore, a study fronted by deep learning techniques is a plausible alternative due to their capability to simultaneously manage models with multiple parameters of stochastic nature. Specifically, deep learning techniques present opportunities for optimisation and appropriate (hyper)parameter tuning (Ning and You, 2019).

Supervised machine learning requires human experts to annotate potential key feature locations. Figure A.1 demonstrates the human ability to locate features from facial appearances outcomes following surgery using an annotation software in preparation for supervised learning.

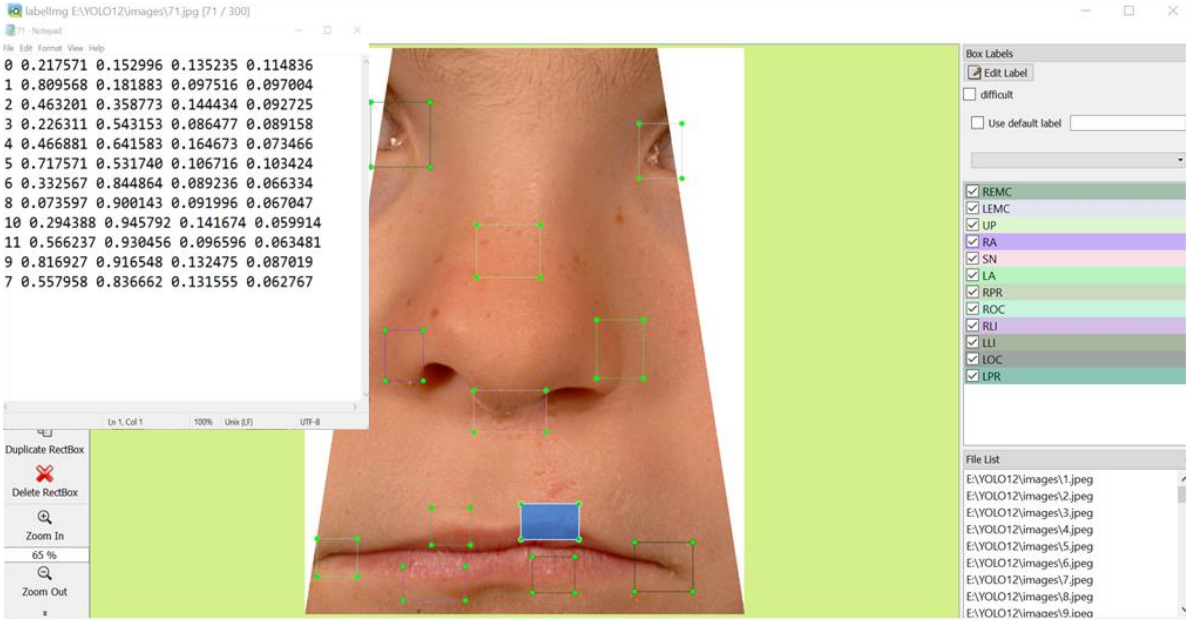

Figure A. 1: Annotation window in LabelImg software. 12 landmarks are annotated. But could have been more. A bounding box is a better feature attributes detector.

Figure A.2 shows the ground truth that serves as model input.

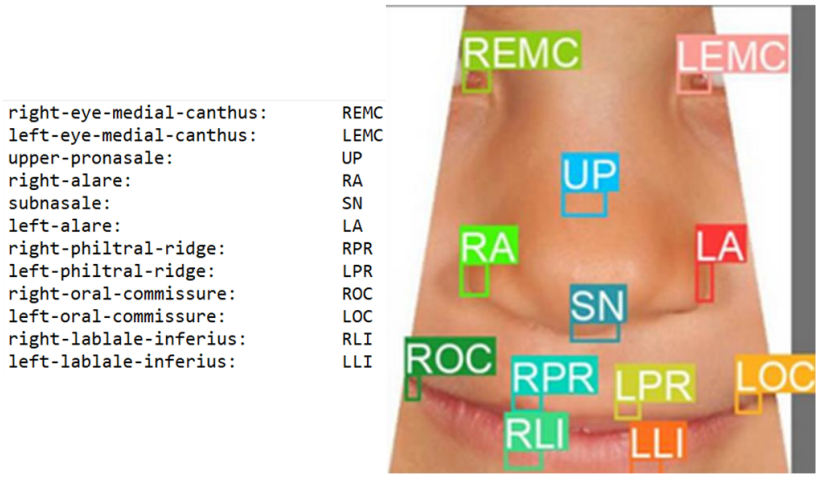

Figure A. 2: Ground truth sample for input into a model. Each key landmark has got a label to guide deep network learning and generate a landmark detection model. 12 landmarks are visualised as annotated.

Using the You Only Look Once (YOLOv3)-based deep learning framework (Figure A.3) (Redmon and Farhadi, 2018), a model was trained.

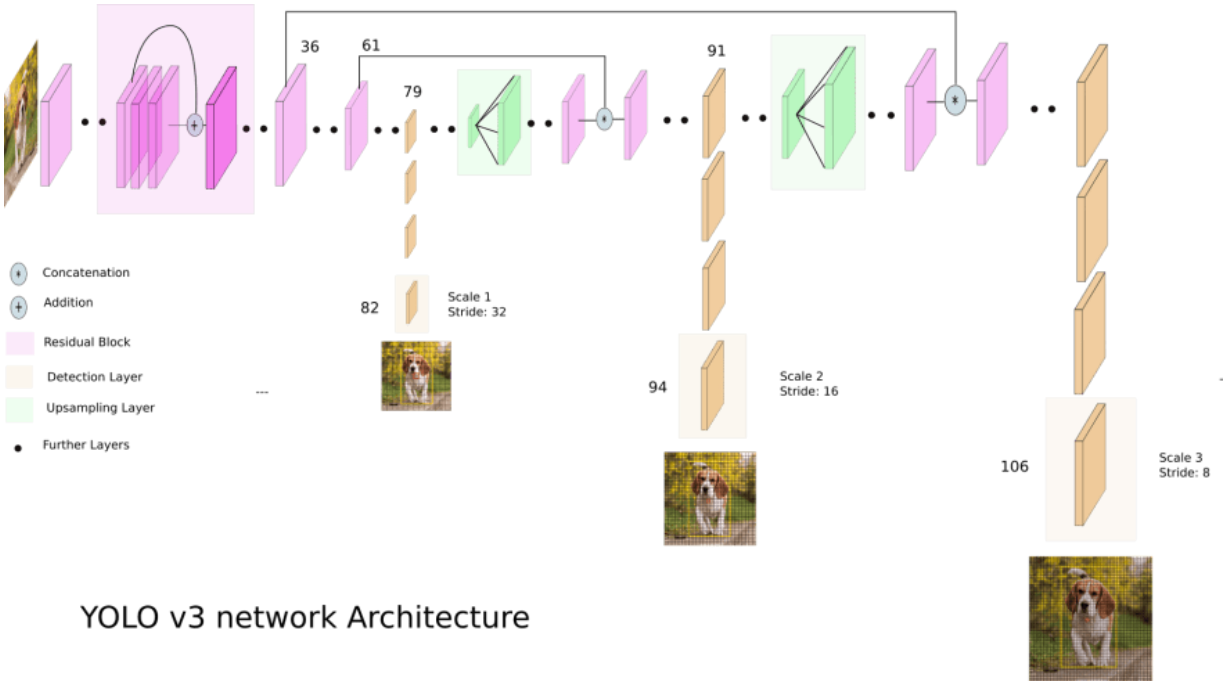

Figure A. 3: How YOLO works (Redmon and Farhadi, 2018). Several residual blocks, detection layers and upsampling layers are modelled. This being only a framework, several other layers may be incorporated.

One of the outputs from the trained model is in Figure A.4.

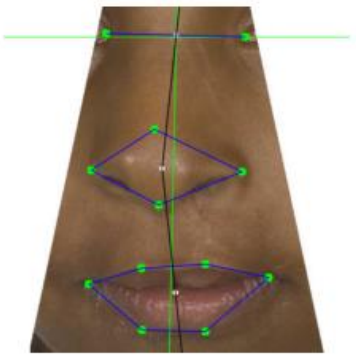

Figure A. 4: Landmarks detected from a trained model (green circles). Detection of key regions of interest, if necessary (blue shapes). Potential symmetric axis based on features (vertical black line).

Different facial images can be used in Figures A.1, A.2, A.3 and A.4 to demonstrate the robust nature of deep learning techniques.

YOLO 'co-works' with Darknet as its implementation back bone. Whereas YOLO can be modified to suit a specific dataset and its parameters, Darknet has been introduced to two architectures Darknet-19 and Darknet-53. It has formed the backbone of several versions, with YOLOv8 as the latest release. Because some researchers and pioneers had reservations around the society impact of their fast object detection framework (YoloV1), some publications were withheld when YoloV3 was released in

2018. Despite this, GitHub source code repositories[9] have been used to study the underlying frameworks.

YOLO is a convolutional neural network (CNN), a traditionally known deep neural network (DNN) suited for exploitation and analysis of (medical) images and videos (Zhou, Greenspan and Shen, 2017).

The reported preliminary findings in Section A4 below are premised on YOLO producing state of the art results for object detection with less computational resources for model training, sometimes outcompeting human experts (Benali Amjoud and Amrouch, 2020). Much as training these models is technically tricky and mostly resource intensive, a YOLO-based framework used in our work is relatively lighter. The base uses a faster and more efficient single spine/single forward propagation approach.

A3. Experimental setup

Once annotation is complete, dataset diversity is achieved through augmentation. This may serve to optimise model convergence during training. A setup in Figure A1 is created using LabelImg, a package of Label Studio[10]. After annotation, a text file containing the bounding box coordinates of each landmark is generated. The process is repeated for every image in the dataset. Normalisation is carried out about every image width and height to ensure uniformity of the bounding box metrics so that they fall between 0 and 1. This ensures that the bounding box (x, y) coordinates are offsets of a particular grid cell location.

$$B = Normalized\ [\ x_{center}, y_{center}, w, h] \quad \text{(Equation 22)}$$

Where:

x and y are pixel values marking the bounding box center from the x-axis and y-axis respectively,

w and h are the width and height of bounding box, respectively.

Sometimes B (from Equation 22) can be a small value and should be scaled in line with the image dimensions with consistency and consideration of other features/objects in the same image. A value in the range of 2 and 4 was used to scale

[9] https://github.com/AlexeyAB

[10] https://labelstud.io/

B. A visualisation of B on the dataset yielded Figure A.5b. In contrast, Figure A.5a represents visualisation of B on the dataset after augmentation by the properties in Table A.1.

Table A. 1: Pre-processing and Augmentation properties.

Grayscale	25% of the dataset
Luminosity Exposure	-20% and + 20%
Gaussian Blur	Up to 4.75 pixels
Resizing	• 416 by 416 with vertical padding • 312 by 416 – regular face

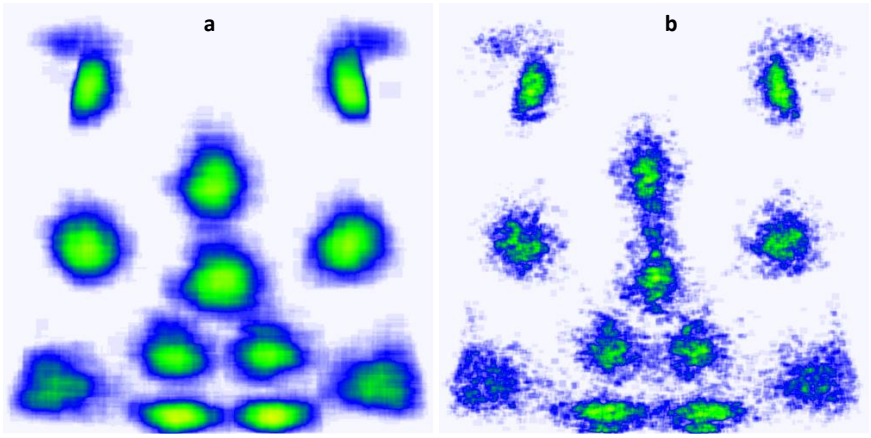

Figure A. 5: Dataset annotation and normalisation. Left: After augmentation, the average heatmap annotation is generated. Right: annotation heatmap generated before augmentation was applied. Therefore, augmentation is useful for a successful model building.

In Figure A.5, each blob represents the 12 landmarks heatmap generated during annotation. Figure A.5 (a) is dataset annotation heatmap after augmentation. An indication of a better training dataset. Figure A.5 (b) is dataset annotation heatmap for the dataset before augmentation.

Following augmentation, 3000 images were generated from the original CCUK dataset of 250 images. Poorly annotated, and unreadable photos were removed, leaving a total of 2857 images.

Figures A.6 and A.7 visualise[11] the same architecture for the YOLO-based CNN framework used to train the object detection model. It is this model that we employ in features localization.

[11] http://alexlenail.me/NN-SVG/LeNet.html

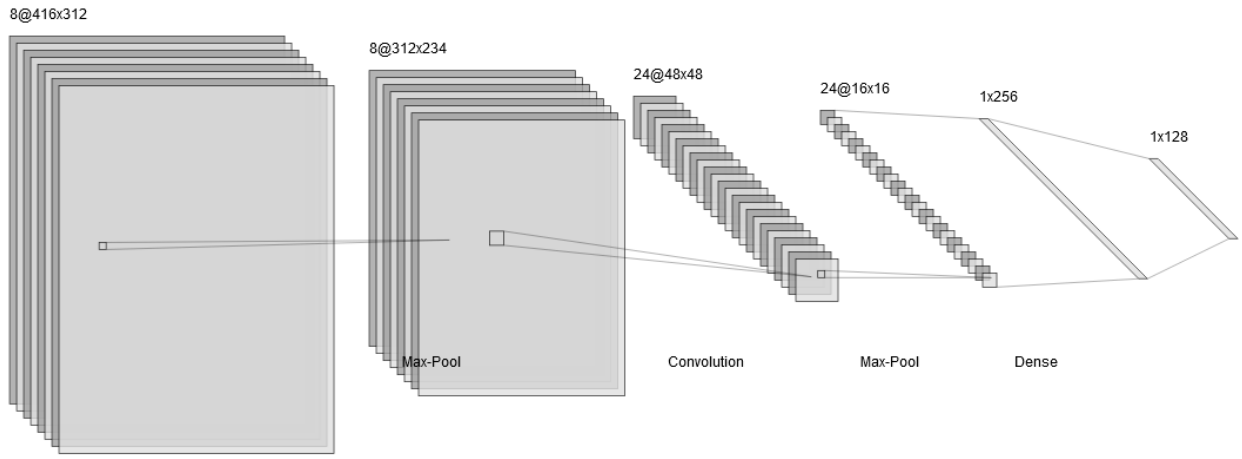

Figure A. 6: Adjusted YoloV3 Architecture using LeNet Style

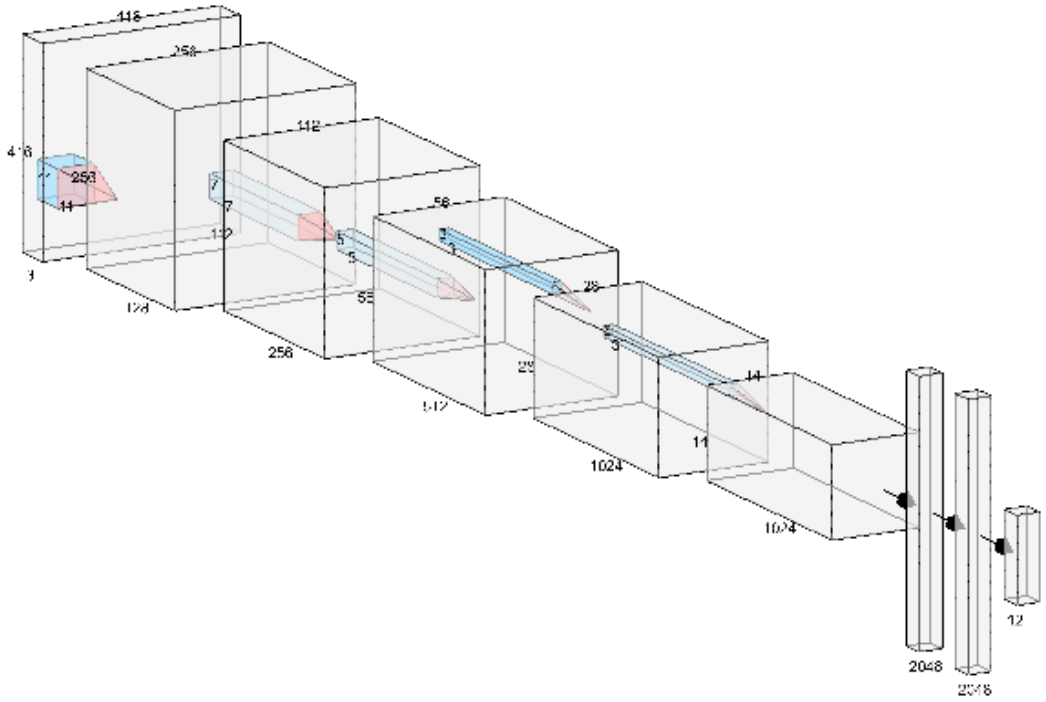

Figure A. 7: Adjusted YoloV3 Architecture using AlexNet Style – with 12 outputs.

The network uses features from the entire image to predict each landmark's bounding box. It also indicates all bounding boxes across all classes for an image simultaneously. This means the network reasons globally about the entire image and learns about all the possible twelve (12) objects/critical landmarks in the image.

There is unification of the separate components of object detection into a single neural network. The network uses features from the entire image to predict each landmark's bounding box. It also indicates all bounding boxes across all classes for an image

simultaneously. This means the network reasons globally about the entire image and learns about all the possible twelve (12) objects/landmarks in the image.

Training and detection often require fine-grained visual information, so network input size is considerably reduced to an acceptable size of 416 by 416, This is further reduced to 312 by 416 due to the 0.25 width padding and 0.75 of the facial height as the width. This is seconded by the approximate rational human facial morphology of 3:4 rule of facial width to size. An ambitious filter of 11 by 11 pixels (default = 3x3 filter) is used on the input image. There are 5 hidden layers and 2 fully connected layers before converging into of 12 nodes (Figure A.7).

A4. Results

The model was trained to predict multiple bounding boxes per grid cell, if any existed. At training time, we only want one bounding box predictor responsible for each object. One predictor is assigned to predict an object based on which prediction has the highest current IOU with the ground truth. Whereas this leads to specialization between the bounding box predictors, special attention is paid to potential bounding boxes overlap for multiple objects in the same image. Figure A.8 presents the model training results. The smooth loss curve following training indicates the model's learning was good.

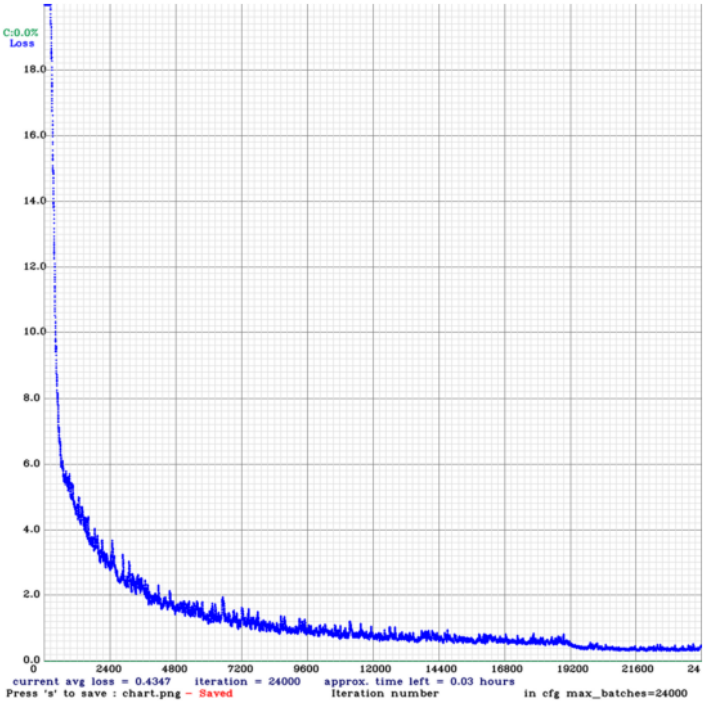

Figure A. 8: Model training results using modified YOLO framework, shows great learning potential. A smooth curve is a welcome outcome.

Following the training, some tensor board outcomes are depicted in Figure A.9

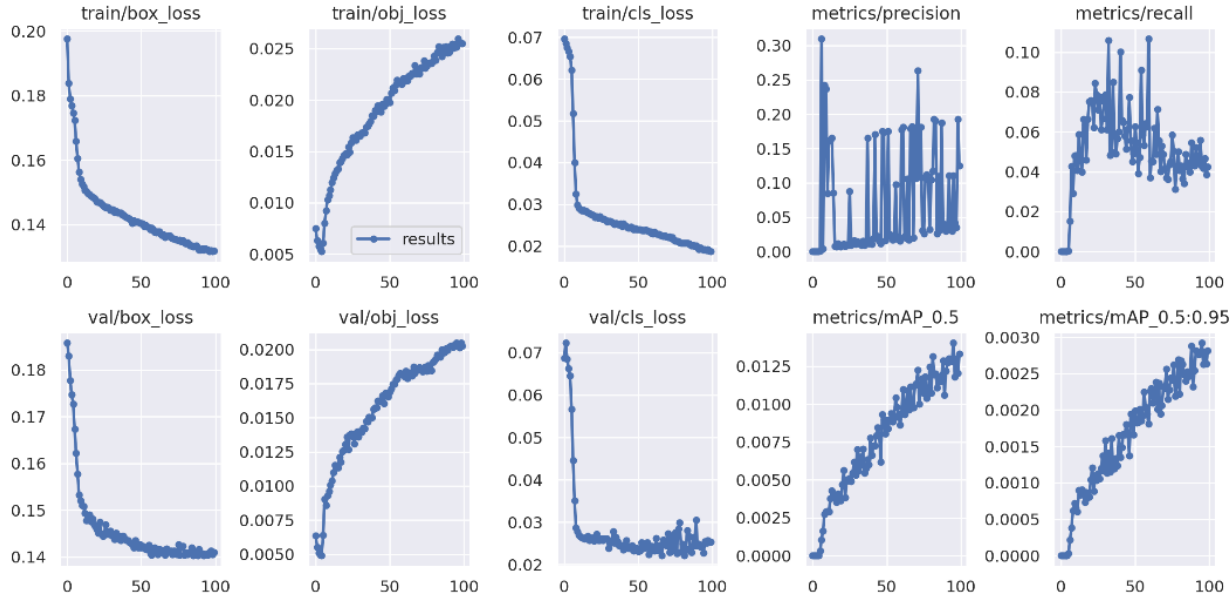

Figure A. 9: Tensor board results from training the modified YOLO. Training was conducted over 100 epochs. From left to right: 1st column: Localisation losses from bounding box coordinate prediction are not changing uniformly. 2nd column: Near matching bounding box prediction for object capture. 3rd column: Model losses from the classification of the objects. 4th and 5th columns represent precisions/average precisions and recall/average precisions respectively, at different intervals.

For each object whose label is read, there is comparison of the actual bounding box between the training dataset and the validation dataset. We used the bounding box for label localisation, the accuracy is calculated. Loss, mean absolute precision are some of the tensor board metrics returned.

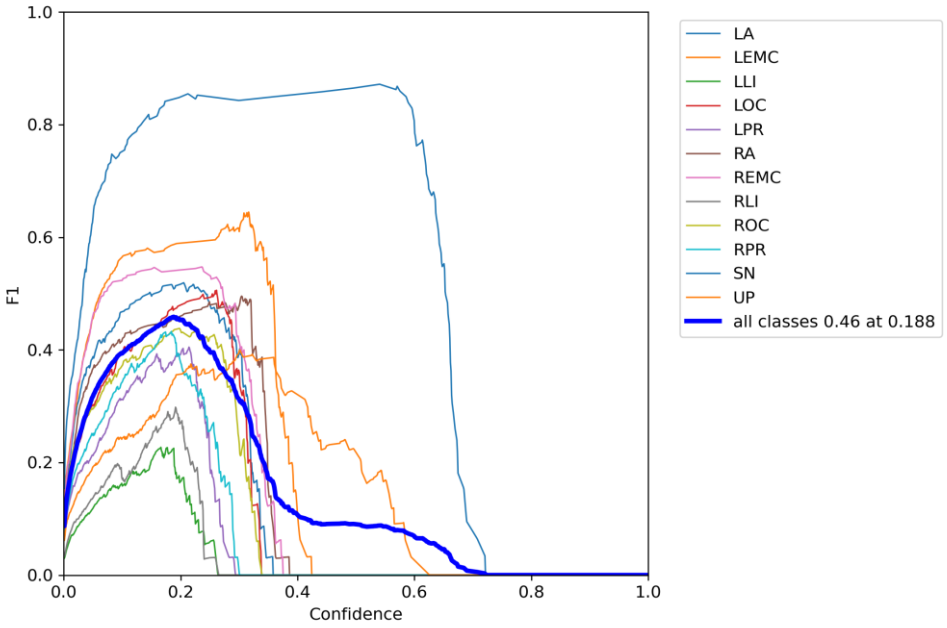

Figure A. 10: F1-Score vs confidence map. Each of the 12 landmarks is considered as a class/ unique category. Best predicted class is RPR (F1>0.8, CI=0.75) and worst predicted class is LLI (F1 < 0.25, CI < 0.25). The bold blue curve is the average for all the classes (F1<0.5, CI<0.75)

Figure A.10 shows the label detection confidence map, where F1-score has been used. The map is plotted for each label, from which an average F1 score is calculated. The best detected label is LA while the worst is LLI.

In conclusion, not all the landmarks could be detected on most images. In fact, the test dataset presented about 90% feature detection accuracy, falling below the expectations. Given a diverse dataset, medical image analysis studies acceptable accuracy ranges between 94 and 99 % (Khalid et al., 2020). The next step is to generate continuous and categorical parameters once the landmarks can be detected. The parameters would consequently serve as inputs for the regression model to output a desired appearance score.

Continuous, categorical, and deeper features/parameters are learnt in a single-spine solution and directly mapped to a single score. A robust regression decay should follow feature extraction in an end-to-end network, not generate parameters for input into an external system.