Development of AI-Assisted Decision Support Tools for Plastic Surgery

Hassan Awad

Thesis submitted to the University of Ottawa in partial fulfillment of requirements for the degree of

Master of Applied Science

in Electrical and Computer Engineering

uOttawa

Ottawa Carleton Institute for Electrical and Computer Engineering

University of Ottawa
Ottawa, Ontario

Abstract

Accurate localization of the Nipple-Areola Complex (NAC) is essential for achieving optimal aesthetic outcomes in chest reconstruction surgeries such as gender-affirming surgeries and post-mastectomy procedures. Traditional methods for NAC placement rely heavily on surgeons' experience and judgement, which are subject to variability and limited reproducibility. This thesis presents a machine learning-based decision support framework for automated NAC localization using pose estimation and regression modeling. A dataset of 102 clinical images from 34 male subjects was processed using OpenPose, an open-source pose estimation algorithm, to extract upper-body landmarks. Normalized distances between selected landmarks were used as input features for six regression models: Decision Tree Regressor, Random Forest Regressor, CatBoost Regressor, Multilayer Perceptron (MLP) Regressor, Support Vector Regressor (SVR), and Linear Regressor. Among these, the Linear Regression model achieved the lowest mean absolute percentage error (MAPE) of 0.06% for normalized distance prediction and 0.69% for average predicted NAC position. To further enhance accuracy, a tilt correction algorithm was implemented, reducing the MAPE from 1.2% to 0.75% (right NAC) and from 0.99% to 0.63% (left NAC) in cases with subject tilt.

To address limitations in dataset size and diversity, a second phase of research introduced a synthetic image generation pipeline using a Large Language Model (Llama 3.2). A prompt engineering strategy enabled the creation of demographically varied anatomically realistic images. Three experiments evaluated the impact of synthetic data on model performance. Results demonstrated that moderate augmentation improved performance and generalization, particularly when tested on demographically diverse or real-world images not seen during training.

This work illustrates the potential of AI-assisted decision support tools in plastic surgery, specifically in standardizing NAC localization through data-driven modeling. The integration of synthetic datasets further addresses challenges in data scarcity and enhances the generalizability of clinical machine learning applications.

Table of Contents

List of Figures

Acknowledgments

First and foremost, I would like to thank **God** for granting me the strength, patience, and perseverance to complete this thesis. This journey would not have been possible without His guidance and blessings.

I am deeply grateful to my **supervisor, Professor Hussein Al Osman**, for his continuous support, insightful feedback, and encouragement throughout this research. His mentorship not only guided the technical direction of this work but also inspired me to grow as a researcher and thinker. Working with him has been both a privilege and a learning experience that I will carry forward in my academic and professional life.

I would also like to sincerely thank my **co-supervisor, Professor Natalie Baddour**, for her thoughtful guidance, constructive criticism, and unwavering support throughout my thesis. Her contributions have been invaluable to the depth and quality of this work.

To my **family in Lebanon**, your love and prayers have been my foundation. Though distance separated us, your unwavering emotional support has always been with me. Thank you for believing in me and for motivating me to reach higher.

This thesis is not only the result of academic effort but also the product of countless sacrifices, encouragement, and faith from those I hold dear. To all who supported me in this journey, thank you from the bottom of my heart.

Chapter 1. Introduction

Plastic surgery encompasses a broad range of reconstructive and aesthetic procedures that demand meticulous planning and intraoperative decision-making to achieve optimal outcomes. In recent years, the proliferation of medical data and advances in computational methods have opened new avenues for clinical decision support. In particular, machine learning, a subfield of artificial intelligence (AI), is increasingly being explored as a tool to assist surgeons in interpreting complex clinical data and guiding surgical decisions[1], [2]. Researchers have highlighted the promise of machine learning-based, data-driven approaches in plastic and reconstructive surgery, suggesting that these techniques can help identify patterns in large datasets and improve key clinical decision processes beyond the limits of human judgment[3], [4]. Early applications of machine learning in plastic surgery have already demonstrated success in several domains. For example, algorithms have been developed to predict burn wound healing time and to suggest optimal reconstructive strategies for complex cases[5], [6].

One challenging context where decision support could be highly beneficial is in reconstructive chest surgeries, such as post-mastectomy breast reconstruction, correction of gynecomastia, and gender-affirming chest masculinization procedures. A critical aspect of these surgeries is the proper localization of the nipple-areola complex (NAC) on the chest wall. Achieving a natural and symmetric NAC position is central to the aesthetic success of the reconstruction [7]. Surgeons typically determine the NAC placement during preoperative planning, drawing on experience and various anthropometric guidelines. However, for male chests and post-mastectomy reconstructions, there are relatively few standardized algorithms, and many are impractical to apply [8]. As a result, surgeons often rely on intuition, experience and individualized judgment for NAC placement in these cases [9].

Recent studies have quantified the extent of variation and difficulty in NAC localization. Even experienced plastic surgeons exhibit significant discrepancies when asked to place the NAC on a given chest wall, indicating that intuitive determination of the NAC is a challenge and prone to inconsistency [10]. In the literature, the ideal NAC position remains a topic of debate: some authors advocate using anatomical landmarks such as the pectoral muscle borders or thoracic circumference as guides, whereas others acknowledge resorting to subjective judgment [11], [12]. This variability can lead to suboptimal outcomes if the NAC is misplaced, as the human eye is

very sensitive to asymmetry [13].

Advancements in medical imaging and computational analysis offer a promising path to address this need. Modern machine learning algorithms, especially deep learning models, can learn complex spatial patterns from imaging data and have advanced automated image analysis in medicine [14], [15]. In fields like radiology and oncology, convolutional neural networks (CNNs) can detect subtle features in images that humans might overlook [16]. Translating these capabilities to surgical planning, researchers envision AI systems that analyze patient-specific anatomical data and propose individualized operative plans [17], [18]. In breast surgery, machine learning tools have been tested for tasks ranging from tumor localization to predicting reconstructive outcomes [19], [20]. Applying machine learning to NAC localization is a natural extension of this trend.
Initial efforts in this direction have shown encouraging results. Ghodratigohar et al. developed a prototype machine learning method to assist NAC localization in chest masculinization surgery, using an artificial neural network trained on a dataset of male torso images [21]. Despite a limited dataset, the model achieved high accuracy, demonstrating the feasibility of using machine learning to support surgeons in NAC placement decisions. However, clinical datasets in this domain are relatively small and often lack diversity due to the challenges of data collection in surgery [22]. Many machine learning models in medical imaging struggle with insufficient or biased training data, which can hinder their reliability when generalized to new patients [23].

1.1 Objectives of the thesis

Accurate and consistent localization of the nipple-areola complex in chest reconstruction surgery remains a significant challenge. During chest reconstructive surgery, surgeons performing post-mastectomy breast reconstructions, gynecomastia corrections, and gender-affirming chest surgeries often lack clear reference points for NAC positioning. Current NAC placement relies on subjective judgment and heuristic guidelines, resulting in high variability among surgeons [9], [10]. Patient-specific anatomical variation further complicates the determination of an ideal NAC location. The absence of reproducible, objective methods introduces the risk of suboptimal aesthetic outcomes and potential need for revision surgeries[7], [13]. The paucity of available data also makes the development of an automated process a challenge. Therefore, the goals of this thesis are:

- To develop a machine learning-based decision support approach to assist in the precise and

consistent localization of the NAC in reconstructive chest surgery.

- To evaluate the effect of augmenting real clinical datasets with synthetic images on the performance and generalizability of the decision support system.

1.2 Motivation

The motivation for solving this problem is both clinical and technical. Clinically, precise positioning of the nipple–areola complex (NAC) is critical to achieve natural chest aesthetics and high patient satisfaction in both gender-affirming and post-mastectomy reconstruction. In chest masculinization surgery, for example, accurate NAC placement is essential to create a convincing male chest contour [24]. Misplacement of the NAC is a well-known pitfall [25], and proper NAC placement significantly impacts postoperative satisfaction and body image, especially in gender-affirming and post-cancer procedures [7], [13]. Patient-reported outcomes confirm this importance, as the analysis of 363 of transgender men's reviews found that NAC position and appearance were the strongest drivers of satisfaction with chest masculinization surgery. Likewise, in female breast reconstruction following cancer, restoring the NAC markedly improves body image and psychosocial health. Women score significantly higher on psychosocial and sexual well-being after undergoing NAC reconstruction than before [26]. In short, the NAC is often described as the "finishing touch" that completes breast reconstruction and strongly influences a patient's sense of normalcy and confidence [26] . Technically, NAC localization represents a complex, image-based anatomical prediction task that challenges current machine learning pipelines due to small, non-diverse datasets [22]. The literature has proposed many different references (the edges of the pectoralis muscle, rib or intercostal spaces, and thoracic width ratios[7]) but these remain largely unstandardized. For instance, surgeons often rely on intuition or experience when deciding NAC placement, rather than a reproducible mathematical rule [7]. This variability makes automated image-based prediction difficult: a recent review notes that although many anatomical markers have been suggested, "placement of the NAC is contentious in the literature" and consensus is lacking [7]. Timmermans et al. likewise observe that an accurate patient-specific model for NAC position is still needed for preoperative planning [27]. Most current approaches are ad hoc: surgeons measure and mark landmarks by hand or use simple formulas. This manual landmarking is time-consuming and introduces variability. Timmermans et al. found that without a standardized algorithm, routine practice led to systematic placement errors

(on the order of 1–2 cm) in both the horizontal and vertical coordinates. The lack of reproducible, data-driven placement methods means existing techniques often lack consistency and generalizability. Developing an effective solution could reduce subjectivity in surgical planning, improve aesthetic outcomes, and serve as a template for broader AI-assisted tools in plastic surgery.

It is worth mentioning that this project was initiated at the request of ***Dr. Kevin Cheung***, a reconstructive surgeon, who identified the need for a more automated and reliable method for NAC localization in surgical planning

1.3 Contributions

This thesis makes the following contributions:

- Development of a fully automated machine learning-based NAC prediction model using pose estimation and normalized distances from anatomical landmarks.
- Implementation of a tilt correction mechanism to improve model accuracy in cases with subject tilt.
- Integration of an automatic prompt generation procedure using a large language model (ChatGPT-4)
- Integration of synthetic data generated using a diffusion model (Emu) to augment dataset diversity and size.
- Experimental evaluation across three scenarios: clinical images, synthetic images with demographic variability, and external real-world public images.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 presents a literature review. Chapter 3 presents the development, methodology, and evaluation of a machine learning-based NAC localization system using pose estimation and regression models. Chapter 4 explores the use of synthetically generated chest images to augment training data and improve model generalization. Chapter 5 outlines the conclusion and an outlook on future work.

Chapter 2. Background and Related Work

This chapter establishes the technical and clinical context of this thesis by surveying the evolution of decision support systems in surgery, with a focus on applications in plastic and reconstructive procedures. It provides an overview of existing methods and technologies that aim to improve operative planning, especially regarding the localization of the nipple-areola complex (NAC) in chest reconstruction. This background introduces the current state of the field and the necessity of developing AI-based solutions for NAC positioning.

2.1 Decision Support Systems in Surgery

In this section, we review decision support systems and their integration into surgical practice. We examine how these systems, ranging from rule-based to AI-powered tools, have contributed to enhancing surgical planning and intraoperative guidance.

2.1.1 Decision Support in Surgical Planning

Clinical decision support systems (CDSS) are computer-based tools designed to assist clinicians in making complex decisions by integrating patient-specific information with a knowledge base[28]. In the context of surgical planning, decision support aims to enhance the surgeon's judgment and planning process with data-driven insights. Since their introduction in the 1980s, CDSS have evolved significantly and are now often embedded in electronic health records to provide real-time recommendations [28]. For surgeons, this means pertinent alerts or suggestions (for example, reminders for prophylactic antibiotics or risk assessments) can be provided during the preoperative planning and perioperative period. Garg et al. demonstrated that such systems can improve adherence to clinical guidelines and provider performance, potentially reducing errors and adverse events [29]. However, successful implementation in surgical workflows requires careful integration so as not to overwhelm or distract the surgical team. User-friendly design and relevance of advice are factors in surgeons' acceptance of these tools.

Decision support in surgery can be broadly categorized by the approach used. Early-generation systems are knowledge-based, relying on explicit clinical rules and guidelines (often encoded as "if-then" statements), whereas newer systems may be non-knowledge-based, employing statistical patterns or artificial intelligence (AI) to generate recommendations [28]. Knowledge-based (rule-based) systems use a predefined knowledge base, for example, a database

of surgical best practices or protocols against which patient data are compared to produce recommendations [28]. These systems have the advantage of transparency and explainability, as the rationale follows human-authored rules. On the other hand, non-knowledge-based systems leverage machine learning to identify complex patterns in data without explicit human-written rules [28]. Such AI-powered decision support can potentially consider a vast array of variables (imaging, lab results, prior cases, etc.) and may uncover insights not evident to human observers. The drawback of using machine learning is that AI models often act as "black boxes" with limited explainability, which raises concerns in high-stakes fields like surgery where accountability and transparency are critical [28], [30]. Despite these challenges, there is growing enthusiasm for integrating AI into surgical decision support because of its promise to improve predictive accuracy and individualized planning [30]. Loftus et al. highlight that artificial intelligence-enabled decision support in surgery is an emerging state-of-the-art, expected to transform how surgeons plan and execute procedures in the future [30].

Overall, decision support systems in surgical planning seek to augment the surgeon's expertise with computational assistance. By combining patient-specific data with medical knowledge, these systems can provide risk assessments, diagnostic suggestions, operative planning guidance, and reminders for best practices. The goal of these systems is not to replace surgeon decision-making but to support it, improving safety, consistency, and outcomes. As more data become available and AI algorithms become more sophisticated, decision support tools are likely to play an increasingly prominent role in surgical planning and intraoperative guidance [30], [31]. Ensuring that these tools are validated and carefully integrated into clinical procedures will be key to realizing their benefits while maintaining the surgeon's authority in decision-making.

2.1.2 Decision Support Systems in Plastic Surgery

Within the field of plastic and reconstructive surgery, decision support systems have been applied to assist surgeons in making complex operative planning choices. Plastic surgery often involves complex decisions, such as selecting the optimal reconstructive technique for a defect or weighing the trade-offs between aesthetic and functional outcomes [32]. Traditional surgical decision-making in this domain has relied heavily on surgeon experience and patient preference. In recent years, formal decision analysis and computerized support tools have been introduced to bring more objectivity and consistency to these choices. For example, Sears et al. present decision analysis as

a structured approach to guide plastic surgeons through complex clinical decisions [33]. This technique involves constructing decision trees that map out possible strategies and outcomes, assigning probabilities and utilities to each, and computing which strategy yields the best expected outcome [33]. The introduction of decision analysis to plastic surgery provides a systematic way to evaluate options like reconstruction versus amputation or various surgical techniques, especially in cases where high-level evidence may be limited. By 2010, decision analysis was being promoted as an important tool for plastic surgeons to complement clinical judgment with evidence-based modeling of decisions [33].

Plastic surgery has also benefited from specialty-specific risk assessment tools that support decision-making. An example is venous thromboembolism (VTE) risk stratification for postoperative patients. The Caprini Risk Assessment Model, originally a generic tool, has been validated in plastic surgery populations as an effective predictor of VTE risk [34]. Pannucci et al. (2011) demonstrated that the Caprini score accurately stratifies plastic surgery patients by VTE risk, which in turn assists surgeons in making prophylaxis decisions (such as prescribing anticoagulants for high-risk patients) [34]. By identifying patients at elevated risk for thrombosis, the surgical team can tailor perioperative management accordingly. This kind of decision support (a point-based statistical risk model) improves patient safety by ensuring at-risk patients receive appropriate preventative measures. In practice, the Caprini score and similar risk models have become part of many plastic surgeons' preoperative planning to guide discussions about prophylaxis and to fulfill informed consent obligations regarding risk.

Another area of decision support in plastic surgery is the development of algorithm-based clinical pathways for specific procedures. These are essentially rule-based decision support systems tailored to common surgical scenarios in plastic surgery. A recent example is the work of Ziani et al. (2025), who developed a digital decision-support application to standardize intraoperative choices during immediate breast reconstruction in patients with large ptotic breasts [35]. This mobile app encodes an institutional surgical algorithm as an interactive guide: at each decision point during the operation, the app prompts the surgeon with questions about the patient's anatomy or surgical findings and based on the preset algorithm, recommends the next step (such as which reconstructive technique to use or whether to employ adjunctive support like acellular dermal matrix) [35]. In a pilot study, use of the app led to consistent decision-making across different surgeons, indicating that a rule-based system can reduce variability in care [35]. This was

a demonstration of a decision support system in plastic surgery that improves adherence to best practices. The algorithm was derived from expert consensus and prior outcomes, and by following it, surgeons achieved standardized results. Such tools are particularly valuable in complex reconstructive procedures where multiple viable options exist because they ensure that decisions are aligned with evidence-based pathways rather than purely individual preferences.

Plastic surgeons are increasingly exploring advanced decision support powered by artificial intelligence as well. Although detailed discussion of AI methods is presented in section 2.1.5, it is worth noting here that plastic surgery has been at the forefront of some AI-driven decision support innovations. For instance, in breast reconstruction planning, machine learning algorithms have been applied to recommend whether an implant-based or autologous tissue reconstruction would be most appropriate for a given patient, based on that patient's unique profile (e.g., body habitus, medical history, and aesthetic goals) [36]. A recent systematic review by Ospina-Gómez et al. highlights several advances in applying AI to breast reconstruction, including predictive models that assist in flap selection and outcome prediction [36]. These AI models serve as decision aids by analyzing large datasets of past patients to predict which surgical approach might yield the best balance of safety and aesthetic outcome for a new patient. [32], [33]. Decision support algorithms have demonstrated promising performance in tasks such as complication risk estimation and flap viability prediction, with reported consistent AUC values between 0.70 and 0.78 for complication prediction models [36] and AUC values of 0.95 (training) and 0.67 (testing) for flap failure prediction [37].

2.1.3 Rule-Based Systems

Rule-based decision support systems rely on explicit knowledge representations, typically in the form of if-then rules or decision trees crafted by medical experts. These systems are a subset of knowledge-based CDSS and were among the earliest successful implementations of computerized decision support [28]. In a rule-based system, a knowledge base of facts and rules is created, often derived from clinical practice guidelines, protocols, or expert consensus. During use, the system matches patient-specific inputs to this knowledge base and triggers recommendations according to the encoded rules. For example, a simple rule in a surgical context might be: "IF the patient has an allergy to penicillin, THEN recommend an alternative antibiotic for prophylaxis." More complex rule sets can guide multi-step decision processes, such as workups for surgical candidates or

perioperative management pathways.

In surgical planning and execution, rule-based CDSS have been used to promote standardization and adherence to best practices [29]. One benefit of rule-based systems is that the rationale for their recommendations is transparent, because the rules can be inspected and understood by clinicians. This transparency helps build trust, as surgeons can verify that the suggestions align with established guidelines. Many hospitals employ rule-based alerts in their order-entry systems (for instance, reminding a surgeon to order deep vein thrombosis (DVT) prophylaxis if a patient's risk factors meet a certain threshold). These interventions have been shown to improve compliance with care protocols and reduce oversight errors, thereby potentially improving patient outcomes [29]. Systematic reviews have found that knowledge-based CDSS can significantly improve practitioner performance in following guidelines, though the impact on patient outcomes can vary case by case [29].

Despite their advantages, rule-based systems have limitations. They are only as comprehensive as the knowledge encoded into them. If a rule base is incomplete or outdated, the system may not recognize special circumstances or novel strategies, limiting its usefulness. Additionally, rigid rule-based advice may not account for context that falls outside the predefined logic. Surgeons sometimes face scenarios where strict guideline adherence must be balanced against individual patient factors, and a simplistic rule might not capture that context. In such cases, overly rigid decision support could be ignored or even potentially misleading. Therefore, maintaining and updating the knowledge base is critical, which would require continual input from experts to keep the rules aligned with current best practices.

Nonetheless, well-designed rule-based CDSS remain highly relevant in surgery. As mentioned in the section 2.1.2, Ziani et al. implemented a rule-based intraoperative guidance system (an algorithmic app) for breast reconstruction [35]. This is essentially a modern expert system where the "if-then" logic was derived from an institutional algorithm vetted by experienced surgeons. The success of that system, which improved decision consistency, illustrates that when expert knowledge is appropriately captured in rules, the CDSS can effectively act as a real-time checklist and guide. Beyond single-institution projects, many surgical specialties have developed consensus-based care pathways (for example, Enhanced Recovery After Surgery (ERAS) protocols in colorectal surgery or algorithms for trauma care) and embedded them into electronic CDS tools. These are rule-based by nature and help ensure that no key step is missed in patient

care.

2.1.4 Statistical Models

As computing power and clinical data availability grew, statistical models became important tools for decision support in surgery. These models typically use patient data and statistical methods (such as regression analysis) to predict outcomes or risks, thereby aiding decisions with quantitative risk-benefit estimates. Unlike simple rule-based logic, statistical models can consider multiple variables simultaneously and provide probabilistic assessments [38]. In surgical planning, one of the most common uses of statistical decision support is in risk prediction calculators. These calculators estimate the likelihood of various outcomes (complications, mortality, etc.) for an individual patient if they undergo a certain procedure. Surgeons can use these predictions to inform treatment planning and to have more informed consent discussions with patients.

An example is the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) Surgical Risk Calculator, a tool developed from a multi-institutional clinical dataset of over a million patients [38]. The NSQIP risk calculator uses logistic regression models to predict the chances of specific complications (like pneumonia, cardiac event, surgical site infection) and overall mortality for a given patient based on their preoperative variables and the planned procedure [38]. Bilimoria et al. (2013), who developed and validated this universal risk calculator, described it as a decision-support tool to facilitate shared decision-making and informed consent [38]. The calculator outputs patient-specific risk percentages for outcomes within 30 days of surgery. In practice, a surgeon planning an operation can input the patient's demographics, comorbidities, and the intended procedure, and the model will provide an estimated risk profile. If the risk for a complication is particularly high, the surgeon might opt to modify the surgical plan (for instance, schedule a higher level of postoperative care, or even reconsider the indication for surgery if nonoperative options exist). On the other hand, a low-risk profile can reassure both surgeon and patient and support proceeding with the planned intervention. The NSQIP calculator exemplifies how statistical models improve decision-making: it brings data-driven risk assessment to the table, countering cognitive biases or anecdotal impressions with hard numbers.

Many other statistical models are in use across surgical specialties. In cardiac surgery, for example, the EuroSCORE and Society of Thoracic Surgeons (STS) risk models predict operative mortality, helping decide candidacy for high-risk surgeries [39]. In plastic surgery specifically, the

field has developed risk models for certain outcomes like flap failure or wound complications. These models often take the form of nomograms or scoring systems derived from retrospective clinical studies. As noted earlier, the Caprini score is a statistically derived risk stratification for venous thromboembolism (VTE); its validation in plastic surgery demonstrates its effectiveness at guiding prophylactic decisions [34]. Another example in plastic surgery is the work by O'Neill et al., who created a machine-learning-based model (a form of statistical predictive model) to predict free flap failure in breast reconstruction using numerous perioperative variables [40]. That model achieved a high accuracy of 0.82 Area Under the Receiver Operating Characteristic Curve (AUROC) in training and 0.63 AUROC on external testing, highlighting the need for robust validation. This demonstrates how more complex statistical, or machine learning models are being explored to support microsurgical decision-making. Surgeons could use such a model preoperatively to identify patients at high risk of flap loss and perhaps choose a different reconstructive strategy or ensure closer monitoring.

Statistical decision support models generally enhance the surgeon's ability to anticipate outcomes. They are particularly useful in facilitating shared decision-making with patients, because they provide concrete estimates (e.g., "Based on our data, you have a 5% chance of a major complication with this surgery"). This information can be important in helping patients weigh the pros and cons of elective surgeries. It also allows personalization, moving away from one-size-fits-all assumptions to acknowledging, for example, that an elderly patient with multiple medical conditions might face a very different risk than a young, healthy patient. The limitation of statistical models is that they rely on the quality and representativeness of the data on which they were developed. If a patient differs significantly from the development population or if a surgery is uncommon, the estimates might be less reliable. Moreover, traditional statistical models assume relationships based on historical data and might not capture complex, non-linear interactions as well as newer machine learning methods. Despite these weaknesses, statistical models remain a used tool in surgical decision support, providing an evidence-based quantification of risk that complements the surgeon's clinical judgment [34], [38]. They serve as an intermediate step between simple rules and AI usage, as they are more flexible than rigid rules, yet more interpretable and established than many machine learning models.

2.1.5 AI-powered systems

The latest generation of decision support systems in surgery are driven by AI and machine learning algorithms. These AI-powered systems represent a shift from explicitly programmed knowledge to learned patterns extracted from large datasets. In surgical planning and decision-making, AI technologies are being employed to tackle complex tasks such as medical image analysis, outcome prediction, surgical skill assessment, and even real-time intraoperative guidance [41]. One of the major advantages of AI systems is their ability to integrate multiple streams of data like radiologic images, electronic health record data, surgical videos, etc. and detect patterns or make predictions that would be challenging for humans to do unaided.

In the preoperative phase, AI has shown promise in processing imaging studies for surgical planning. For instance, computer vision algorithms using deep learning can automatically segment anatomical structures on Computed Tomography (CT)/Magnetic Resonance Imaging(MRI) scans, which is valuable for planning reconstructions or oncologic resections [37]. In plastic and reconstructive surgery, this capability has been used to map out perforator vessels for flap surgeries: AI models can analyze a CT angiogram and identify the optimal perforators for a microsurgical flap, saving the surgeon time and increasing accuracy in planning the incision and flap design [37]. By improving preoperative vessel mapping, such AI decision support can reduce intraoperative guesswork and potentially lower the risk of flap failure or operative time. Similarly, AI-driven 3D modeling can help craniofacial surgeons by converting patient imaging into virtual surgical plans like planning osteotomies in orthognathic surgery or simulating post-reconstruction facial appearance [37]. These AI tools augment the surgeon's planning by providing a virtual environment to rehearse or fine-tune surgical steps, effectively serving as a sophisticated decision aid when choosing how to execute a surgery.

Machine learning models have also been created to predict surgical outcomes and complications with high complexity, sometimes outperforming traditional statistical models [42]. These models (ranging from random forests to neural networks) can incorporate hundreds of variables to forecast outcomes like complications, length of stay, or success of a graft. In the realm of plastic surgery, recent studies have used machine learning to predict outcomes such as breast reconstruction success or the likelihood of postoperative complications in cosmetic surgery patients [36], [37]. Because machine learning can capture non-linear relationships, it may detect subtle combinations of risk factors that classical regression would miss. Comprehensive AI-

powered decision platforms that combine multiple aspects are emerging. For example, an AI system might simultaneously analyze patient risk factors (predicting complications), anatomical considerations (from imaging), and even surgeon-specific performance data to recommend an optimal surgical plan [37]. Mansoor et al. (2025) describe such integrated platforms where AI provides a recommendation on surgical approach by synthesizing patient-specific risk assessment, expected outcomes, and anatomical analyses [37]. In their vision, a plastic surgeon planning a breast reconstruction could receive an AI-generated recommendation on whether to use an implant or a particular flap, which donor site to use, and how to execute the reconstruction to maximize symmetry and minimize complication risk, all tailored to the patient's unique profile [37]. This kind of comprehensive decision support goes far beyond what earlier rule-based systems offered.

AI is also making its way into intraoperative decision support. In the operating room, advanced systems can monitor the surgical field via computer vision. For example, in laparoscopic surgery, AI algorithms can identify anatomical landmarks on the video feed and alert the surgeon if instruments are approaching critical structures (as has been prototyped in bile duct identification during cholecystectomy) [30]. These systems essentially act as a co-pilot, offering real-time "observations" and safety checks. Another frontier is the use of large language models for immediate decision support. Recent studies have evaluated models like ChatGPT in answering intraoperative queries from surgeons. Atkinson et al. tested ChatGPT's ability to provide accurate responses to urgent surgical questions (specifically in the context of a complex plastic surgery flap procedure) and found that while the AI could rapidly supply information, its reliability varied, and it occasionally gave incorrect or incomplete advice [43]. This highlights both the potential and the current limitations of AI, as a language model can sift through medical knowledge in seconds to assist a surgeon, but it might lack the judgment or updated knowledge base to be fully trusted without oversight [43]. Ongoing research is comparing different AI models (for example, OpenAI's GPT-4 vs. other models) in how well they support surgical decision-making, with improvements occurring rapidly in this field [44].

While AI-powered decision support offers powerful new capabilities, it also brings challenges that need to be addressed for safe adoption in surgery. Surgeons must be aware of issues like algorithm bias (AI models trained on non-diverse data may not generalize to all patient populations) and lack of explainability in many AI recommendations [28], [30]. Moreover, surgeons and patients may be hesitant to accept an AI-driven suggestion without a clear rationale

or evidence of proven benefit. For these reasons, Loftus et al. stresses that AI systems in surgery should function as adjuncts, providing supportive information rather than autonomous decision-makers [30]. The surgeon remains the final arbiter of any decision, using AI input as one additional source of insight. With appropriate validation, regulatory oversight, and user training, AI-powered systems have the potential to enhance surgical planning and execution. Early outcomes from AI implementations are promising, but further studies and experience will determine how these tools can be best integrated into surgical practice for improved patient outcomes [30][37]. What is clear is that the use of both surgical expertise and artificial intelligence holds promise for advancing surgical care.

2.2 NAC Localization and Placement Techniques

This section examines the various approaches developed to determine ideal NAC positioning during chest surgeries. It describes the historical and traditional methods, then progresses to more structured geometric and manual measurement techniques and finally evaluates recent innovations involving artificial intelligence and machine learning. This progression reflects how the field has evolved from heuristic rules to increasingly personalized, quantitative planning strategies.

2.2.1 Traditional Methods

In the early development of breast surgery, surgeons focused on techniques to preserve the NAC rather than precise quantitative positioning. In 1922, Thorek introduced the use of free NAC grafts (removing and reattaching the nipple) to reposition the complex [45]. This approach allowed flexibility in placement but often sacrificed innervation and carried risks of graft loss. It was soon largely supplanted by pedicled flap techniques, which are techniques that allow tissue transfer while maintaining the NAC blood supply. For example, Schwarzmann in 1930 described a periareolar de-epithelialization method, and Gillies and McIndoe in 1939 developed a skin–gland flap mammaplasty to transpose the NAC while keeping it vascularized [45]. These early methods established that the NAC could be moved to a more natural location on the reshaped breast mound, but the placement was determined mostly by the surgeon's experience and anatomical landmarks rather than by measurements.

By the mid-20th century, efforts began to standardize the ideal NAC position on the breast. Penn in 1955 proposed a simple rule based on anthropometry: in an aesthetically ideal breast, the

suprasternal notch and the two nipples should form the vertices of an equilateral triangle with sides approximately 21 cm in length [46], [47]. This guideline implied that the nipple-to-nipple distance and the notch-to-nipple distance should each be about 21 cm in young women with proportionate figures [47]. However, this guideline did not include any statistical validation across diverse body types, and was not adjusted for patient height, weight, chest width or Body Mass Index (BMI). In 1956, Wise introduced a preoperative "keyhole" skin resection pattern designed to achieve this symmetric placement of the NAC on the breast mound [48]. The Wise pattern provided a reproducible template (resulting in the classic inverted-T scar) that helped surgeons position the NAC consistently while preserving blood supply and sensation [48]. Over time, surgeons refined reduction mammaplasty markings like Pitanguy's 1960s technique using a dermoglandular pedicle, however, they generally still adhered to similar landmark-based NAC positioning (e.g. a sternal notch-to-nipple distance around 20–22 cm) [49]. Traditional female breast reduction markings often placed the new nipple approximately 7 cm above the inframammary fold on the breast meridian and centered over the breast mound [49]. This typically corresponded to the level of the inframammary fold projected onto the breast surface, which surgeons would locate by palpation or with calipers [47]. Such methods, while simple, introduced subjectivity. For example, feeling the fold through a large breast is imprecise and did not account for variations in breast ptosis[47].

In male chest surgery (gynecomastia and female-to-male transgender mastectomy), early practices likewise relied on fixed anatomical landmarks. Surgeons historically placed the NAC at roughly the level of the 4th intercostal space (around the 4th rib) along the lateral border of the pectoralis muscle, which in an adult male corresponds to the usual nipple line[50]. For instance, Murphy et al. in 1994 recommended positioning free nipple grafts for severe gynecomastia at the fourth rib and just medial to the anterior axillary line[51]. The horizontal position was often taken as the midclavicular line on each side, yielding an internipple distance similar to an average male's chest width [50]. Many surgeons simply "eyeballed" the placement to mimic a cisgender male chest, sometimes using the one-third rule of the clavicle or aligning with the pectoral muscle contour [50]. For example, Monstrey et al. suggested locating the new nipples one-third of the clavicle length from the sternum on each side and slightly above the inframammary line [50]. A common rule of thumb that emerged was to position male nipples about 19–20 cm down from the sternal notch and approximately 22 cm apart from each other, based on average male

anthropometry[50]. These set values were used in many standard “top surgery” cases and are in line with measured norms (cisgender male averages around 19.3 cm notch-to-nipple and 22.3 cm internipple distance)[52]. However, such one-size-fits-all guidelines often failed to account for varied body types and weights. Clinicians observed that strictly following Penn’s 21 cm triangle or analogous fixed numbers could produce a less natural look in older or larger patients, whose nipples often sit lower and farther lateral preoperatively [47]. In the context of female to male (FTM) chest reconstructions, early outcomes sometimes had the nipples placed too high on the chest or too close together because surgeons applied female-based distances without tailoring to the patient’s torso [53]. As Maas et al. noted in a 2020 review, these traditional methods using chest-wall landmarks or fixed measurements often lacked a patient-specific approach and yielded inconsistent aesthetic results in many cases [24]. Over time, it became clear that a more objective, individualized strategy was needed to improve the accuracy of NAC localization.

2.2.2 Geometric and Manual Measurement Techniques

To introduce more precision into NAC placement, surgeons and researchers developed geometric and measurement-based techniques in the late 20th and early 21st centuries. Rather than relying on rough estimates, these methods use anthropometric measurements and simple math or tools to determine the optimal nipple position for each patient. One early example was by Aufricht in 1949, who advocated an “empiric and geometric planning” for pendulous breast reductions [47]. In the modern era, several groups performed detailed anthropometric studies to derive reproducible landmarks. Beckenstein et al. (1996) measured 52 males and reported normative distances for male NAC location, finding for instance an average internipple distance of about 22 cm and a nipple-to-midline distance around 11–12 cm in their cohort [54]. They also documented average areola sizes and noted little variation with age or body mass, establishing baseline parameters for male chest aesthetics. Around the same time, Shulman et al. (2001) analyzed male chest measurements and suggested that the midclavicular line and inframammary fold could serve as reliable guides for horizontal and vertical nipple position, respectively [12], [50]. These manual measurement approaches provided surgeons with concrete numbers (e.g. distance from the sternum or clavicle) to use during preoperative marking.

A significant advance came from applying analytical geometry to NAC localization. Beer et al. (2001) developed a pair of equations to calculate the ideal nipple coordinates based on a

patient's thoracic dimensions[11], [24]. In their method, the chest circumference and the sternal length are measured, and then simple formulas yield the recommended nipple-to-midline distance and nipple height on the chest wall. This was one of the first attempts to create a predictive model for NAC placement: Beer's formula could be used to personalize the internipple distance and the vertical level for an individual male patient [24]. Although relatively straightforward, the accuracy of this approach depended on the assumption that chest shape scales predictably with those two measurements. Subsequent studies by other authors worked to improve such models. Hage et al. (1995), in a comprehensive review of FTM chest contouring, also emphasized taking patient measurements into account and compiled earlier data to guide nipple positioning in transgender men[55]. They highlighted that previous algorithms with fixed values did not fit everyone and called for more tailored techniques [56].

Other teams devised geometric constructions to locate the NAC. In female breast surgery, Khan et al. (2008) introduced an innovative method using the geometry of a triangle. They postulated that the two nipples and the sternal notch should form an equilateral or short isosceles triangle on the chest [47]. In their technique, the surgeon measures the base width (across the chest) and the vertical distance from the notch, then uses the Pythagorean theorem to calculate the hypotenuse, which corresponds to the precise distance from the notch to each new nipple position. This "geometrical calculation method" objectively determines where to mark the nipple on the breast surface. Khan et al. reported applying it in 40 breast reduction patients with satisfactory placement of the NAC in all cases. They noted that the method was quick, required no special instruments, and gave trainees confidence by reducing work that required subjective judgement. Similarly, in male chest surgery, Mett et al. (2017) proposed the Mohrenheim-Estimated Tangential Tracking Line (METT-Line) technique as an intuitive geometric guide. By evaluating male chest photographs and surgeons' freehand nipple drawings, Mett's team found that the NAC in men consistently lies along a line drawn from the infraclavicular fossa (Mohrenheim's pit) and about 4 to 4.5 cm above the inframammary fold level [8]. This finding was used to define the METT-Line, along which the nipples can be placed using a combination of metric measurement and the surgeon's visual judgment. The METT-Line approach bridges empirical intuition with a fixed geometric reference, and the authors reported it could reliably recreate a natural male NAC position in their series of patients.

Several manual measurement algorithms have also been developed. Atiyeh et al. advocated

another approach by invoking the golden ratio ($\varphi \approx 1.618$) in determining NAC coordinates. They observed that previous guidelines often yielded NAC placements that were "too high and too far medially" on the male chest and sometimes involved cumbersome calculations [12]. In response, Atiyeh's method uses the divine proportion to simplify the process: key horizontal and vertical distances on the torso are partitioned according to the golden ratio to locate the nipples in a more aesthetically pleasing position. For example, one of the relationships they mention is between the internipple distance and the distance from the umbilicus to the axilla, which they found to follow a golden ratio in idealized physiques. By applying φ, their technique yields a quick two-step measurement for re-positioning nipples without complex tools. Atiyeh's study, which presented this method along with clinical examples, argued that it produces a balanced look and is easier for surgeons to use in the operating room. Another example of a simplified measurement tool is the use of templates or positioning calipers that some surgeons employ. For instance, a template might ensure the nipples are marked at a set distance from midline and at the same height bilaterally before making incisions [24]. These manual aids, while not algorithmic, represent the practical extension of geometric principles in surgery.

In recent years, hybrid approaches combining patient-specific data and simple formulas have emerged. Tanini et al. (2021) examined a cohort of athletic male subjects and noted that the ideal male NAC has a consistent spatial relationship with the pectoralis major muscle. They found that in well-built physiques the NAC is typically about 3 cm medial to the lateral border of the pectoralis and roughly 2.5 cm above the inferior insertion of that muscle on each side [50]. Using these observations, Tanini's group described a quick intraoperative technique: after identifying the edges of the pectoral muscle on the chest wall, the surgeon can mark the new nipple positions by measuring a few centimeters inward from the muscle edge and up from its lower margin [51]. This method was demonstrated in a pilot study of young male patients who were water-polo players with fit bodies. However, the authors acknowledged that their cohort was relatively homogenous, and results might differ in patients with higher body fat or less defined muscular tone.

All these geometric and manual measurement techniques aim to improve the accuracy and reproducibility of NAC localization. They introduce objective criteria, whether mathematical equations, proportional constants, or fixed landmark offsets, to guide surgeons in planning nipple position. For instance, when applying Beer's formula or Khan's triangle in practice, surgeons can achieve nipple symmetry within a few millimeters of the intended spot, as opposed to variations

of a centimeter or more with purely visual estimation [47]. The push for measurable precision also has introduced the idea of validating these methods statistically. Timmermans et al. performed meta-analyses and external validations to compare different algorithms' performance [27]. In one study, Timmermans compiled data from 150 cisgender men to create a two-step regression model predicting optimal internipple distance and sternal notch-to-nipple distance for a given individual. This model achieved a mean absolute error under 5% for both key distances and proved reproducible across external cohorts [27]. The effort to statistically validate such algorithms shifts this field from traditional "rules of thumb" to evidence-based, quantitative planning. Despite these improvements, these approaches remain constrained by assumptions of anatomical regularity and limited adaptability to diverse body types. Moreover, the reliance on manually derived measurements and fixed geometric heuristics restricts personalization, especially in complex or non-normative chest anatomies. These limitations emphasize the need for more dynamic, data-driven solutions, such as AI-based systems, that can generalize across variability and optimize NAC localization beyond the constraints of rule-based methods.

2.2.3 AI-Based Techniques

Precisely positioning the NAC is critical for achieving a natural male chest contour. Surgeons performing chest masculinization or gynecomastia correction surgeries have historically relied on subjective judgment. As Mett et al. observed, "to date, there are no standards" for male NAC placement and "only a few, metric and often impractical algorithms" are available [8]. Recent studies have therefore explored computer vision and machine-learning solutions to standardize NAC localization.

In one example, Kenig et al. developed a YOLOv3-based neural network to detect key breast landmarks (including the NAC) in frontal photographs of breast-reconstruction patients [57]. While their model detected all NACs and achieved a landmark detection rate of 97.74%, the method was limited to postoperative analysis and did not provide predictive capabilities for preoperative NAC localization or guidance for initial placement. Similarly, Park et al. created the S-BEST system, a DenseNet-264 network trained on annotated frontal breast photos, to automatically detect 30 standardized breast landmarks including both nipples [58]. In validation, S-BEST's landmark measurements showed no statistically significant differences from manual physical exam measurements for most distances (except for a high bias in nipple to inframammary

fold distance), indicating high localization accuracy [58]. However, the system is limited to postoperative aesthetic evaluation and does not provide predictive guidance for preoperative NAC positioning, particularly in cases requiring reconstruction or gender-affirming surgery. Other work has applied deep learning to related tasks. For instance, Guo et al. (2022) proposed a fully automated deep-learning pipeline for breast cosmesis evaluation, using convolutional networks to segment breast contours and extract landmarks [59]. Their system achieved performance comparable to state-of-the-art clinical tools (such as BCCT.core[60]) without manual inputs. However, the primary focus of this method is on postoperative cosmetic assessment rather than on predictive NAC placement. As such, while these machine learning-based approaches demonstrate that neural networks can detect NAC locations on patient images with high sensitivity (≈98% detection rates [57]), they do not address anatomical variability or offer decision-support capabilities for guiding initial NAC localization during surgical planning.

Ghodratigohar et al. developed a machine learning method to predict the nipple–areola complex (NAC) location for chest masculinization procedures using an artificial neural network trained on body ratios extracted from 173 frontal images of male torsos [61]. Their model was designed to estimate the horizontal and vertical positions of the NAC using normalized distances from key anatomical landmarks, including the suprasternal notch, umbilicus, and anterior axillary folds. These landmarks were manually annotated by the authors for each image in the dataset. By converting these distances into dimensionless ratios, the authors aimed to minimize the effect of scale and intersubject variability. The neural network, composed of fully connected layers with ReLU activations and dropout regularization, was trained to regress the vertical and horizontal NAC coordinates independently. On the test set, the model achieved root mean square errors (RMSE) of 0.0617 and 0.0560 for the vertical and horizontal ratios respectively. When translated into pixel-based distances normalized by body width, the RMSE for horizontal NAC position was 0.0254 (left) and 0.0314 (right), while the vertical position error was 0.0383. Despite these results, several limitations constrain the generalizability and clinical readiness of this approach. The dataset consisted of publicly available images of adult male torsos, which may not reflect the diversity in body types or anatomical variability encountered with real life examples. Moreover, the dataset size was limited, further constraining the generalizability of their model. In addition, the anatomical landmarks used as model inputs were manually identified, which introduces subjectivity and makes the process more labor-intensive, limiting its applicability in real-world

clinical settings.

2.3 Datasets for AI Decision Support Systems in Plastic Surgery

This section addresses the importance of data in the development of AI-driven decision support tools. It outlines the challenges of assembling and curating surgical datasets and discusses how dataset quality and diversity affect model performance. This section sets the stage for the data-centric methodology introduced in Chapter 4 of the thesis.

2.3.1 Overview of Existing Datasets

A variety of image datasets have been used to train AI in plastic surgery, though most are limited to specific tasks or small cohorts. In facial plastic surgery, for example, researchers have compiled pre- and post-surgery photograph collections. Rathgeb et al. describe the IIIT-Delhi Plastic Surgery Face Database, which originally contained 1,800 frontal face images (900 individuals) with "before" and "after" views for various procedures (rhinoplasty, blepharoplasty, brow lift, skin peeling, rhytidectomy)[62]. This database is publicly accessible via provided links, but many original images are now unavailable. An additional set Rathgeb et al. used, the Hochschule Darmstadt University of Applied Sciences Plastic Surgery Face Database, contains 638 subjects (540 female, 98 male), with roughly 100 image pairs each for five common facial surgeries. Such face datasets have been used in biometric and surgical outcome studies. These databases illustrate the types of clinical photographs used for AI tasks: typically, 2D frontal views standardized for size and pose.

For breast and aesthetic surgery, other clinical photograph collections have been used. Park et al. developed a deep-learning tool (S-BEST) for breast aesthetic scoring using a dataset of frontal breast photographs from 100 women with breast cancer[58]. They annotated each image with 30 anatomical landmarks (sternum, nipple, etc.) and trained their model to measure symmetry and distances. Their model achieved very high agreement with physical exam measurements on most landmarks, though one distance (nipple-to-inframammary fold) showed some bias (coefficient of determination ~0.38–0.42). Park et al. therefore demonstrate that relatively small, well-annotated real-image datasets (100 patients, 2D photos) can train useful AI tools for specific plastic surgery metrics. Similarly, small surgical image sets have been used for tasks like ptosis evaluation or symmetry assessment in reconstructive breast surgery (though these are usually proprietary to each

study, often on the order of dozens of patients). Keefer et al. conducted a systematic analysis of data sharing practices across the seven highest-ranked plastic surgery journals between 2018–2023, revealing that only 1.51% (11 out of 727) of published clinical articles included a data-sharing statement [63]. The prevalence of statements varied significantly by journal, ranging from 0.43% to 4.42%, and was more common in clinical trials (1.92%) than cohort studies (1.16%). Follow-up with authors who promised data availability yielded no responses. These results indicate that very few plastic surgery datasets are accessible or even acknowledged.

Singh et al. reviewed ethical factors related to artificial intelligence in plastic surgery and highlighted multiple dataset-related concerns, including bias and patient privacy risks [64]. The authors analyzed 63 AI-related publications (15 selected for in-depth review), concluding that most datasets lack diversity, particularly in terms of skin tone, gender, and age distribution, and that sensitive clinical images are often collected without standardized consent or anonymization. They emphasized the need for transparent, secure, and ethically governed dataset practices guided by regulatory frameworks such as HIPAA (Health Insurance Portability and Accountability Act) and for the deployment of explainable AI models. These observations demonstrate the significant ethical constraints and gaps in currently available datasets for plastic surgery AI models.

In addition to human-derived images, medical imaging datasets tangentially related to plastic surgery have been released. For instance, the fastMRI-Breast dataset provides multi-institutional contrast-enhanced breast MRI scans from 381 patients (over 2,000 MRI examinations) for AI research in breast imaging[65]. While fastMRI-Breast is aimed at cancer screening, such volumetric imaging resources could potentially support reconstruction planning or implant assessment. However, most explicitly plastic surgery AI work relies on 2D clinical photos as above.

For surgical planning (e.g. craniofacial or orthognathic surgery), some groups have collected limited 3D scan or radiographic sets (for example, the test set of 53 3D breast scans used in a computational modeling study [66]), but no large public repository is known. Spoer et al. reviewed 1,820 AI-related publications and found only 44 met criteria for clinical relevance, with nearly all still in early validation phases (phases 0–2) and only one study reaching phase 3 (clinical evaluation)[67]. The authors emphasized that inadequate dataset scale, structure, and standardization were major barriers to progress and recommended collaborative development of standardized plastic surgery datasets to support clinical deployment of AI

2.3.2 Challenges in Collecting Real Datasets

Building robust AI models for plastic surgery is hampered by numerous data challenges. Plastic surgical data are intrinsically sensitive and hard to share. Keefer et al. note that patient privacy concerns and unclear data-sharing conditions were primary barriers to data sharing [63]. Routine clinical photography in plastic surgery is invaluable for monitoring outcomes, but each image is potentially identifiable. As Keefer et al. explain, incorporating photography into routine care heightens legal and privacy risk, as digital images may expose patient identity if not rigorously de-identified. Informed consent for photography is another hurdle, as inconsistency in obtaining valid photo consent can lead to incomplete datasets [68]. Thus, IRB and ethical constraints often limit both the size and availability of real surgical image sets.

Even when data can be collected, dataset size and diversity are often insufficient. Plastic surgery is highly individualized and elective; assembling large cohorts is slow. Studies frequently use only dozens of cases per procedure. Although larger datasets are needed for more accurate results, plastic surgery patients inherently yield small samples [64]. This scarcity is aggravated by the fact that many procedures are uncommon or only performed on one side (e.g. rare reconstructive cases). As an example, the IIIT-Delhi face database, though sizable for a face dataset, still had imbalances. Some surgery types (otoplasty) had far fewer examples, and overall images per patient varied widely [62]. Consequently, AI models trained on such data risk overfitting and may not generalize to broader populations.

Another challenge is data heterogeneity and quality control. Real clinical photos vary in lighting, pose, resolution, and background. Rathgeb et al. report that the IIITD face dataset contained many low-quality images, duplicates, and inconsistent framing [62]. These issues make it hard to isolate the effect of surgery from trivial variations. For instance, Nguyen et al. found that differences in inter-ocular distance and head tilt (unrelated to surgery) could confuse models trained on the dataset [69]. Schlett et al. examined web-scraped facial image datasets and found that hundreds to thousands of duplicate images persisted even after deduplication measures, indicating that unintended factors like repeated head positions or framing can influence model outcomes in face-recognition systems, which could generalize to other applications [70]. This lack of standardization is typical, as photos taken for the same dataset years apart may use different cameras or patient positioning, adding noise. Ensuring uniform photographic protocols and image preprocessing is labor-intensive.

Closely related is the annotation bottleneck. Many AI tasks (e.g. landmark detection, symmetry scoring) require expert labeling. Plastic surgeons must mark key points or rates outcomes, which is time-consuming. As Duong et al. advise, collaboration between surgeons and AI teams is essential: clinicians should "assist AI researchers by granting access to clinical data and aiding in the annotation of datasets for AI model training"[71]. In practice, however, busy surgeons and institutions have limited capacity to label hundreds of images. This often forces studies to rely on automated or semi-automated annotations, which may be less accurate, or to limit study size to what is manually feasible.

Finally, bias and representativeness of collected data pose critical challenges. Many datasets over-represent certain demographics (e.g. young, healthy females for cosmetic procedures) while under-representing others (older patients, males, diverse skin tones). Singh et al. emphasize that "biased datasets can lead to discriminatory practices"[64]. For example, if a plastic surgery AI model is trained mostly on light-skinned patients, it may perform poorly on darker-skinned individuals. Similar issues arise if most data come from one geographic region or ethnicity. Mitigating bias is difficult without deliberately curating balanced data, which in turn may not be feasible given the limited pool. Kenig et al. succinctly state that "the model [is] only as good as the data set used to train [it]"[57], highlighting that any bias or gaps in the data will directly affect model performance.

Real-world plastic surgery datasets face major hurdles: patient privacy/legal issues limit data sharing [63]; small, imbalanced samples reduce model robustness [62], [64]; image quality/standardization varies widely [62]; and annotation/bias concerns complicate model training [71]. These challenges explain why many current studies rely on small internal cohorts and motivate the exploration of synthetic data and novel data-collection strategies.

2.3.3 Importance of Diversity and Scale

Data scale and diversity are critical for the reliable performance of machine learning in plastic surgery decision support. In general, larger training datasets improve statistical robustness and reduce model variance, whereas insufficient data increase overfitting and estimation error [72], [73]. For instance, Althnian et al. found that "large datasets lead to better classification performance and small datasets may trigger over-fitting" [72]. Overfitting occurs when a model learns noise or outliers in the training set rather than the underlying pattern. It is much harder to

avoid overfitting with few examples because the model has freedom to fit every point. Althnian et al. explain that small datasets "cannot generalize patterns in training data" and that over-fitting can even spill over into validation sets [72]. In other words, with limited data, even cross-validation performance may not reflect true generalization.

Similarly, Gütter et al. found that very small training sets produced less robust models, with rapidly rising error when any label noise or variation was introduced [73]. Small-scale datasets yield high error variance. The model's predictions swing dramatically with minor data changes, making the machine learning model unreliable. By contrast, large datasets (with diverse examples) average out random noise, yielding smoother decision boundaries and lower test error [72], [73]. Gütter et al. report that models require "sufficiently large datasets" or they will "likely overfit to the training data" [73]. These principles follow from the bias–variance tradeoff: with small data the estimation variance is high, so predictions vary widely and error rates are elevated. Conversely, increasing sample size tends to reduce variance and improve generalization, lowering the expected error on new patients [72], [73]. Althnian et al. note that increasing dataset size almost always improves predictive accuracy, whereas reducing data can sharply degrade it [72]. In their experiments, classifiers trained on small medical datasets (e.g. a few hundred examples) showed much higher variability and lower accuracy than those trained on larger sets. For example, on one subset with only 98 samples, the average accuracy was as low as 62%, whereas on larger subsets it approached 99%. This broad range (62–99%) illustrates how limited data can lead to extreme performance swings depending on dataset scale.

Biased training sets also inflate sampling error and bias the model. When key subgroups are underrepresented, a model may learn false correlations. Class imbalance is a well-known phenomenon: if one class (e.g. a specific breast shape or facial phenotype) is rare, a machine learning model tends to favor the dominant class, often yielding poor sensitivity for the minority class [72]. In plastic surgery, rare anatomies (e.g. unusual craniofacial variations or large reconstruction cases) may be under-sampled, leading to predictions that systematically omit or misestimate those cases. Similarly, non-uniform sampling of patient demographics (e.g. an imbalance between young vs. older patients) creates sampling bias in the dataset. This means the training data do not fairly represent the target population, causing the model to perform well on common cases but worse on atypical ones. Statistical learning theory dictates that robust generalization requires representative sampling; skewed data can violate assumptions and thus

degrade performance [72]. In practice, algorithms trained on limited or biased data "may perform worse," exhibiting inflated errors even on held-out validation examples [72].

Plastic surgery involves patients of all skin tones, ages, genders, and body types, so an AI model must work well across this spectrum to be safe and fair. However, most existing datasets (in any medical domain) are dominated by light-skinned individuals and narrow demographic slices. For example, facial image datasets often comprise roughly 80% or more lighter-skinned faces [74], leaving darker-skinned patients underrepresented. Buolamwini et al. demonstrated that commercial facial analysis systems trained on such skewed data had higher error on dark-skinned women (up to 34.7% error) than on light-skinned men (as low as 0.8% error)[74]. The authors reported that in standard test sets the error rates in determining the gender of light-skinned men were never worse than 0.8%, whereas for dark-skinned women the error rates were at 20–34.7% [74]. These findings, though in a general vision context, directly illustrate how intersectional underrepresentation (skin tone × gender) can yield inequitable performance.

2.3.4 Synthetic Datasets in Plastic Surgery

The application of artificial intelligence to plastic surgery is limited by the scarcity and sensitivity of real clinical data. Unlike many other domains, large-scale annotated datasets of surgical images, outcomes, or patient records in plastic surgery are rare due to privacy concerns, ethical restrictions, and the heterogeneity of procedures [75], [76]. Because real plastic-surgery images are scarce, recent studies have explored synthetic data generation. For example, Ozmen et al. argue that AI in surgery is limited by "scarcity of large, diverse outcomes datasets." Generating synthetic data using AI can mitigate these barriers by augmenting existing datasets or creating fully artificial cohorts. Synthetic images can capture a broad spectrum of patient anatomies and surgical scenarios, thereby enhancing model training without compromising privacy [75]. For instance, generative adversarial networks (GANs) have been proposed to generate "realistic synthetic data, capturing a broad spectrum of patient profiles and surgical scenarios" for outcome prediction [75]. This approach enables training of AI models on enriched datasets that more comprehensively represent the variability encountered in plastic surgery.

Generative methods extend beyond image augmentation. Recent work emphasizes conditional GANs, variational autoencoders (VAEs), and style-based networks tailored to medical imaging. Ozmen et al. describe using conditional GANs [77] and Wasserstein GANs [78] to model

the statistical properties of surgical data [75]. This approach has been explored in specific domains: for example, Chandaliya et al. developed PlasticGAN, a GAN-based framework that generates post-operative facial images by simulating aesthetic procedures [79]. PlasticGAN is designed to produce realistic "after" surgery faces given "before" inputs, aiding tasks like outcome visualization or recognition. They report that even state-of-the-art face recognition systems struggle with these synthetic post-surgery images, emphasizing their complexity. Flynn et al. generated a repository of 1000 synthetic facial images using an open-source GAN, then analyzed attributes (e.g. age, gender, landmarks) via computer vision [80]. These synthetic faces adhered to aesthetic principles of horizontal thirds and vertical fifths, demonstrating "excellent correspondence" to real photographs [80]. In other words, the GAN-produced faces mirrored the statistical facial proportions of real patients while providing unrestricted research samples. Similarly, Hayajneh et al. introduces CleftGAN, which they used to create photorealistic children's faces with cleft lip deformities [81]. Hayajneh et al. report that CleftGAN can generate "an almost unlimited number of high-fidelity facsimiles of cleft lip facial images" from only 514 real training photos [81].

In breast reconstruction and augmentation, synthetic data enables outcome simulation and risk modeling. Chartier et al.'s BreastGAN trained a convolutional network on paired preoperative and postoperative breast photographs to simulate surgical results [82]. The model produced synthetic postoperative images from preoperative photos, and these AI-generated images were "comparable to real surgical results" according to qualitative evaluation [82]. This demonstrates that synthetic image generation can model breast augmentation outcomes without requiring expensive 3D simulation hardware. Furthermore, Ozmen et al. note that in breast reconstruction contexts, GANs could synthesize data capturing the full "spectrum of patient features, tumor characteristics, and treatment variables" to predict oncologic and aesthetic outcomes [75].

To the best of my knowledge, no large public repository of synthetic plastic-surgery images yet exists, but these proof-of-concept studies suggest potential. In microsurgery research, Ozmen et al. showed that augmenting a small outcomes dataset with high-quality GAN-generated cases can improve AI predictive modeling of surgical success [75]. Similarly, Lim et al. have proposed using text-to-image models (like DALL·E) to simulate cosmetic procedures in underrepresented groups, generating synthetic faces to balance training data [83]. These examples illustrate that the approach of creating synthetic plastic surgery images is emerging, even if fixed "datasets" of

synthetic images are not yet standardized.

Chapter 3. Enhancing Nipple Positioning Accuracy in Chest Reconstruction Surgery: An Automated Machine Learning Approach

This chapter addresses thesis objective 1 of developing a machine learning-based decision support approach to assist in the precise and consistent localization of the NAC in reconstructive chest surgery. This chapter presents a complete version of the paper accepted for publication in the proceedings of the IEEE International Symposium on Medical Measurements and Applications (IEEE MeMeA). It addresses the problem of automated nipple–areola complex (NAC) localization in chest reconstruction surgery. The content is reproduced in its original form without any modifications.

3.1 Abstract

Accurate placement of the Nipple-Areola Complex (NAC) is critical for the aesthetic success of chest reconstruction surgery. Traditional methods rely on the surgeon's experience and subjective judgment, presenting a need for a more objective and reliable approach. This study introduces a machine learning solution to estimate the NAC position on a male chest wall. In this study, we feed a dataset composed of 102 images of 34 male subjects of different ages and body types into an open-source pose estimation algorithm to detect upper body key points that are common to both biological sexes such as shoulders and elbows. A selected subset of those key points is then used to form a normalized feature set which was fed into six regression models: Decision Tree Regressor, Random Forest Regressor, CatBoost Regressor, Multilayer Perceptron Regressor, Linear Regressor, and Support Vector Regressor to predict the NAC position. The lowest mean absolute percentage error (MAPE) between the real and predicted normalized body ratios (0.69%) was achieved with Linear Regression. Subsequently, the chosen model was used to predict the positions of both nipples. A tilt-correction mechanism was introduced to improve accuracy when patient posture was misaligned. Without tilt correction, the MAPE for right and left NAC positions was 1.2% and 0.99%, respectively, which was reduced to 0.75% and 0.63% after correction. Our results suggest that machine learning can be used to assist surgeons performing chest reconstruction surgeries.

3.2 Introduction

The precise localization of the Nipple-Areola Complex (NAC) is vital for the success of chest masculinization surgeries for female-to-male (FtM) transgender patients, as well as individuals undergoing reconstructive procedures following breast cancer surgery. These procedures typically involve the removal of the NAC, followed by its reconstruction at the end of the surgery, positioning the NAC in a location aligned with that of a biological male chest. Ghodratigohar et al.[21] underscore the significance of accurate NAC positioning in achieving a male-appearing chest, a critical determinant of the procedure's aesthetic success. The inherent differences in the location and shape of the NAC between male and female chests necessitate innovative approaches for repositioning, which include the removal of breast tissue, excess skin, and the strategic reshaping of the NAC to minimize scarring while maintaining its masculine characteristic shape.

The complexities of NAC repositioning are highlighted by Kääriäinen et al.[10] who observed that 40.4% of FtM transgender patients required secondary aesthetic correction. Their work divided patients into two groups, on which either a transverse incision approach or a concentric circular approach was used for the mastectomy procedure. It was shown that 27.6% of the concentric circular group and 3.6% of the transverse incision group needed corrections to the shape and size of the NAC, which highlights the necessity for precise NAC repositioning and potential aesthetic revisions. Ayyala et al. [9] emphasized the importance of accurate NAC localization in chest masculinization and breast reduction surgeries to achieve appropriate masculine chest contours. This is typically done using traditional methods, where a triangular reference is manually constructed using the suprasternal notch, the left or right nipple, and the corresponding inframammary fold projected to the midline as vertices. In a follow-up study by Van de Grift et al. [84], 80% of patients who underwent double-incision mastectomy reported satisfaction with chest symmetry, but only 60% were satisfied with nipple symmetry. These findings suggest that improving methods for precise NAC localization is necessary to enhance postoperative outcomes.

Beer et al. [11] observed that the nipple-areola complex in men can be localized by manually measuring some anatomical reference points on the chest. NAC localization was done relative to the lateral edges of the pectoralis major muscles and the midline of the body. The localization was

obtained with the use of a mathematical formula involving measurements like the height of the pectoral muscle fold and the distance to the midline. However, the results indicate that although the NAC could be localized using this formula, its accurate placement could vary depending on factors such as patient's Body Mass Index (BMI), height, and age, and as such, some adjustments might be necessary. For example, patients with higher BMI typically required the NAC to be positioned slightly lower and more laterally compared to patients with a lower BMI.

Atieh et al. [12] proposed another method to localize the NAC based on the golden ratio (Phi), using two manually measured distances; the umbilicus-suprasternal notch distance (U-SN) and the umbilicus-anterior axillary fold distance (U-AX). They also examined the correlation between the calculated NAC coordinates and factors such as weight, height, and BMI and observed that while the inter-nipple distance (N-N) did not significantly correlate with these factors, U-AX and U-SN distances were correlated. This suggested that the vertical positioning of the NAC may vary with patient body size and shape, necessitating manual correction. Due to their manual and labor-intensive nature, traditional methods for NAC localization often prove clinically impractical.

Ghodratigohar et al.[21] introduced an image-based method utilizing an Artificial Neural Network (ANN) for automatic NAC localization. The method involved manually labelling various body landmarks such as the suprasternal notch (SN), anterior axillary fold (AX), and umbilicus (U) and then creating normalized distances using the distances between those landmarks as inputs to the ANN. However, this approach did not achieve full automation of the localization process.

In response to the inaccuracies involved in the traditional NAC localization methods, and the manual labor-intensive nature of the aforementioned approaches, this paper proposes a novel, fully automated machine learning approach for NAC localization, building upon the work of Ghodratigohar et al.[21] while incorporating an anti-tilt mechanism to further refine the precision of surgical planning and execution in chest reconstruction surgeries. Our work builds upon the work of Ghodratigohar et al.[21] and addresses the mentioned limitations, demonstrating that artificial intelligence can be used to support decision-making in chest reconstruction surgeries.

3.3 Methodology

3.3.1 Dataset

The anatomical structure of the biological male chest differs from that of the biological female chest. Since the goal of chest masculinization reconstruction surgery is to create a male-

appearing chest, our machine learning model needed to learn from examples that reflect the final aesthetic goal. To achieve that, we collaborated with the Children's Hospital of Eastern Ontario (CHEO) to collect a dataset featuring 34 biological male subjects. Each subject was photographed in three poses to reflect varying real-world scenarios, as can be seen in **Fig. 3. 1**. This clinical dataset was assembled after obtaining the ethics approval certificate H-03-20-5662.

(a)

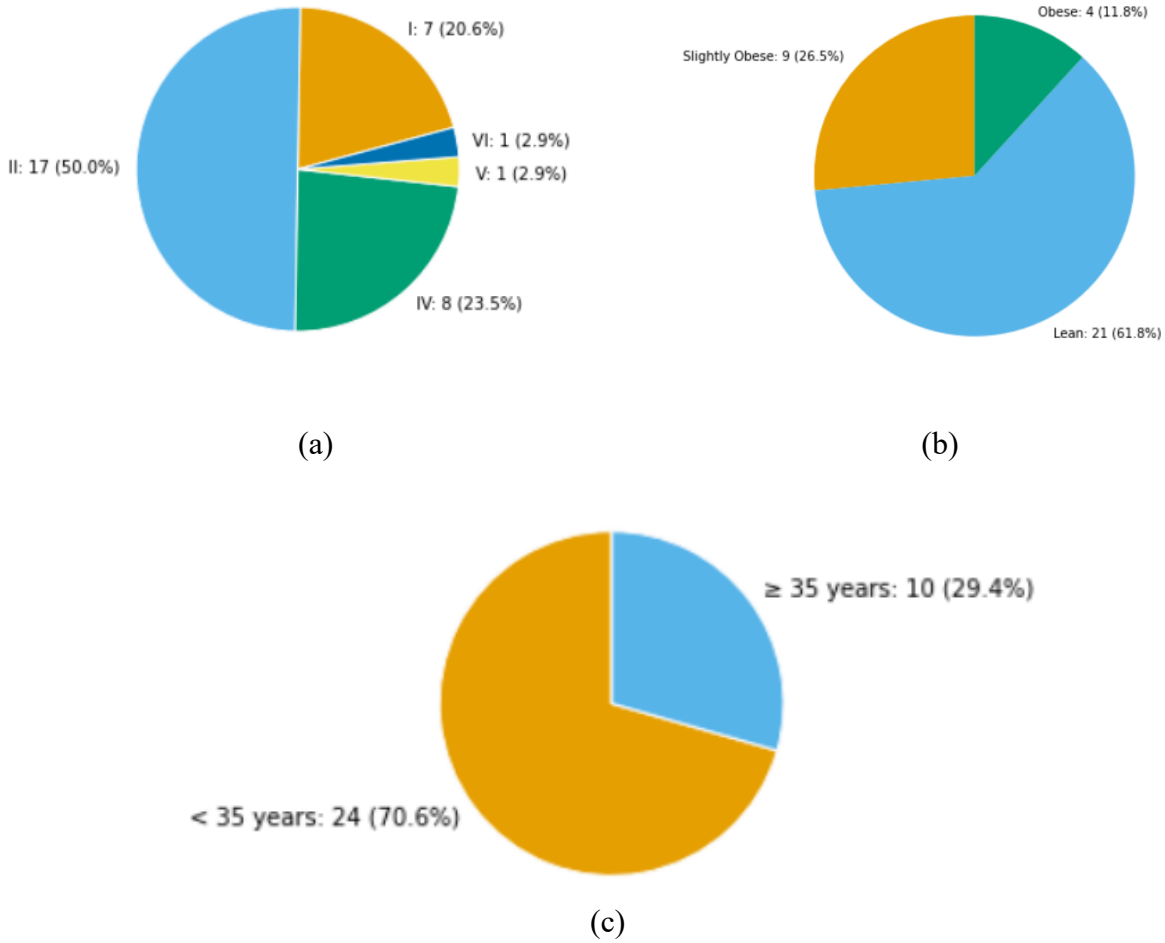
(b)

(c)

Figure 3. 1(a): Patient standing in an anterior-posterior orientation, with arms away (b): Patient standing in an anterior-posterior orientation, with arms to the side (c): Patient in supine position, with arms away.

All of the patient images were compiled to form our dataset, consisting of 102 images in total. This dataset was later split into training (80%) and testing (20%). To prevent overlap between the training and testing sets, we randomly selected seven subjects and allocated all their images exclusively to the testing set. As a result, our dataset was split into 81 images for training and 21 images for testing.

The patients' skin color, body type, and age groups were recorded in the pie charts in **Fig.3.2** below. The patient's skin colors were classified according to the Fitzpatrick scale[85] highlighted in **Fig.3.3.**[85]



(a) (b)

(c)

Figure 3. 2 Pie-Charts showing the distribution of the clinical dataset in terms of a) Skin Color b) Body type and c) Age groups

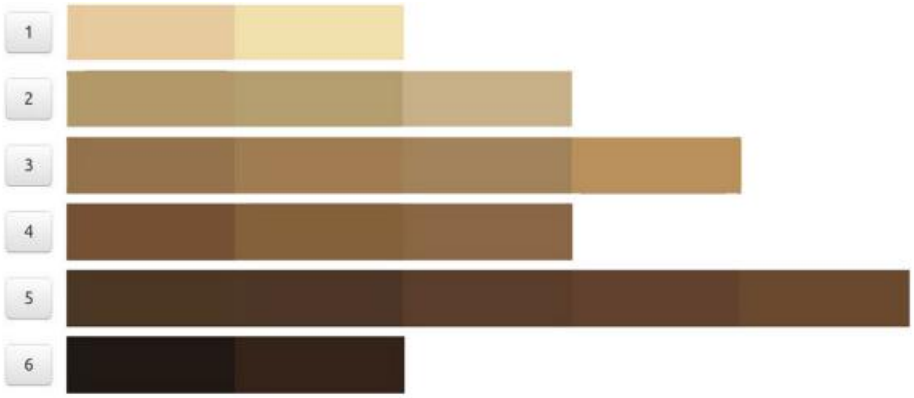
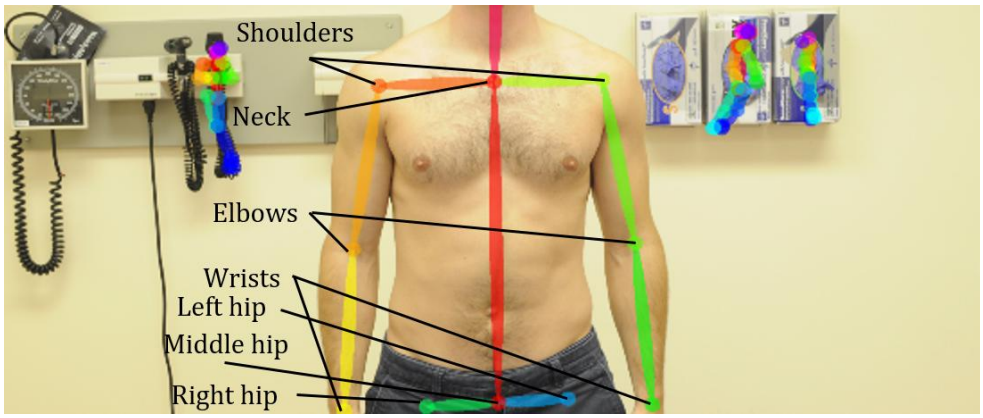


Figure 3. 3 The Fitzpatrick skin phototype scale (Types I–VI), classifying human skin by its response to ultraviolet (UV) exposure[85]

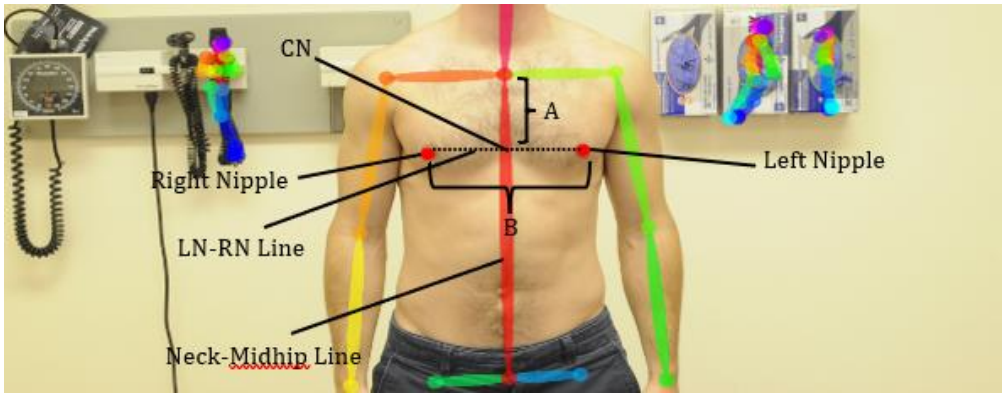
3.3.2 Landmarks

To automate NAC localization in chest reconstruction surgeries, we leveraged OpenPose, an advanced open-source pose estimation library[86] to automate the extraction of anatomical landmarks, an advancement over prior manual selection methods. We fed our data into OpenPose which superimposes color-coded points on the input images (**Fig. 3.4-a**) like shoulders and elbows. For every image, a JSON file that details the coordinates (x and y) of each detected anatomical landmark detected for every detected object in the image is generated. As our goal is to localize NAC coordinates, we subsequently labelled the dataset by manually annotating the coordinates of the left and right nipples (LN and RN) (**Fig. 3.4-b**).

For some images in our dataset, the OpenPose algorithm mistakenly identified some inanimate background objects like a glovebox or a thermometer hanging on the wall of the clinical examination room as human subjects, as can be seen in **Fig. 3.4**. To eliminate these erroneous detections, we leveraged the fact that our object of interest, the patient, is in the foreground while other unwanted objects are smaller and confined to the background. Hence, we calculated the vertical distance between the highest and lowest detected landmarks for each detected object and selected the object with the greatest vertical distance.



(a)



(b)

Figure 3. 4 (a): The raw output from OpenPose detecting 4 objects in the image, one of the objects is the subject we are interested in, the other 3 are background noise. (b): The output after manually adding the Left and Right nipples.

For this study, we used the OpenPose's Body_25 model, which detects 25 anatomical landmarks (**Table 3.1-a**). However, as some of our dataset images exclude the subjects' legs, certain landmarks, including the left and right knees and ankles, are not detectable. Additionally, inconsistencies were noted in the detection of some landmarks, such as the left and right wrists and hips, across images processed by OpenPose.

To address this, a team of three researchers reviewed each of the images and selected six landmarks

that were consistently well-detected across all images (**Table 3.1-b**). Additionally, we manually labelled two additional landmarks: the left nipple and the right nipple, as shown in **Table 3.1-b**.

Table 3. 1-a All the output features from Openpose

All Landmarks from OpenPose
Nose
Left Eye; Right Eye
Left Ear; Right Ear
Neck
Left Shoulder; Right Shoulder
Left Elbow; Right Elbow
Left Wrist; Right Wrist
Left Hip; Right Hip; Mid Hip
Left Knee; Right Knee
Left Ankle; Right Ankle
Left Big Toe; Right Big Toe
Left small Toe; Right small Toe
Left Heel; Right Heel

Table 3. 2-b Reduced Openpose features chosen, in addition to the 3 features we added

Final Chosen Landmarks (FCLs)	
Neck	Detected by OpenPose
Left Shoulder	
Right Shoulder	
Left Elbow	
Right Elbow	
Mid Hip	
Left Nipple (manually added) (LN)	Manually annotated
Right Nipple (manually added) (RN)	

3.3.3 Feature Extraction

Using the coordinates obtained from the pose estimation phase, we calculate the pixel distance between each Final Chosen Landmark (FCL) point and every other point in that dataset. The coordinate system used is defined with its origin at the top-left corner of the image, where the positive x-axis extends horizontally to the right and the positive y-axis extends vertically downward. Images in our dataset may display the subject at different distances from the camera, hence making pixel distances a poor choice of measurement for NAC localization. To address this,

we normalized these pixel distances by dividing them by a reference distance. To select the reference distance, a committee of three researchers reviewed the images and observed that the inter-shoulder distance (ISD), the distance between the left and right shoulders, was well defined in all OpenPose-processed images.

Since the goal is to determine the coordinates of the NAC, it is essential to define both a vertical and a horizontal reference distance to establish a coordinate system on the body. For the y-coordinate, a reference point aligned horizontally with the right and left nipples (RN and LN) was needed. Given that the FCLs include the Neck and Mid Hip along a central vertical axis passing through the chest, an additional point, the center of the inter-nipple distance (CN), was introduced along this same axis (**Figure 3.4-b**). The normalized distance from the Neck to CN, denoted as A, was selected as the vertical reference for the model. For the x-coordinate, the most relevant horizontal distance is the RN-LN distance, as it directly represents the natural separation between the two anatomical landmarks. This normalized distance, denoted as B, was therefore chosen as the horizontal reference as shown in **Table 2**. The rest of the inputs to the model are referred to by W_i, where i is an integer ranging from 0 to 14.

Table 3. 2 Normalized distances as inputs or outputs of the machine learning model

Normalized Distance Alias	**Distance Between Landmarks/ Inter-Shoulder Distance (ISD)**	**Input/Output**
$W_i, i \in \{0,1,\dots,14\}$	All the distances between the Landmarks, normalized by dividing them by ISD, except the two outputs below (15 normalized distances)	Input
A	(Neck – CN) / ISD	Output
B	(RN – LN) / ISD	Output

3.3.4 Machine Learning Model

Following the image dataset processing by OpenPose, six machine learning regression models were evaluated for localizing the NAC. Each model received 34 input features, representing the normalized distances from the selected FCL set. The goal was to determine which model performed best in predicting the NAC position by calculating the MAPE. The six models tested

are Random Forest, Decision Tree Regressor, CatBoost, Multi-layer Perceptron, Linear Regression, and Support Vector Regression (SVR), each chosen for their different approaches to regression tasks.

3.3.4.1 Decision Tree Regressor

The Decision Tree Regressor [87], implemented using the **DecisionTreeRegressor** module in scikit-learn [88], splits the dataset's numerical features into subsets at each node by minimizing variance [89]. This process creates a sequence of decision nodes that yield a predicted value for NAC localization. Decision trees are effective at capturing feature interactions and can adapt well to smaller datasets. However, they are prone to overfitting, especially when dealing with complex or noisy data [90]. This model was chosen as a baseline to evaluate the ability of simple algorithms to extract patterns from the normalized distance data.

3.3.4.2 Random Forest Regressor

The Random Forest Regressor [91], available via the **RandomForestRegressor** module in scikit-learn, builds an ensemble of decision trees and averages their predictions to achieve robustness. This model reduces overfitting seen in individual trees and can capture complex, nonlinear relationships[92] present in the normalized distances derived from the OpenPose output. Random Forest was selected for its ability to handle feature interactions and its robustness, making it a strong candidate for NAC localization.

3.3.4.3 CatBoost Regressor

CatBoost [93], implemented using the **CatBoostRegressor** module, applies gradient boosting to iteratively refine predictions. It is particularly optimized for categorical data but also performs well on numerical inputs such as the normalized distances derived from our dataset. CatBoost's strengths include its resistance to overfitting [94] and its ability to achieve high accuracy with minimal tuning [95]. However, its training process can be resource intensive. CatBoost was selected for its high predictive performance and its efficiency in leveraging the structured numerical dataset.

3.3.4.4 MultiLayer Perceptron (MLP) Regressor

The MLP Regressor [96], implemented with the **MLPRegressor** module in scikit-learn, is a feedforward neural network capable of capturing intricate non-linear relationships [97] in numerical datasets. By adjusting weights iteratively, it learns to approximate complex functions

mapping input distances to NAC coordinates. The model's performance depends heavily on hyperparameter tuning and requires substantial computational resources [98]. MLP was chosen for its flexibility in learning from intricate patterns in the normalized distances, although its sensitivity to dataset size posed a limitation.

3.3.4.5 Linear Regressor

Linear Regressor, implemented using the **LinearRegression** module in scikit-learn, models the relationship between input distances and NAC coordinates as a simple linear function. Its straightforward implementation and minimal computational requirements, combined with its high predictive power [99], make it a promising choice for our NAC localization task.

3.3.4.6 Support Vector Regressor (SVR)

Support Vector Regressor (SVR) [100], available as **SVR** in scikit-learn, is based on support vector machines and finds a hyperplane that best fits the data within a specific margin of tolerance. SVR's ability to handle high-dimensional spaces and outliers [101] makes it useful for prediction tasks, though it may require careful tuning of parameters like the regularization and kernel. SVR was chosen because of its robustness to outliers and its potential ability to generalize well in the normalized distance feature space.

3.3.5 NAC Localization

3.3.5.1 Preliminary Localization

To localize the NAC, normalized distance A (Neck-CN) and normalized distance B (RN-LN) were used. The Neck coordinates provided by OpenPose served as the starting point, and then the coordinates of CN were determined using Equations (3.1) and (3.2).

$$y_{CN} = y_{Neck} + (A \cdot ISD) \qquad (3.1)$$

$$x_{CN} = x_{Neck} \qquad (3.2)$$

In this process, the neck is used as a starting point. Then, the y-coordinate of the CN is calculated by moving vertically downward by the unnormalized Neck-CN distance (3.1). The x-coordinate of the neck is assigned to that of CN (3.2). Next, the x and y coordinates of the nipples (RN, LN) are calculated using Equations (3.3), (3.4), and (3.5).

$$y_{RN} = y_{LN} = y_{CN} \quad (3.3)$$

$$x_{RN} = x_{CN} - \frac{(B \cdot ISD)}{2} \quad (3.4)$$

$$x_{LN} = x_{CN} + \frac{(B \cdot ISD)}{2} \quad (3.5)$$

Here, the LN and RN are assigned the same y-coordinate as that of CN. Then, proceeding horizontally from CN, the left and right nipple x coordinates are found by traversing to the left and to the right by half the predicted unnormalized inter-nipple distance respectively.

While straightforward, this method has a limitation: if the subject in the image is not positioned vertically along the y-axis, but rather appears tilted, the predicted NAC coordinates will reflect this tilt. Consequently, the inter-nipple distance will be angled from the x-axis by the same degree as the subject's tilt from the y-axis.

To further refine the localization in case of tilted images, a tilt-correction algorithm was implemented, detailed in section 3.3.5.2.

3.3.5.2 Tilt Correction

To address the issue of tilted images in the dataset, where the body's orientation deviates from the vertical and horizontal axes (**Figure 3.5-a**), a tilt correction algorithm was developed. This algorithm uses the body's orientation, determined through specific landmarks detected by OpenPose, to localize the NAC regardless of the subject pose (**Figure 3.5-b**).

The following equations provide a more accurate localization of the CN by adjusting for the subject tilt. First, the position of CN (RCN) is calculated by moving vertically along the Neck-Mid Hip line by the predicted Neck-CN unnormalized distance (Eq. 3.7). Then, the left and right nipple coordinates are found by moving horizontally from CN (Eq. 3.9).

To find the CN coordinates, the directional unit vector (rN-MH) along the Neck-Mid Hip Line is calculated from the Neck-Mid Hip vector (RN-MH) using Equation (3.6).

$$\boldsymbol{r}_{N-MH} = \frac{\boldsymbol{R}_{N-MH}}{\|\boldsymbol{R}_{N-MH}\|} \qquad (3.6)$$

CN Coordinates are then calculated using equation 3.7.

$$\boldsymbol{R}_{CN} = \boldsymbol{R}_{neck} + (A \cdot ISD) \cdot \boldsymbol{r}_{N-MH} \qquad (3.7)$$

To accurately identify the left and right nipples' positions correcting for tilt, the normalized left shoulder to right shoulder vector ($\mathbf{r}_{\text{LS-RS}}$) was calculated in Equation (3.8) from the vectors joining the origin to the right shoulder ($\mathbf{R}_{\text{RS}}$) and the left shoulder ($\mathbf{R}_{\text{LS}}$).

$$\boldsymbol{r}_{LS-RS} = \frac{\boldsymbol{R}_{LS-RS}}{ISD} \qquad (3.8)$$

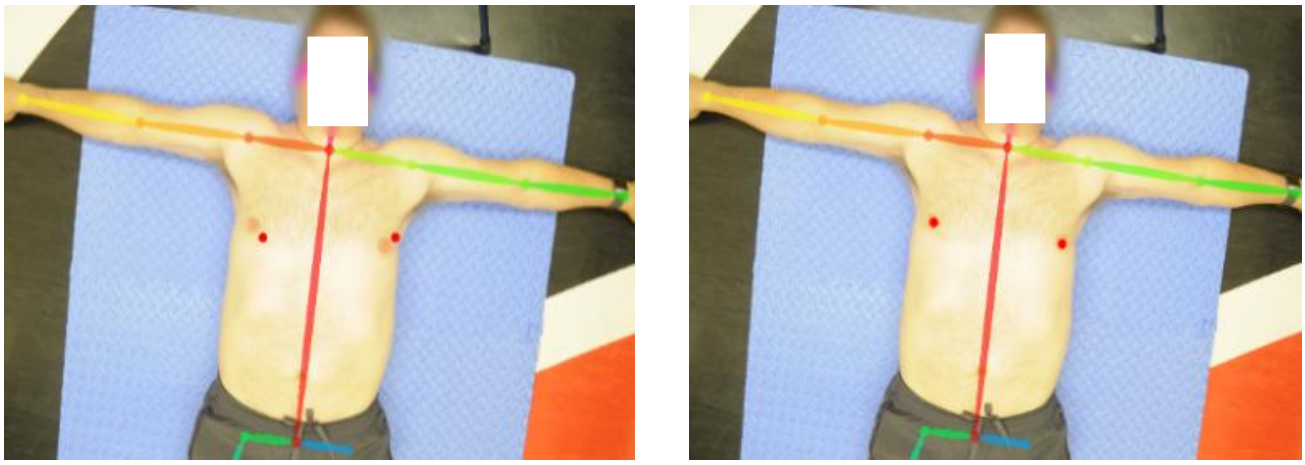
To calculate the coordinates of LN vector ($\mathbf{R}_{\text{LN}}$) and RN vector ($\mathbf{R}_{\text{RN}}$), the results of Equations (3.7) and (3.8) are substituted into Equation (3.9).

$$\boldsymbol{R}_{LN,RN} = \boldsymbol{R}_{CN} \pm \left(\frac{B \cdot ISD}{2}\right) \cdot \boldsymbol{r}_{LS-RS} \qquad (3.9)$$

This approach ensures accurate NAC localization even in images where the subject is not aligned with the image's vertical and horizontal axes, as can be seen in **Fig. 3.5-b**.

3.4 Results

The comparative analysis of all models is shown in Table 3.3, summarizing their performances using the normalized distance mean absolute percentage error (MAPE) metric. This was implemented using Equation (3.10), where n is the number of data points, $Y_i$refers to the observed values, and $\hat{Y}_i$ refers to the values predicted by the models.



(a) (b)

Figure 3. 5 (a) Before applying the tilt-correction mechanism and (b) after applying it

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \cdot 100 \qquad (10)$$

The results shown in Table 3.3 were obtained on the testing set, using an 80%-20% train-test split. Based on these results, the Linear Regression model demonstrated the lowest MAPE, indicating the highest accuracy among the evaluated models. Consequently, Linear Regression was selected as the final model for the application.

Table 3. 3 Normalized distance MAPE comparison across tested models.

Model	**Normalized Distance MAPE**
Random Forest	2.2%
Decision Tree Regressor	5.19%
CatBoost	2.64%
Multi-layer Perceptron	9.75%
Linear Regression	0.06%
SVR	6.95%

One likely reason for linear regression outperforming the other models is the relatively

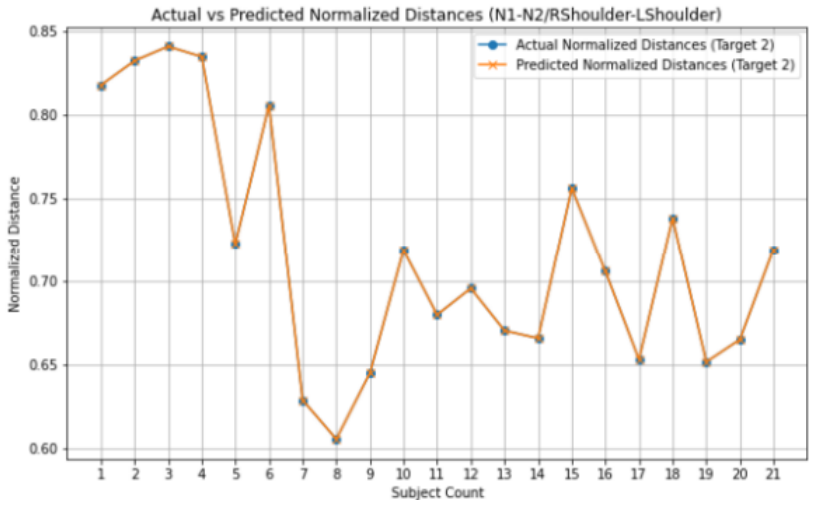
small size of our dataset. Linear regression is known to perform robustly on smaller datasets because of its simplicity and low variance[102], whereas more complex models such as tree-based methods or neural networks tend to require larger sample sizes to generalize effectively.

A second reason could relate to the nature of the relationships within the data itself. More complex models attempt to capture nonlinear or higher-order interactions in addition to linear patterns. While this flexibility is advantageous when such complexity is present, it can introduce unnecessary noise or overfitting when the underlying relationships are predominantly linear. Our results suggest that the distances and ratios derived from the OpenPose keypoints are best described through linear associations, which would explain why linear regression consistently outperformed the other methods in this study.

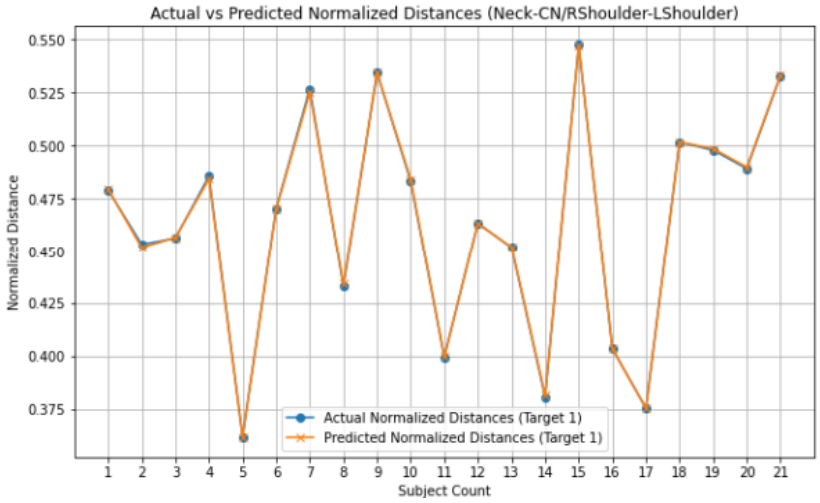
For the selected Linear Regression model, the MSE was calculated for the right and left nipple (RN and LN) position in pixels. Without applying the tilt correction algorithm, the RN MAPE was 1.2% and the LN MAPE was 0.99%. With the tilt correction algorithm, the RN MSE was reduced to 0.75% and the LN MSE was reduced to 0.63%.

To improve the reliability of the model's generalizability given the small nature of the dataset, a 5-fold grouped cross-validation was used, using MAPE as the scoring metric. This choice of cross validation groups each of the three images of the same patient into one group, making 21 groups in total, then splits the entire groups into folds, ensuring no overlap between the training and the testing dataset. This approach also reduces the risk of overfitting and assesses the model's performance across different data subsets to estimate the RN-LF and Neck-CF normalized distances accurately. **Fig. 3.6** shows the MAPE of the two model's output normalized distances across all the test subjects, which are 0.12% and $1.40e^{-14}$ % respectively.

In addition to the normalized distances, **Fig. 3.7** presents a comparison between the actual and estimated NAC coordinates for each subject's images (4288 x 2848 pixels) in the dataset. The results show a MAPE of ***0.63%*** for the left NAC position and 0.75% for the right NAC position. Additionally, the average inter-nipple distance error was calculated as $\boldsymbol{3.61e^{-14}}$%, and the average angle between the actual and predicted NAC positions was found to be ***1.42*** degrees.



(a)



(b)

Figure 3. 6 Comparison of actual and estimated data for: (a) normalized distance C MAPE of 0.12% and (b) normalized distance B MAPE of 1.4e-15

3.5 Discussion and Conclusion

This study introduces a machine learning-based approach to improve the accuracy of NAC localization in chest reconstruction surgeries. Results indicate that the Linear Regression model demonstrated the lowest prediction MAPE compared to other tested models. The integration of a tilt correction algorithm effectively addressed issues related to image orientation, making the model applicable to a range of imaging conditions beyond ideal setups.

The Linear Regression model was trained using normalized distances between key anatomical landmarks, such as the Shoulders, Elbows, Neck, and Mid Hip, allowing it to predict two essential normalized distances: the vertical normalized distance between the neck and the center of the inter-nipple distance and the horizontal normalized distance between the left and right nipples. From these predicted normalized distances, the NAC coordinates were found. The model was trained on 80% of the dataset provided by CHEO hospital, consisting of 34 male subjects, each with 3 images. When tested on the remaining 20% of the dataset, the model achieved a MAPE of 0.06% for the output normalized distances and a MAPE of 0.69% between actual and predicted nipple positions.

These findings have significant implications for fields such as chest masculinization surgery and breast cancer reconstruction, where accurate NAC placement is vital for patient satisfaction and overall surgical success. However, a limitation of this study is the small dataset size, with only 3 images per subject for 34 subjects. This limited dataset restricts the ability to capture the diversity of body types, sizes, and ethnicities, which is essential for broader applicability. Future work should focus on expanding the dataset to include a wider range of body types and demographics to improve generalizability. This study highlights the potential of machine learning to support surgeons in making informed decisions in the planning phase of reconstructive surgery, contributing to improved outcomes in NAC localization.

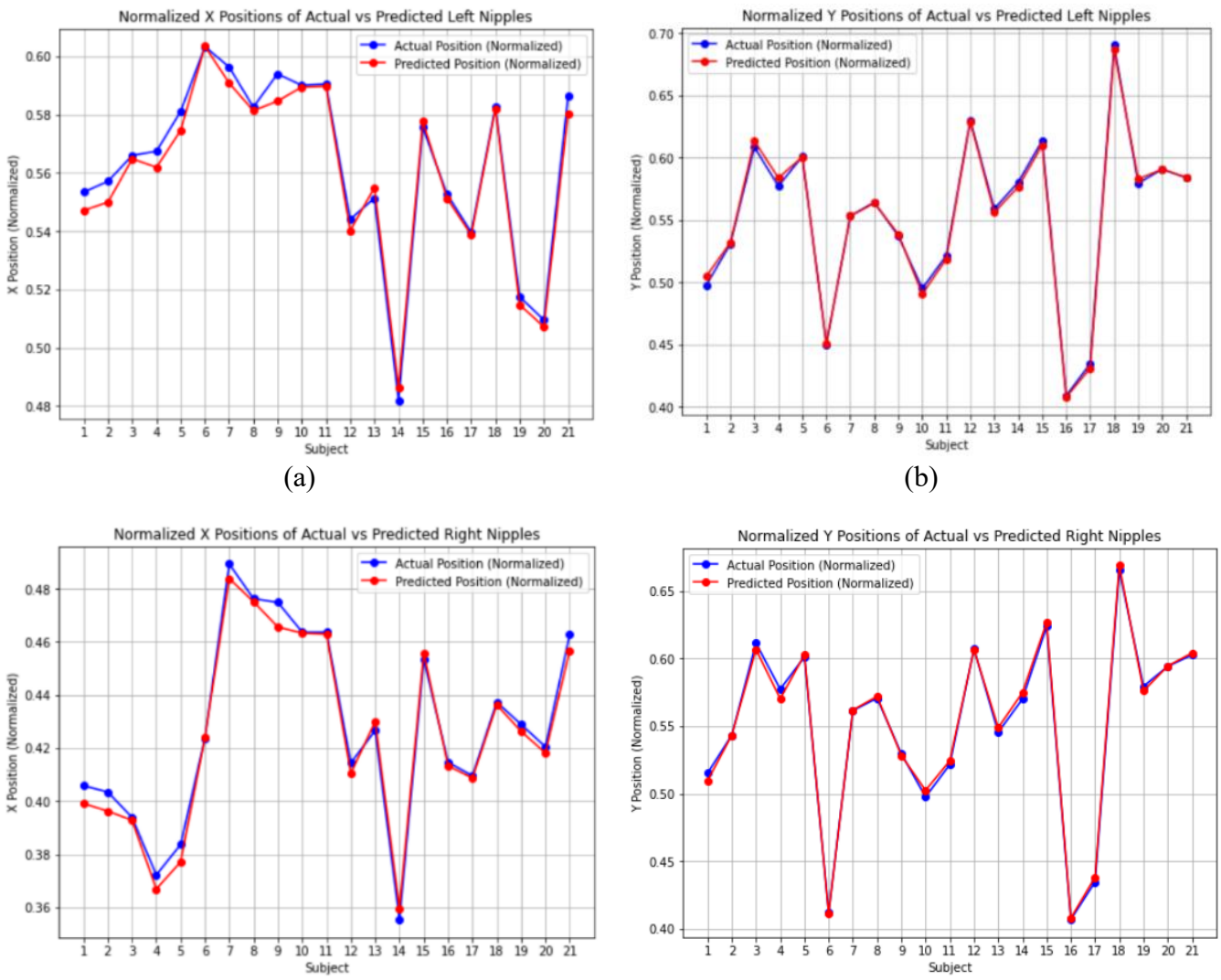


Figure 3. 7 (a) and (b) comparison of actual and estimated x and y coordinates of Left nipple. (c) and (d) comparison of actual and estimated x and y coordinates of Right nipple

While it may appear useful to compare our machine learning framework with a hybrid approach that combines OpenPose-based feature extraction and geometric rule-based methods from the literature, such a comparison is not feasible. The geometric methods described by Beer et al.[11] and Ghotradigohar et al.[21] rely on landmarks such as the xiphoid, the anterosuperior iliac spine, and the anterior axillary folds. These landmarks are not captured by OpenPose, which is limited to a standardized set of keypoints (primarily joints and facial points) that do not encompass the anatomic references required for geometric calculations of NAC position. Without these landmarks, implementing those geometric approaches within an OpenPose-driven pipeline is impractical.

Moreover, attempting to supplement OpenPose with manual annotation of the missing landmarks would contradict the central aim of this research: to provide an automated, reproducible method for NAC localization that reduces dependence on manual expertise. The strength of the presented machine learning approach lies in its ability to work directly with the reliably available keypoints from OpenPose, without requiring additional operator input. Introducing manual intervention would undermine both scalability and the clinical value of automation. For these reasons, a hybrid comparison would not only be impractical but also misaligned with the scope and objectives of this study.

Chapter 4. Enhancing NAC Localization Accuracy via Synthetic Image Augmentation Using Large Language Models and Diffusion Models

This chapter presents an investigation into the use of synthetic data to improve the generalizability of the selected machine learning model in chest reconstruction surgery decision support. It builds on Chapter 3 and contributes to the development of data-driven tools for clinical decision-making. This chapter addresses thesis objective 2 of evaluating the effect of augmenting a real clinical dataset with synthetic images on the performance of the decision support approach.

4.1 Abstract

Accurate localization of the Nipple-Areola Complex (NAC) is essential for achieving optimal outcomes in chest reconstruction surgery. However, the development of machine learning models for this task is limited by the scarcity and demographic narrowness of clinical imaging datasets. In this study, we address this challenge by introducing a synthetic dataset generated through large language model (LLM)-assisted prompt engineering and text-to-image diffusion models. We evaluate the impact of synthetic augmentation on a linear regression-based NAC localization model using normalized anatomical distances extracted from OpenPose. Three experiments assess model performance under increasingly diverse and challenging test conditions: real-to-real, real-to-synthetic, and real-to-out-of-distribution image generalization. Results show that moderate augmentation with synthetic data can slightly improve model performance on real images, while the model's error on challenging, real-world images dropped from 1.63% to 0.91% after introducing 130 synthetic samples, representing a 44% reduction in prediction error. These findings demonstrate the potential of combining LLM-guided prompt generation with diffusion-based synthetic image creation to enhance the robustness and generalizability of machine learning models in clinical applications.

4.2 Introduction

Accurate Nipple-Areola Complex (NAC) localization is essential for chest reconstruction surgeries, including those for female-to-male (FtM) transgender patients and individuals undergoing breast cancer reconstruction. NAC reconstruction was introduced in the 1940s and has evolved alongside breast reconstruction [103]. It involves the reconstruction of both the nipple and the areola. Ideal reconstruction aims for symmetry in position, size, shape, texture, and pigmentation, as well as permanent projection [104].

Accurate NAC localization is not only a technical necessity but also plays a critical role in patients' psychological well-being and satisfaction with surgical outcomes. In chest masculinization surgeries, particularly for FtM transgender individuals, the appearance of the chest, including symmetric and appropriately positioned NACs, significantly affects body image and gender affirmation [13]. Inaccuracies in NAC placement can lead to dissatisfaction and the need for revision surgeries [10].

Traditional methods are subjective and heavily rely on the surgeon's experience. Recent advances in machine learning have introduced automated approaches to NAC localization [21]. Our previous work proposed a fully automated regression-based model to enhance the precision of surgical planning and execution [105].

The shortage of annotated clinical data is a common problem when training a machine learning model in the medical field, largely due to privacy regulations, data heterogeneity, and annotation costs [106]. In our prior work [105], we used a dataset collected in collaboration with the Children's Hospital of Eastern Ontario (CHEO). This dataset includes 102 images from 34 biological male subjects. Male subjects were used since the anatomical structure of the biological male chest is distinct from that of the biological female chest and the target application of this work is chest masculinization reconstruction surgery. While our previous model showed strong predictive accuracy on the collected dataset, its generalizability remains in question due to the limited size of the dataset. The dataset lacked sufficient variability in skin tone, chest size, and posture, which are critical to building a robust model that can adapt to real-world clinical diversity. As such, extending the dataset beyond a narrow demographic profile became a crucial objective

to enhance our model.

To address this challenge, this paper explores the use of synthetic image data generated by text-to-image diffusion models via Large Language Model (LLM)-engineered prompts for NAC localization. In this paper, we evaluate multiple LLMs for prompt generation and multiple text-to-image diffusion models for image generation. Furthermore, we investigate the role of prompt engineering in improving image realism and diversity. The synthetic dataset is then used to augment our existing training set and to assess the impact of this augmentation on the performance of the NAC localization model.

4.3 Background and Related work

4.3.1 Synthetic Data in Medical Imaging

There are several significant challenges in collecting real clinical images for training machine learning models [4], [22], [106]. One major challenge lies in the source of the images [106], which typically falls into two categories: research-collected images and routinely collected clinical images.

Research-collected images are usually captured under controlled conditions with participant consent. However, such datasets are often small in size and lack diversity, which limits the generalizability of the models trained on them. This limitation arises because the acquisition environment is highly constrained and does not reflect real-world variability.

On the other hand, routinely collected clinical images are those generated during routine clinical care. They are more abundant but come with their own set of challenges. These images are highly heterogeneous due to differences in imaging equipment, vendor specifications, and clinical settings, leading to inconsistency in the data. Additionally, leveraging such data requires complex ethics approvals and extensive anonymization procedures.

Privacy concerns also play a major role. Regulations such as the General Data Protection Regulation (GDPR) in Europe, the Health Insurance Portability and Accountability Act (HIPAA) in the US, and the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada strictly govern how patient data can be processed and shared [106]. Even when data is

anonymized, certain types such as 3D reconstructions of head Magnetic Resonance Imaging (MRI)/Computed Tomography (CT) scans, still pose a risk of re-identification. Researchers must also navigate complex data governance protocols and obtain multiple approvals before accessing clinical data [4].

Augmenting real datasets with synthetic data, whether by completely generating new samples or augmenting data through techniques such as rotating and flipping, presents a promising solution to many of these challenges[15]. It eliminates patient re-identification risk, enabling data sharing without violating privacy regulations. Furthermore, it enhances model generalizability by introducing variations not found in limited real datasets.

Recent studies have demonstrated that in some cases, synthetic data alone can be used to train effective machine learning models [107]. Synthetic medical images are becoming increasingly common in areas such as X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET) [15]. Synthetic data has also been used in applications like body part localization and segmentation. For example, Saviolo et al. [108] developed a framework for human body part segmentation using automatically generated synthetic images of human limbs, showing promising applications in surgery and medical image analysis.

However, for anatomical landmark detection, there is still under-representation in medical datasets due to large anatomical variability, which increases the difficulty of accurate landmark localization [109]. This limitation was also highlighted in [105]as a primary challenge in training a robust machine learning model for NAC localization in chest reconstruction surgeries. The lack of dataset diversity, especially in terms of body types, sizes, and ethnicities, limits model generalization and hinders clinical applicability.

4.3.2 Image-Generative Models and Automatic Prompt Generation

The use of synthetic data in medical imaging has evolved from early methods involving basic image augmentation to more advanced approaches that leverage generative models such as Generative Adversarial Networks (GANs) to produce realistic data [110]. However, as noted in [111], GANs tend to produce images with lower fidelity and limited diversity. To address this limitation, the authors proposed a class-conditioned diffusion model as a more effective

alternative.

Recent advancements in text-to-image generation have significantly improved the quality and realism of synthetic human images [112]. For instance, Cevik et al. [23] found that diffusion-based models such as DALL·E2 and Midjourney produced images with greater visual diversity in gender and skin tone compared to those generated by GANs. Their study assessed AI-generated depictions of surgeons and noted reduced representational bias in diffusion model outputs, suggesting improved demographic inclusivity in image generation. In this study, four diffusion models were tested for image generation: DALL·E 2, Midjourney, Meta AI's Emu, and Google's Imagen 4.

The strength of using LLMs in the process of image generation lies in their capacity for effective prompt engineering, which plays a crucial role in enhancing image quality. Well-crafted prompts lead to more detailed and contextually appropriate images by incorporating background information and common-sense relationships [113]. Building on the importance of prompt quality, recent research has explored automatic prompt generation, a process in which language models autonomously construct, refine, or optimize textual prompts based on minimal user input or predefined objectives[114]. Unlike manual prompt engineering, which relies on human expertise to craft effective instructions, automatic prompt generation uses machine learning techniques, such as reinforcement learning, supervised fine-tuning, or prompt templates, to produce more descriptive, contextually rich, and model-compatible queries[114]. This capability is particularly valuable in image generation tasks, where subtle differences in phrasing can lead to substantial variations in visual output. For example, Cao et al. introduced BeautifulPrompt, a deep generative model that transforms simple raw descriptions into high-quality prompts, resulting in more aesthetically pleasing and diverse images through Stable Diffusion and similar models [115]. Rosenman et al. proposed NeuroPrompts, an adaptive framework fine-tuned to enhance user prompts by employing constrained decoding with language models, thereby improving image fidelity and alignment with human expectations[116]. Furthermore, Yaru Hao et al. developed a prompt optimization method combining supervised fine-tuning and reinforcement learning (called "prompt adaptation") that outperforms manual prompt crafting on metrics of aesthetic quality and user preference[114]. Collectively, these approaches demonstrate that automatic prompt engineering can reduce human effort, increase image quality, and support scalability across

varying models and tasks. In this study, three large language models were tested for prompt engineering: ChatGPT-3.5, Llama 3.2, and Gemini 1.0 Pro.

4.3.3 Nipple Areola Complex (NAC) Localization

Nipple-Areola Complex (NAC) and its accurate placement are critical for the success of chest reconstruction surgery. It is vital for the success of chest masculinization surgeries for female-to-male (FtM) transgender patients as well as individuals undergoing reconstructive procedures following breast cancer surgery. Ghodratigohar et al. [21] introduced an image-based method using an Artificial Neural Network (ANN) for automatic NAC localization. This method involved manually labeling body landmarks and creating normalized distances as inputs to the ANN model. However, this approach did not achieve full automation of the localization process. The work in [105] built upon this method by introducing a machine learning solution to improve the accuracy of NAC localization. The work also incorporated an anti-tilt correction mechanism to improve accuracy when the patient's posture was misaligned. One of the main challenges facing the proposed method is the limited size of the existing dataset, which was composed of 102 images of 34 biological male subjects, each photographed in three poses to reflect varying real-world scenarios.

4.4 Synthetic Data Generation

4.4.1 Model Selection for Image Generation

To augment our clinical dataset with synthetic images, a selection process was conducted to identify a large language model capable of generating photorealistic images aligned with the study's requirements. The task was defined as generating an image of a male subject who is shirtless and facing the camera. The subject's chest had to be exposed, and the lower body had to be clothed in pants. One body type and one ethnicity were selected for homogeneity across all models to be tested. Five default prompts were used for uniformity across four selected models as shown below, Midjourney, Imagen4 through Google Gemini 1.0 Pro's web interface, Emu through the Llama 3.2 web interface, and DALL-E2 through the OpenAI ChatGPT 4 web interface.

1. Generate a high-resolution image of an obese African American man, shirtless and standing frontally. He is wearing pants and the full upper body should be visible

2. Generate a front-facing image of an obese African American male subject who is shirtless and clothed from the waist down, with the full upper body and head visible.
3. Generate a photorealistic image of an African American male, obese build, standing shirtless and facing the camera. His bare skin is visible from head to crotch, and he is wearing dark pants.
4. Generate a photorealistic image of a male subject facing the camera. The male subject is shirtless, and his bare skin is exposed at the top. He is wearing pants. The male subject is African American and obese.
5. Generate a high-resolution image of an obese African American man, shirtless and standing frontally. He is wearing pants and the full upper body should be visible.

In this section, we will show the results of the models when prompted with prompt 4 listed above. The model selection process is outlined in **Fig. 4.1** below.

Figure 4. 1 Model selection process for synthetic image generation

4.4.1.1 Midjourney

Midjourney was chosen due to its widespread popularity and high-quality output in the domain of creative image generation.[117]. Eom et. al ranks Midjourney among the top four used generative AI tools globally.[117] The default prompt was issued to Midjourney and the model returned an image that was stylistically consistent with digital illustration rather than true photorealism. The subject was rendered with smooth skin gradients and non-photorealistic facial proportions, indicating that the image was more aligned with cartoon-like digital artwork (**Fig. 4.2**). A review was conducted by a committee of three researchers, who independently determined that the image lacked sufficient anatomical realism to be used for data augmentation in medical applications. As a result, Midjourney was rejected from further use.

Figure 4. 2 Output synthetic image generated by Midjourney using default prompt 4

4.4.1.2 Imagen4

While Imagen4, accessed through the web interface of Gemini 1.0 Pro, is generally capable of generating images from text prompts[118], it failed to produce an output when provided with the standardized instruction used for model comparison. Instead, the interface returned the message: *"I'm a text-based AI and can't assist with that."* This response suggests that the system's automated content safety filters prevented it from processing prompts involving shirtless human subjects, even when framed in a clinically relevant context. Since no image was generated, Imagen4 was excluded from further consideration.

4.4.1.3 DALL-E2

DALL-E2, accessed through the ChatGPT4 web interface, has been reported to have good human image-generation capabilities [119]. When prompted through the interface, ChatGPT declined to generate an image in response to the default prompt. The response consistently returned by the interface was: "*I couldn't generate the image because the request violates our content policies. If you'd like, I can help create a different image—just let me know what you'd like to see instead.*"

This behavior reflects the application of OpenAI's safety and content policy enforcement mechanisms. Due to this restriction, ChatGPT was excluded from the model selection process.

4.4.1.4 Emu

Emu was accessed through the Llama 3.2 interface. The interface was prompted with the same default prompt, and the model returned an image that was determined to be visually realistic by a committee of three researchers. The generated image presented anatomically coherent body proportions, realistic lighting and skin texture, and accurate rendering of the specified demographic features (**Fig. 4.3**). The outputs were reviewed by the same three-member committee and were unanimously deemed suitable for use in synthetic dataset construction. Based on this evaluation, Emu was selected as the image generation model for the synthetic data augmentation pipeline.

Figure 4. 3 Output synthetic image generated by Emu using default prompt 4.

4.4.2 Prompt Engineering

Following the selection of Emu as the interface model for image generation, an automatic prompt generation procedure was implemented to replace the initial default prompt with a set of more diverse and dataset-relevant instructions. This procedure was structured into two complementary phases: text-based prompt engineering and image-based prompt engineering (**Fig. 4.4**). In the text-

based phase, candidate language models were prompted to generate a text-to-image prompt meeting predefined requirements; the resulting prompt was then used to generate an image via the Emu diffusion model. In the image-based phase, the models were provided with a real image from the clinical dataset and tasked with producing a descriptive caption. The LLM was then tasked to refine the caption into a prompt suitable for Emu-based image synthesis. The outputs from both phases were evaluated, and the more effective prompt (determined by visual analysis of the resulting images by a committee of three researchers) was selected for generating the final synthetic dataset.

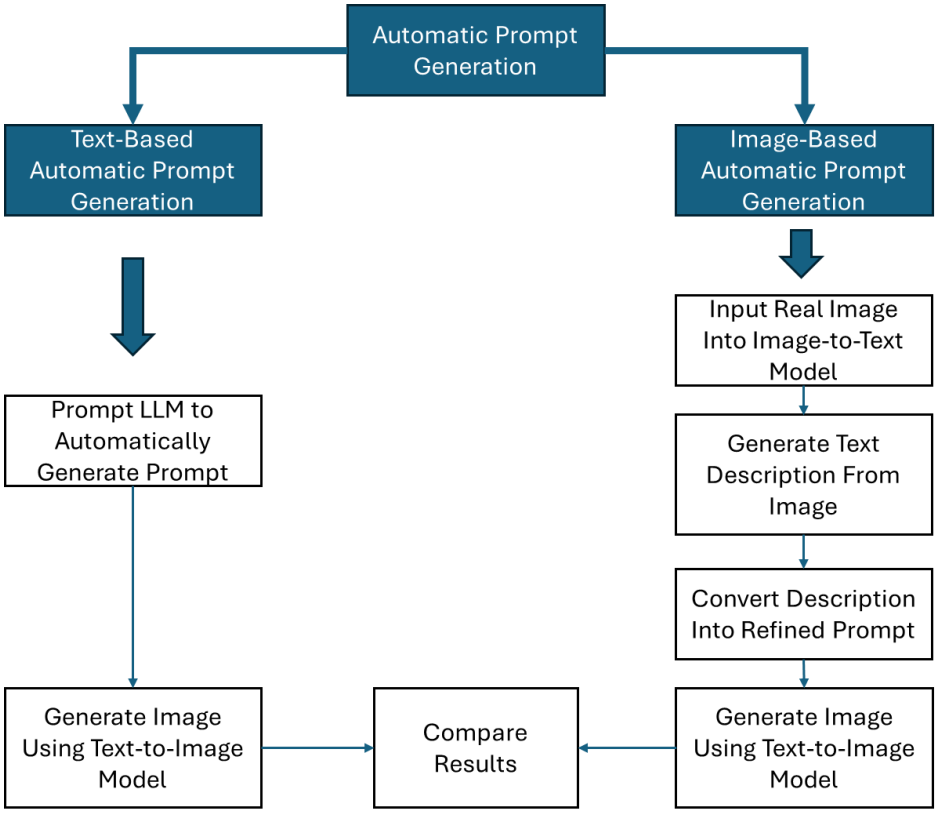


Figure 4. 4 The prompt engineering process for synthetic image generation.

4.4.2.1 Phase 1: Text-Based Automatic Prompt Engineering

In the first phase, the goal was to explore the effectiveness of the selected models in generating prompts for image synthesis based on high-level textual instructions. The three candidate LLM's -ChatGPT, Gemini, and Llama 3.2- were queried using an instruction to generate 10 prompts each that could produce images of shirtless male subjects facing the camera with varied ethnicities, ages, and body types.

For example, when each model was asked to produce a descriptive prompt that, when given to an LLM, would prompt it to generate an image of a shirtless white male, 25 years of age with an obese build, standing face towards the camera. Gemini 1.0 Pro responded with a prompt that restated the input without offering any additional descriptive detail "*A shirtless, 25-year-old white male with an obese build stands facing the camera*". This output closely mirrored the original instruction and did not introduce any elements, such as background, lighting, posture, or contextual details, that would enhance the effectiveness of the prompt for image generation.

ChatGPT-4 generated a slightly more elaborate version of the input, adding background and lighting context. However, the increase in descriptive value remained marginal: "*Generate a photorealistic image of a 25-year-old white male subject standing and facing the camera. The subject has an obese body build and is shirtless, with his upper torso exposed. He is wearing pants and standing in a neutral pose against a plain studio background with even lighting.*" Although more elaborate, the prompt still did not offer significant enhancements to realism or visual instruction.

Llama 3.2 responded with the following prompt: "*Create an image of a 25-year-old White man with an obese build standing shirtless in front of a plain background, facing directly towards the camera with a neutral expression.*" Similar to the aforementioned responses, this response included only minimal contextual or anatomical expansion beyond the original prompt.

These results suggest that text-based prompt engineering provided little to no added value over human-authored prompts. The output from all three models was constrained by the semantic limitations of the original instruction and failed to introduce clinically or visually relevant details that could guide image generation to the desired dataset-level quality.

4.4.2.2 Phase 2: Image-Based Automatic Prompt Engineering

To overcome the limitations identified in Phase 1, an image-based prompt engineering approach was implemented. Real images from our previously collected dataset in collaboration with CHEO were uploaded to ChatGPT and Gemini Web Interfaces, the only systems at the time with image input capabilities. The two models where then asked to generate a detailed description of the uploaded images. Gemini returned a restriction message: "*Sorry, I can't help with images of people yet.*" As such, no descriptive output was obtained from Gemini.

ChatGPT, in contrast, returned highly detailed descriptions. The descriptions included subject gender, pose, posture, facial expression, body type, clothing, accessories, and environmental context, such as the presence of medical equipment mounted on the clinical room wall. These descriptions were parsed and refined to exclude unrelated information, such as glasses, clothing, and clinical equipment, while retaining critical physical descriptors such as gender, pose, body type, and orientation.

ChatGPT was then asked to transform this refined description into a generalized prompt template that incorporated variability across three dimensions:

- Age: 25 years - 55 years
- Body Type: Obese, Skinny, Lean
- Ethnicity: Black, White, East Asian, Hispanic, Indian, Middle Eastern

This resulted in a list of prompts representing every possible permutation of the three dimensions mentioned above. Each prompt was manually reviewed, and inconsistent descriptions- such as those describing subjects having both a "*obese build*" and a "*flat midsection*"- were excluded. This resulted in 36 unique prompts listed in section 4.4.2.3. These prompts were then used cyclically to generate images for the new synthetic dataset (**Fig. 4.5**).

Figure 4. 5 Output synthetic image generated through Llama 3.2 web interface (diffusion model: Emu) using the engineered prompt.

4.4.2.3 Engineered Prompts List

The following is the list of 36 unique prompts generated after following the prompt engineering methodology described in section 4.4.2.2.

1. Generate an image of a 25-year-old Black man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
2. Generate an image of a 25-year-old Black man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the

subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

3. Generate an image of a 25-year-old Black man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
4. Generate an image of a 25-year-old White man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
5. Generate an image of a 25-year-old White man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
6. Generate an image of a 25-year-old White man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

7. Generate an image of a 25-year-old East Asian man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
8. Generate an image of a 25-year-old East Asian man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
9. Generate an image of a 25-year-old East Asian man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
10. Generate an image of a 25-year-old Hispanic man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
11. Generate an image of a 25-year-old Hispanic man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim

and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

12. Generate an image of a 25-year-old Hispanic man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
13. Generate an image of a 25-year-old Indian man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
14. Generate an image of a 25-year-old Indian man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
15. Generate an image of a 25-year-old Indian man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly

illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

16. Generate an image of a 25-year-old Middle Eastern man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
17. Generate an image of a 25-year-old Middle Eastern man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
18. Generate an image of a 25-year-old Middle Eastern man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
19. Generate an image of a 55-year-old Black man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

20. Generate an image of a 55-year-old Black man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
21. Generate an image of a 55-year-old Black man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
22. Generate an image of a 55-year-old White man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
23. Generate an image of a 55-year-old White man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
24. Generate an image of a 55-year-old White man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined

upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

25. Generate an image of a 55-year-old East Asian man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
26. Generate an image of a 55-year-old East Asian man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
27. Generate an image of a 55-year-old East Asian man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
28. Generate an image of a 55-year-old Hispanic man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm,

evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

29. Generate an image of a 55-year-old Hispanic man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
30. Generate an image of a 55-year-old Hispanic man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
31. Generate an image of a 55-year-old Indian man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
32. Generate an image of a 55-year-old Indian man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

33. Generate an image of a 55-year-old Indian man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
34. Generate an image of a 55-year-old Middle Eastern man with an obese build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a broad and smooth upper body and a rounded and smooth midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
35. Generate an image of a 55-year-old Middle Eastern man with a skinny build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a slim and undefined upper body and a flat and slim midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.
36. Generate an image of a 55-year-old Middle Eastern man with a lean build standing in a clinical examination room. The man has a neutral expression and is standing facing forward with his arms resting naturally by his sides. He is shirtless, showcasing a moderately defined upper body and a toned but balanced midsection, and is wearing black drawstring pants. The room has a pale-colored wall, and the lighting is bright and warm, evenly illuminating the subject. Focus on the subject's overall body type and neutral demeanor, keeping the setting simple and clean to emphasize the subject.

4.4.3 New Dataset Composition

The obtained synthetic dataset was generated by cycling through the 36 unique prompts listed in section 4.4.2.3 to produce a balanced and diverse collection of images representing six ethnicities (**Fig. 4.6a**), three body types (**Fig. 4.6b**), and two age groups (**Fig. 4.6c**). Each prompt instructed the model to generate a shirtless male subject standing facing the camera in a clinical examination room with his arms resting naturally by his sides. The pose, facial expression, and room setting were kept consistent to ensure control over non-essential variables.

However, the distance of the subject from the camera was not constrained in the prompt design. This introduced variability in the framing of the generated images. Some images are close-ups that exclude the subject's head, while others capture more of the upper body with varying degrees of anatomical detail.

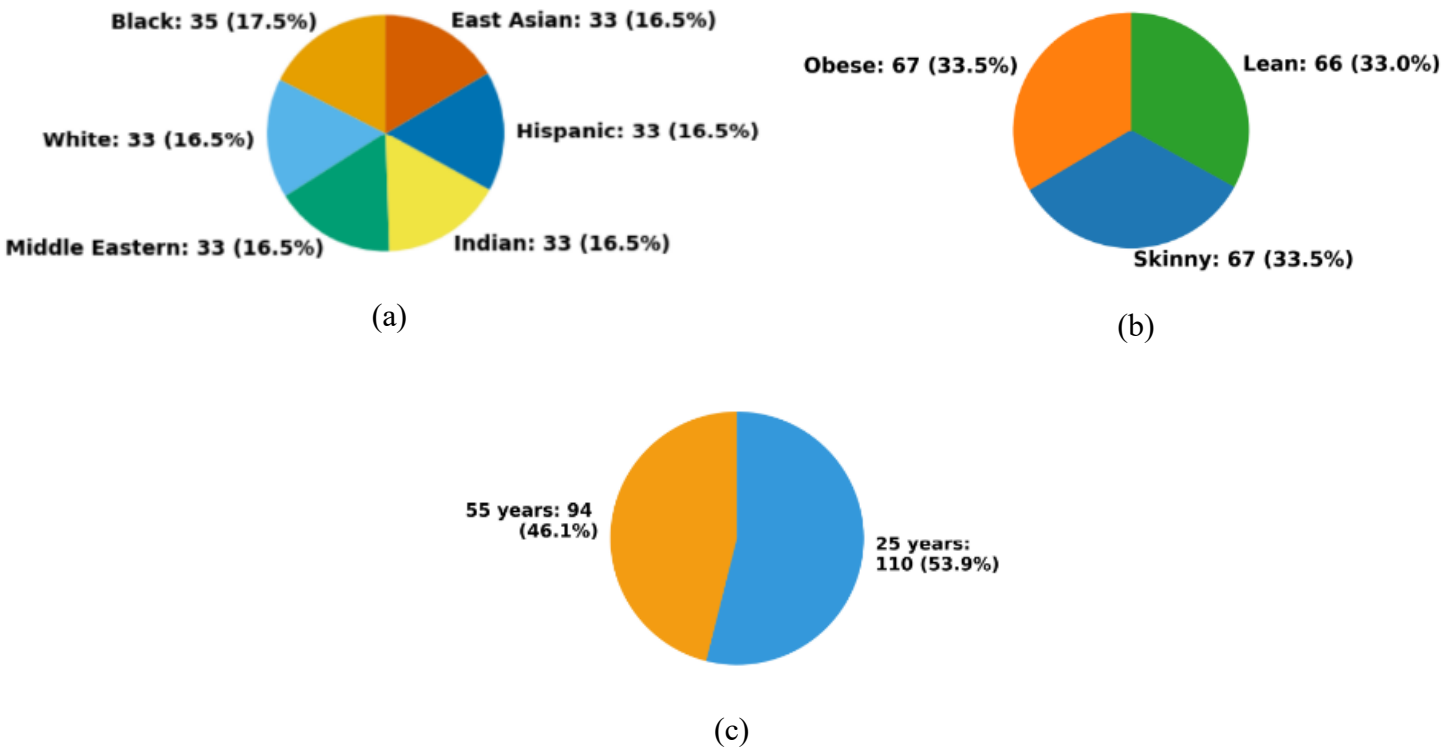


Figure 4. 9 Pie charts showing the distribution of the synthetic images by (a) ethnicity (b) body type and (c) age

Despite the balanced design of prompt features, it was observed that the dataset exhibits a bias toward muscularity. Even in prompts explicitly describing obese body types, subjects frequently displayed some degree of muscular definition (**Fig. 4.7**). This likely reflects an inherent bias in the training datasets used by text-to-image models, which overrepresent physically fit individuals due to the higher availability and visibility of such imagery online.

Figure 4. 10 Output synthetic image generated by Emy through the Llama 3.2 web interface showing muscular definition on a subject with obese build

4.5 Methodology

4.5.1 Model Architecture

This section will provide a recap of the previous NAC estimation machine learning model and evaluation (Target 1: Neck to chest and Target 2: Nipple to Nipple) done in our previous work [105].

The baseline model architecture presented in this study builds upon our prior work on Nipple-Areola Complex (NAC) position estimation in chest reconstruction surgery [105]. In the original study, a dataset of 102 clinical images from 34 male subjects was used. Each subject

contributed three images taken in varying poses. OpenPose [86], an open-source pose estimation framework, was employed to extract 25 body key points. A team of three researchers narrowed these to a subset of six consistently detected anatomical landmarks (Neck, middle hip, left shoulder, right shoulder, left elbow and right elbow), two manually annotated points (left and right nipples), and one calculated point (the center of the inter-nipple distance).

From these landmarks, 34 normalized distances were calculated relative to the inter-shoulder distance (ISD), which served as a reference length. These 34 normalized distances served as inputs for the models to be tested. Two output distances were computed: the vertical distance between the neck and the midpoint between the nipples (denoted as Target 1), and the horizontal distance between the two nipples (denoted as Target 2). These targets formed the outputs of the predictive model.

Six machine learning regression models were tested: Decision Tree Regressor, Random Forest Regressor, CatBoost Regressor, Multilayer Perceptron (MLP), Linear Regression, and Support Vector Regressor (SVR). Each model was evaluated for its ability to predict the NAC coordinates using normalized distance inputs. Linear Regression was identified as the best-performing model, achieving the lowest Mean Absolute Percentage Error (MAPE) given by (4.1) across all evaluations and was thus selected for the applications of this chapter. In this equation, n is the number of data points, Y_i refers to the observed values, and $\hat{Y}_i$ refers to the values predicted by the models.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \cdot 100 \qquad (4.1)$$

To further enhance accuracy, especially in cases where the subject's posture was misaligned (tilted), a tilt correction algorithm was implemented. This algorithm adjusted predictions based on the orientation of the torso using vector normalization techniques involving the neck-mid-hip and shoulder-shoulder vectors. The correction reduced MAPE values from 1.2% (right NAC) and 0.99% (left NAC) to 0.75% and 0.63%, respectively.

The model's performance was validated using a grouped 5-fold cross-validation that grouped each subject's images to prevent overlaps between training and testing sets. The final average inter-nipple distance error (error between the distance between the real left and right nipples and the distance between the predicted left and right nipples) was calculated as 3.61e-14%, and the average angle between the real and predicted nipples was 1.42 degrees.

4.5.2 Model Training and Evaluation

Building on the results from [105] as discussed in Section 4.5.1, in which linear regression was identified as the most accurate model for NAC localization, this section outlines the experimental setup used to evaluate its performance under different training and testing conditions. Table 4.1 below summarizes the types of training and testing datasets involved in the experiments that follow in this section.

Table 4. 1 Dataset composition for each experiment

Experiment	Training Set	Testing Set
1	Real, Synthetic	Real
2	Real, Synthetic	Synthetic
3	Real, Synthetic	External Real-World

4.5.2.1 Experiment 1

To evaluate the impact of augmenting the training set with synthetic data, a controlled experiment was conducted using the existing clinical dataset consisting of 102 real subjects' images, (each of the 34 subjects contributed three different images). The first 7 subjects (21 images) were initially selected as the test set, while the remaining 81 real images were used for training. A baseline was first established by training the model using only these 81 real images and recording the mean absolute percentage error (MAPE) for Target 1 and Target 2.

Following the baseline, synthetic data was incrementally introduced into the training dataset. At each step, 10% of the original training dataset size (8 synthetic images per iteration) was added until the complete set of 200 synthetic images was integrated. After each addition, the model was retrained and the MAPE for both targets were recorded.

The experiment was repeated using a sliding window approach across the entire real dataset. For each repetition, the subsequent group of 7 subjects (21 images) was designated as the test set, and the remaining real images were used as the training set to be augmented by synthetic data. The synthetic data injection into the training dataset was repeated until the synthetic dataset was exhausted, resulting in multiple folds of testing. In each fold, the MAPE for both targets were computed after every synthetic data injection.

Finally, the average MAPE and the standard deviation across all folds were calculated for each target at each level of synthetic data inclusion to visualize the change in performance and variability introduced by the synthetic data augmentation.

4.5.2.2 Experiment 2

Given the limited demographic diversity in the original clinical dataset, a second experiment was conducted to evaluate model performance on a more varied test set. To accomplish this, a subset of the synthetic dataset, which included broader variability across ethnicity, age, and weight class, was used as the testing dataset.

In Experiment 2, the training dataset initially consisted of all 102 real images collected from the original dataset. To evaluate the model's generalizability on a demographically diverse population, a subset of the synthetic dataset was used as the testing set. The first 20 synthetic images were selected as the initial test set. From the remaining 180 synthetic images, synthetic samples were incrementally added to the training set in steps of 10% of the original training set size (i.e., 10 synthetic images per iteration). After each augmentation step, the model was retrained, and the mean absolute percentage error (MAPE) for both target outputs was recorded.

A sliding window approach was employed in the testing phase to iterate over the entire synthetic dataset. In each fold, a new block of 20 synthetic images was selected as the test set, while the remaining synthetic images were incrementally added to the training set following the

same augmentation strategy. At each iteration, the MAPE for both targets was computed, allowing assessment of the model's performance as synthetic data volume increased.

This procedure was repeated until the entire synthetic dataset was exhausted for testing in 20-image segments. At the end of the experiment, the average MAPE and standard deviation across all folds were calculated for each target. These values were used to analyze how model performance and prediction variability changed as a function of synthetic training data volume, particularly when tested against a more diverse and demographically realistic dataset.

4.5.2.3 Experiment 3

In a third experiment, the performance of the model was evaluated against a more challenging test set composed of real-world images collected from public sources. This dataset will be referred to by "external real-world dataset". Given the synthetic nature of the previous testing dataset and the limited demographic and weight class diversity in the original clinical dataset, this external real-world test set was selected by a team of three researchers who unanimously accepted that the images were suitable for the desired objective and that the trained model had not encountered similar images.

Ten images of shirtless male subjects were manually selected from online image repositories. The selected images varied in ethnicity, age, and body type, and featured a broad range of poses and backgrounds. Unlike the controlled posture of the synthetic dataset, where subjects had their arms extended straight at their sides, this new dataset included subjects with arms placed on their hips, hands resting on their bellies, or arms flexed upwards as seen in **Fig. 4.8**. Additionally, the background environment varied from indoor to outdoor settings, further diverging from the clinical examination room background used in synthetic image generation.

To evaluate the impact of synthetic augmentation on model performance, the full set of 102 clinical images was used as the initial training dataset. Synthetic images were then incrementally added to this training set in batches of 10 images, corresponding to 10% of the clinical dataset size per increment. After each addition, the model was retrained and evaluated on the external real-world test set. The mean absolute percentage error (MAPE) was recorded after every training round to quantify performance changes. This procedure enabled a stepwise assessment of how the inclusion of synthetic data influenced the model's ability to generalize to

images that are significantly different from those seen during initial training.

(a) (b)

Figure 4. 14 Examples of selected images for the external real-world dataset showing (a) a subject with arm flexed upwards, (b) a subject with hands placed on his hips.

4.6 Results and Discussion

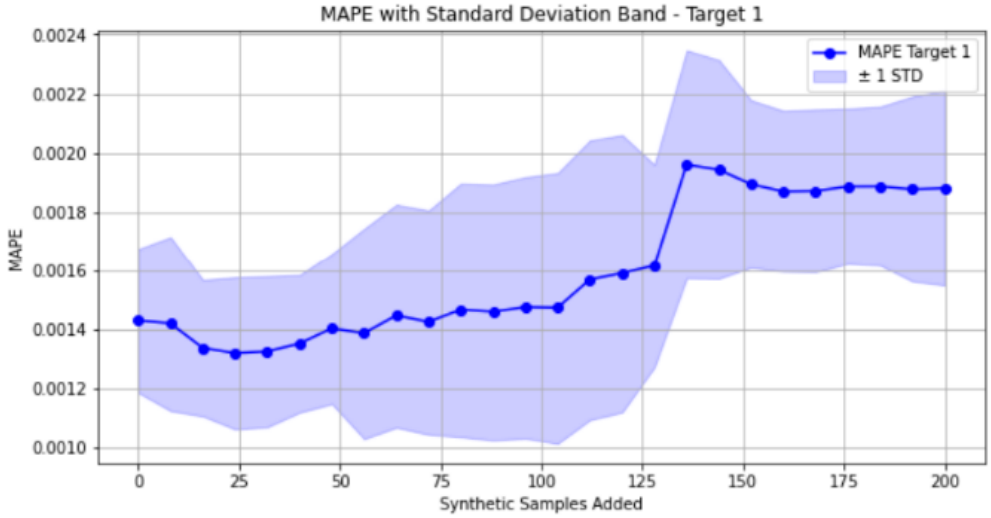
4.6.1 Experiment 1 Results

Experiment 1 evaluated the effect of progressively augmenting the real clinical training dataset with synthetic images generated via Emu. The model was initially trained on 81 real images and tested on 21 held-out real images per fold. Synthetic images were added incrementally in 10% steps, and performance was assessed at each stage using Mean Absolute Percentage Error (MAPE) for two target distances. The goal was to determine whether synthetic augmentation could improve or maintain model performance on real clinical data.

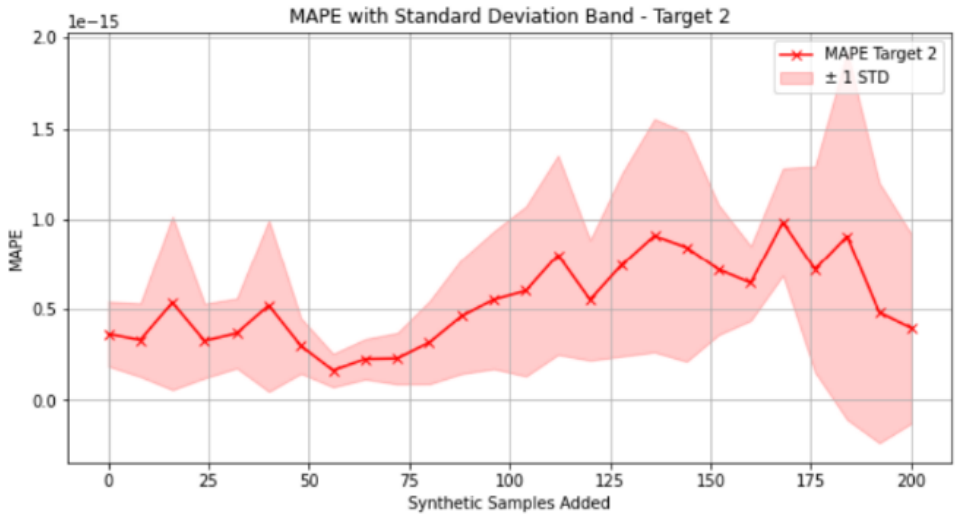
Fig. 4.9 illustrates the MAPE for Target 1 and Target 2 of experiment 1 averaged across five folds, with shaded regions indicating ±1 standard deviation across folds. Target 1 displays complex behavior in regards to the synthetic dataset augmentation. At baseline (no synthetic data),

the model achieved a stable MAPE of approximately 0.00145%. As synthetic samples were gradually introduced, performance improved slightly, with a minimum MAPE of 0.0013% at around 25 added synthetic images (~20% augmentation). However, beyond 100 synthetic images, the MAPE began to increase, and the variance across folds widened significantly, highlighting reduced model robustness. These observations suggest that while synthetic augmentation provides initial benefits, there exists a saturation point, after which further augmentation may impair model generalization.

While the downward trend of the MAPE in the final phase could be an indication of a potential minimum beyond the 200 synthetic images mark, more synthetic images are needed to confirm this behavior.



(a)



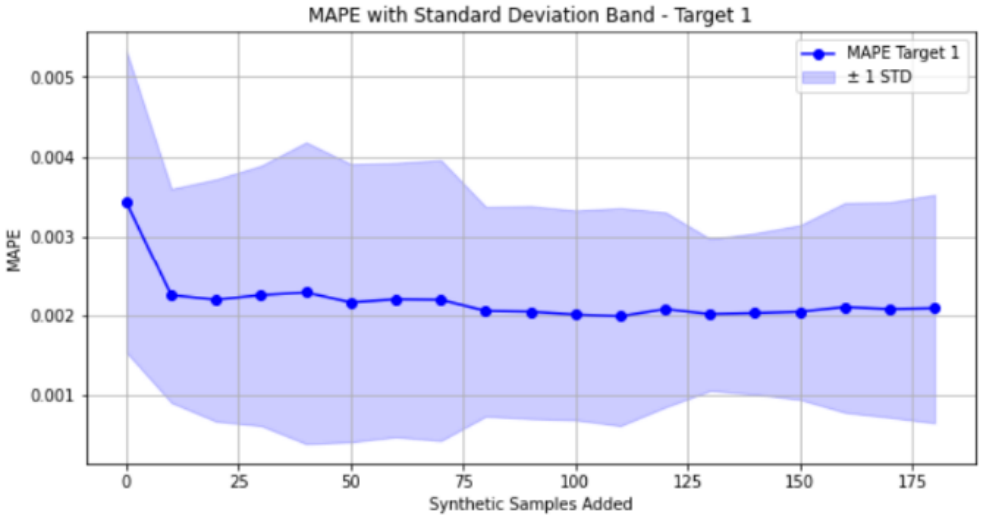
(b)

Figure 4. 18 A Graph showing the MAPE and standard deviation of (a) Target 1 (b) Target 2 across 5 folds for experiment 1.

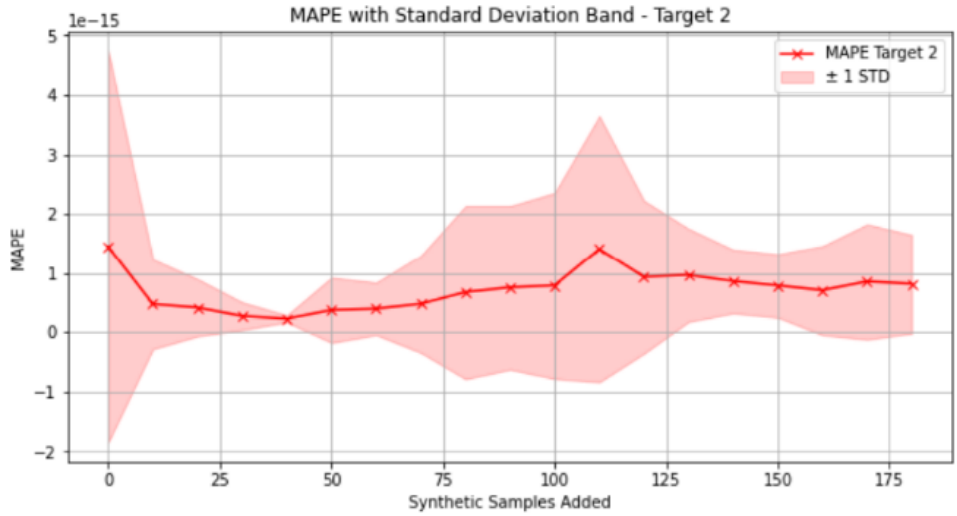
The error is observed to be consistently close to zero for Target 2 across all folds of the experiment. This suggests that the linear regression model has achieved a near-optimal generalization for this output (Target 2) with the initial 81 real images, suggesting that target 2 exhibits a linearly predictable spatial relationship that is less sensitive to individual variability. The standard deviation across all folds for target 2 remains negligible, which suggests that the synthetic augmentation of the dataset has little to no measurable impact on the prediction MAPE of Target 2.

4.6.2 Experiment 2 Results

The results of experiment 2 are detailed in **Fig. 4.10** below.



(a)



(b)

Figure 4. 22 A Graph showing the MAPE and standard deviation of (a) Target 1 (b) Target 2 across 10 folds for experiment 2.

In Experiment 1, the test set consisted of real patient images similar to the training set, and the model achieved low baseline MAPE. In contrast, experiment 2 highlights the model's initial limited generalizability when evaluated on demographically diverse synthetic images, as

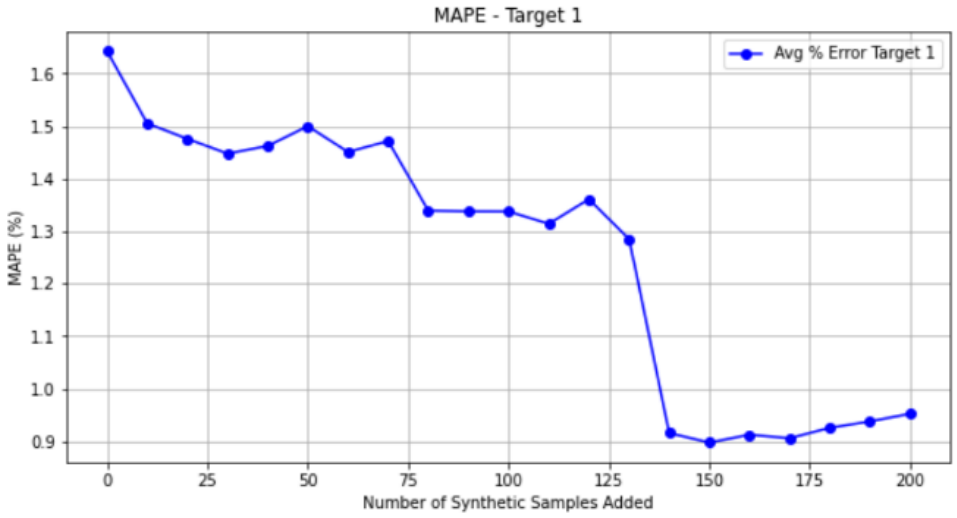
evidenced by the higher initial error of 0.0035%. However, performance improves steadily with synthetic augmentation and stabilizes around 0.002%. The most noticeable improvement occurs in the first 10 synthetic samples added, after which performance gains taper off. This suggests diminishing returns beyond a certain augmentation threshold. The variance (shaded band) is relatively large throughout, especially in early iterations. This reflects model inconsistency when first exposed to demographically diverse test data. However, as synthetic data is introduced into training, the variance narrows slightly, indicating improved generalization and reduced fold-to-fold variability.

When it comes to Target 2, the results in experiment 2 further confirm the ability of the linear regression model to accurately predict Target 2 and its insensitivity to the synthetic data augmentation process. The MAPE of Target 2 is consistent with the results of experiment 1, showing a negligible error across all folds.

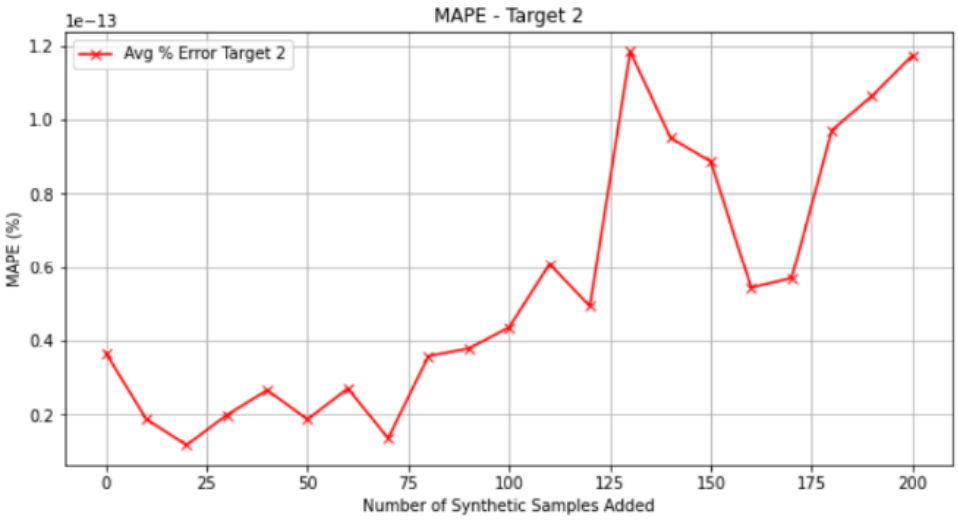
4.6.3 Experiment 3 Results

This experiment assessed the model's ability to generalize to out-of-distribution data by evaluating it on a challenging external real-world dataset collected from public sources. These images featured greater variability in pose, background, body type, and ethnicity compared to the clinical training set. The model was trained on the full set of real clinical images (102 images), and synthetic images were incrementally added to the training set. Performance was measured using MAPE to determine whether synthetic augmentation could improve model robustness in real-world scenarios.

The results of experiment 3 are outlined in **Fig 4.11** below.



(a)



(b)

Figure 4. 26 A Graph showing the MAPE and standard deviation of (a) Target 1 (b) Target 2 for experiment 3.

Fig. 4.11 presents the MAPE for Target 1 in Experiment 3, where the model was evaluated on an external real-world test set composed of images collected from public online sources. Unlike the synthetic or CHEO clinical images used in previous experiments, these images featured highly varied poses, body types, ethnicities, and backgrounds. The baseline model, trained only on CHEO

data, exhibited a relatively high MAPE of 1.63%, highlighting its inability to generalize to diverse, real-world scenarios. However, as synthetic data was incrementally added to the training set, the model's performance steadily improved. A significant drop in MAPE was observed after 130 synthetic samples, reaching a minimum error of 0.91%. This steep improvement indicates that synthetic data augmentation was able to bridge the domain gap between clinical and real-world imagery. The results also suggest that the model begins to generalize and recognize patterns more effectively beyond a certain threshold (~130 synthetic images). The findings of this experiment suggest that for Target 1, synthetic data can enhance generalization to out-of-distribution samples, provided a sufficient volume is included. However, the improvement is non-linear and requires surpassing a certain threshold before benefits materialize.

The MAPE of target 2 remains negligible across this experiment as well, indicating that the model has already generalized well to target 2 under varied testing conditions. The small fluctuations observed in MAPE of target 2 are not significant, as the overall error is in the magnitude of e-13%.

In interpreting the error magnitudes obtained, it is important to note that several results fell within the order of e-15% to e-13%. Errors in the e-15 to e-13 range can be reasonably interpreted as zero for all practical purposes. In other words, these outcomes indicate that the predicted and target values are functionally identical within the resolution of the image data and the normalization scheme employed. This interpretation is consistent with the pixel-level granularity of the dataset, where sub-pixel differences are not observable in practice.

4.7 Conclusion

This study addressed a key limitation, the scarcity and demographic narrowness of real clinical imaging datasets, in machine learning-driven NAC localization for chest reconstruction surgery. By introducing a synthetic dataset generated via Llama 3.2, through customized prompt engineering, we demonstrated the potential of large language models to produce realistic, demographically varied training images. A series of three experiments were conducted to evaluate

the impact of synthetic data augmentation on model performance under increasingly diverse and challenging testing conditions.

In Experiment 1, the model was trained on real clinical images and tested on a subset of real patient images. The training set consisted of 81 real images, and the testing set consisted of 21 real images per fold. Synthetic images were incrementally added to the training set to assess their impact. Results revealed that moderate augmentation of the original dataset with synthetic images could slightly improve or maintain model performance. However, performance gains plateaued or slightly declined when synthetic data exceeded a certain threshold, suggesting that balance is essential to avoid overfitting or introducing noise from overly synthetic content.

In Experiment 2, the model was again trained on the full set of 102 real clinical images but tested on synthetic images that were purposefully designed to represent a more demographically diverse population in terms of age, body type, and ethnicity. As synthetic data was gradually added to the training set, results indicated a significant improvement in model generalization. This confirmed that synthetic augmentation helped bridge the demographic gap in the training data and enabled the model to perform better on a more varied input distribution.

In Experiment 3, the model was trained on the real clinical images and evaluated on a challenging test set composed of real-world images collected from online sources. These test images differed substantially from the training data in terms of pose, background, and subject diversity. Initially, the model exhibited a high MAPE due to the out-of-distribution nature of the test set. However, as synthetic images were incrementally introduced into the training set, model performance significantly improved. A notable drop in error occurred after approximately 130 synthetic images were added, reducing the MAPE by nearly 50%. This demonstrated that synthetic data can significantly enhance model robustness and domain generalization, especially under challenging and heterogeneous testing conditions.

These findings show the value of synthetic image generation using LLMs in overcoming real data limitations. Carefully designed and demographically diverse synthetic datasets can improve the accuracy and generalizability of machine learning models in clinical applications such as NAC localization.

Chapter 5. Summary and Conclusions

This thesis set out to explore the development of machine learning-based decision support systems for plastic surgery, with a particular focus on automating and improving the precision and reproducibility of nipple-areola complex (NAC) localization in reconstructive chest procedures. The work was guided by two primary objectives:

- To develop a machine learning-based decision support approach to assist in the precise and consistent localization of the NAC in reconstructive chest surgery.
- To evaluate the effect of augmenting a real clinical dataset with synthetic images on the performance and generalizability of the decision support system.

To address the first objective, Chapter 3 presented an NAC localization pipeline that utilized OpenPose for upper-body landmark extraction and applied six regression models for prediction. The models included Decision Tree Regressor, Random Forest Regressor, CatBoost Regressor, Multilayer Perceptron Regressor, Support Vector Regressor, and Linear Regressor. The linear regression model achieved the best performance, with a mean absolute percentage error (MAPE) of 0.06% for normalized distances and an average NAC positional error of 0.69%. A tilt correction algorithm was implemented to address misalignment in body posture, reducing the MAPE from 1.20% to 0.75% for the right NAC and from 0.99% to 0.63% for the left NAC in affected cases. These results demonstrated that the selected models can predict NAC position with high accuracy in normalized image space, offering a significant improvement over subjective manual estimation. This work builds on the work of Ghodratigohar et al. [21] by replacing manually annotated landmarks with a fully automated pipeline that uses pose estimation to extract anatomical keypoints. Moreover, it uses a different regression model (linear regression vs. the proposed artificial neural network) to achieve a higher localization accuracy.

Chapter 4 addressed the second objective by evaluating the impact of synthetic data augmentation on model performance and generalizability. A total of 200 synthetic images were generated using a controlled prompt engineering pipeline and incrementally added to the real dataset across three experiments. In Experiment 1, the model was trained on 81 real images and tested on 21 real images in the first fold. Synthetic images were then gradually introduced to the

training set to assess their impact on performance. Results showed that moderate augmentation slightly improved or maintained accuracy, but performance declined when synthetic data exceeded a certain threshold, indicating the importance of balance. In Experiment 2, the model was initially trained on all 102 real clinical images and evaluated on a demographically diverse set of synthetic test images, which included variability in age, body type, and ethnicity. To assess the effect of synthetic augmentation on model generalization, synthetic images were incrementally added to the training set in 10% steps relative to the original training size. After each addition, the model was retrained and re-evaluated on a rotating subset of 20 synthetic test images. This process was repeated across multiple folds until the full set of 200 synthetic images was utilized. Results showed a consistent decrease in prediction error as more synthetic data was incorporated, indicating that the inclusion of demographically varied synthetic images improved the model's ability to generalize beyond the narrower distribution of the original clinical dataset. In Experiment 3, the model was evaluated on a challenging external real-world dataset composed of images collected from public sources. These images featured subject characteristics, such as pose, background, and body types, that were not present in the original clinical dataset, making them unfamiliar to the model during training. To assess whether synthetic augmentation could improve performance under such conditions, the model was trained on all 102 real clinical images, and synthetic images were incrementally added to the training set in batches of 10% of the size of the training set (10 images/batch). After each increment, the model was retrained and tested on the same external dataset. Initially, the model exhibited relatively high prediction error due to the unfamiliarity of the test images. However, after approximately 130 synthetic images were introduced into the training set, the MAPE dropped by 44%. These results confirm that when real datasets lack diversity, strategically increasing synthetic data volume can enhance model robustness and generalization across broader real-world scenarios.

Collectively, the findings of this thesis support the following five contributions:

- A fully automated machine learning-based NAC prediction model was developed, leveraging pose estimation and normalized anatomical distances.
- A tilt correction algorithm was implemented to improve model robustness in cases of subject misalignment.

- An automatic prompt generation workflow using ChatGPT-4 was designed to support diverse synthetic image creation.

- Synthetic data generated via the Emu diffusion model was successfully integrated to expand dataset diversity and scale.

- The model was experimentally evaluated across three scenarios, real clinical images, synthetic images with demographic variation, and external public datasets, which allowed for the demonstration of improvements in generalization.

Finally, this thesis has demonstrated that machine learning models can achieve high-precision NAC localization (≤0.75% MAPE) using structured anatomical features, and that synthetic data can improve generalization when combined with real datasets. These results contribute toward automated, scalable, and reproducible AI systems for plastic surgery decision support, and lay the foundation for future clinical integration and model refinement.

In conclusion, this thesis presented a machine learning–based framework for automated nipple–areola complex (NAC) localization from upper body keypoints detected by OpenPose. While it may appear intuitive to compare this method against prior geometric approaches from the literature, such a comparison is not practical. Those methods depend on landmarks such as the xiphoid process, anterior superior iliac spine, or anterior axillary folds, which OpenPose does not detect. Incorporating these landmarks through manual annotation would undermine the very objective of automation, making such comparisons misaligned with the scope and goals of this study.

A limitation of this study lies in the composition of the clinical dataset. The sample predominantly included lean individuals, which does not adequately represent the full spectrum of body types encountered in the general population. As a result, the model may not fully capture the variability in anatomical landmarks and chest morphology that can occur in individuals with higher body mass index (BMI) or different body compositions. This imbalance could introduce bias into the predictive framework, limiting its generalizability and accuracy when applied to more diverse patient groups. Addressing this limitation will require the inclusion of a larger and more heterogeneous real dataset that encompasses a broader range of body types, thereby improving the

robustness and clinical applicability of the proposed method.

Another limitation lies in the bias introduced through the use of synthetic data augmentation. The synthetic dataset was generated using models that tend to emphasize muscular body forms, which could reduce the diversity of body types represented and affect generalizability. Future datasets should incorporate a broader range of body compositions to mitigate this bias and improve robustness across populations.

Looking ahead, this system can be envisioned as a component within a larger surgical planning system. One immediate direction is its integration into a mobile application where the images are taken by the surgeon through the application and the predicted measurements are provided directly to the surgeon, who can then manually translate them onto the patient's body during preoperative planning. A more advanced pathway is the incorporation of this predictive framework into an augmented reality (AR) environment. In such a system, the surgeon could visualize the projected NAC locations directly on the patient's chest in real time, providing an intuitive and precise guide to support surgical decision-making. These directions highlight the translational potential of the current work and its role in shaping more automated, accessible, and clinically useful tools in reconstructive surgery. Initial feedback from ***Dr. Kevin Cheung*** confirmed the potential usefulness of this work in surgical contexts, supporting its translational relevance.

References

[1] J. Kanevsky, J. Corban, R. Gaster, A. Kanevsky, S. Lin, and M. Gilardino, "Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation," *Plast Reconstr Surg*, vol. 137, no. 5, pp. 890e–897e, May 2016, doi: 10.1097/PRS.0000000000002088.

[2] N. Murali[1] and N. Sivakumaran[2], "Artificial Intelligence in Healthcare-A Review," *International Journal of Modern Computation, Information and Communication Technology*, vol. 1, no. 6, pp. 2581–5954, 2018, doi: 10.13140/RG.2.2.27265.92003.

[3] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat Med*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[4] A. S. Coyner *et al.*, "Synthetic Medical Images for Robust, Privacy-Preserving Training of Artificial Intelligence," *Ophthalmology Science*, vol. 2, no. 2, p. 100126, Jun. 2022, doi: 10.1016/j.xops.2022.100126.

[5] E.-K. Yeong, T.-C. Hsiao, H. K. Chiang, and C.-W. Lin, "Prediction of burn healing time using artificial neural networks and reflectance spectrometer," *Burns*, vol. 31, no. 4, pp. 415–420, Jun. 2005, doi: 10.1016/j.burns.2004.12.003.

[6] B. Stephanian *et al.*, "Role of Artificial Intelligence and Machine Learning in Facial Aesthetic Surgery: A Systematic Review," *Facial Plast Surg Aesthet Med*, vol. 26, no. 6, pp. 679–705, Nov. 2024, doi: 10.1089/fpsam.2024.0204.

[7] W. A. Cohen, R. S. Maisner, N. Hazim, H. S. Ayyala, and J. D. Keith, "Crowdsourcing the Ideal Nipple-Areolar Complex Position for Chest Masculinization Surgery," *Plast Reconstr Surg Glob Open*, vol. 8, no. 8, p. e3070, Aug. 2020, doi: 10.1097/GOX.0000000000003070.

[8] T. R. Mett *et al.*, "Optimal Positioning of the Nipple–Areola Complex in Men Using the Mohrenheim-Estimated-Tangential-Tracking-Line (METT-Line): An Intuitive Approach," *Aesthetic Plast Surg*, vol. 41, no. 6, pp. 1295–1302, Dec. 2017, doi: 10.1007/s00266-017-0941-9.

[9] H. S. Ayyala, T. J. Mukherjee, T.-M. Le, W. A. Cohen, M. Luthringer, and J. D. Keith, “A Three-Step Technique for Optimal Nipple Position in Transgender Chest Masculinization,” *Aesthet Surg J*, vol. 40, no. 11, pp. NP619–NP625, Oct. 2020, doi: 10.1093/asj/sjaa150.

[10] M. Kääriäinen, K. Salonen, M. Helminen, and U. Karhunen-Enckell, “Chest-wall contouring surgery in female-to-male transgender patients: A one-center retrospective analysis of applied surgical techniques and results,” *Scandinavian Journal of Surgery*, vol. 106, no. 1, pp. 74–79, May 2017, doi: 10.1177/1457496916645964.

[11] G. M. Beer, S. Budi, B. Seifert, W. Morgenthaler, M. Infanger, and V. E. Meyer, “Configuration and Localization of the Nipple-Areola Complex in Men,” *Plast Reconstr Surg*, vol. 108, no. 7, pp. 1947–1952, Dec. 2001, doi: 10.1097/00006534-200112000-00015.

[12] B. S. Atiyeh, S. A. Dibo, and A. H. El Chafic, “Vertical and Horizontal Coordinates of the Nipple-Areola Complex Position in Males,” *Ann Plast Surg*, vol. 63, no. 5, pp. 499–502, Nov. 2009, doi: 10.1097/SAP.0b013e3181953854.

[13] C. A. Agarwal, M. F. Scheefer, L. N. Wright, N. K. Walzer, and A. Rivera, “Quality of life improvement after chest wall masculinization in female-to-male transgender patients: A prospective study using the BREAST-Q and Body Uneasiness Test,” *Journal of Plastic, Reconstructive and Aesthetic Surgery*, vol. 71, no. 5, pp. 651–657, May 2018, doi: 10.1016/j.bjps.2018.01.003.

[14] M. J. Willemink *et al.*, “Preparing Medical Imaging Data for Machine Learning,” *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020, doi: 10.1148/radiol.2020192224.

[15] L. R. Koetzier *et al.*, “Generating Synthetic Data for Medical Imaging,” May 2024. doi: 10.1148/radiol.232471.

[16] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” *Sci Rep*, vol. 9, no. 1, p. 12495, Aug. 2019, doi: 10.1038/s41598-019-48995-4.

[17] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles, “Artificial Intelligence in

Surgery: Promises and Perils," *Ann Surg*, vol. 268, no. 1, pp. 70–76, Jul. 2018, doi: 10.1097/SLA.0000000000002693.

[18] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat Med*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[19] K. J. Geras, R. M. Mann, and L. Moy, "Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives," *Radiology*, vol. 293, no. 2, pp. 246–259, Nov. 2019, doi: 10.1148/radiol.2019182627.

[20] S. E. Braun, L. M. Sinik, A. M. Meyer, K. E. Larson, and J. A. Butterworth, "Predicting Complications in Breast Reconstruction," *Ann Plast Surg*, vol. 91, no. 2, pp. 282–286, Aug. 2023, doi: 10.1097/SAP.0000000000003621.

[21] M. Ghodratigohar, K. Cheung, N. Baddour, and H. Al Osman, "A Machine Learning Method for Nipple-Areola Complex Localization for Chest Masculinization Surgery," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 479–485. doi: 10.1007/978-3-030-18305-9_48.

[22] K. Magudia, C. P. Bridge, K. P. Andriole, and M. H. Rosenthal, "The Trials and Tribulations of Assembling Large Medical Imaging Datasets for Machine Learning Applications," *J Digit Imaging*, vol. 34, no. 6, pp. 1424–1429, May 2021, doi: 10.1007/s10278-021-00505-7.

[23] J. Cevik *et al.*, "Assessment of the bias of artificial intelligence generated images and large language models on their depiction of a surgeon," *ANZ J Surg*, vol. 94, no. 3, pp. 287–294, May 2024, doi: 10.1111/ans.18792.

[24] A. K. Moorefield *et al.*, "Analysis of Nipple-Areola Complex Localization Using Male Cadavers: Considerations for Gender-Affirming Surgery," *Aesthet Surg J Open Forum*, vol. 3, no. 4, Dec. 2021, doi: 10.1093/asjof/ojab032.

[25] C. A. Agarwal, V. T. Wall, S. T. Mehta, D. P. Donato, and N. K. Walzer, "Creation of an

Aesthetic Male Nipple Areolar Complex in Female-to-Male Transgender Chest Reconstruction.," *Aesthetic Plast Surg*, vol. 41, no. 6, pp. 1305–1310, Dec. 2017, doi: 10.1007/s00266-017-0935-7.

[26] M. R. Bykowski, P. I. Emelife, N. N. Emelife, W. Chen, N. J. Panetta, and C. de la Cruz, "Nipple-areola complex reconstruction improves psychosocial and sexual well-being in women treated for breast cancer.," *J Plast Reconstr Aesthet Surg*, vol. 70, no. 2, pp. 209–214, Feb. 2017, doi: 10.1016/j.bjps.2016.10.009.

[27] F. W. Timmermans *et al.*, "The ideal location of the male nipple-areolar complex: A pinpointing algorithm," *Int J Transgend Health*, vol. 22, no. 4, pp. 403–411, Oct. 2021, doi: 10.1080/26895269.2021.1884926.

[28] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digit Med*, vol. 3, no. 1, p. 17, Feb. 2020, doi: 10.1038/s41746-020-0221-y.

[29] A. X. Garg *et al.*, "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes," *JAMA*, vol. 293, no. 10, p. 1223, Mar. 2005, doi: 10.1001/jama.293.10.1223.

[30] T. J. Loftus *et al.*, "Artificial Intelligence–enabled Decision Support in Surgery," *Ann Surg*, vol. 278, no. 1, pp. 51–58, Jul. 2023, doi: 10.1097/SLA.0000000000005853.

[31] A. J. Navarrete-Welton and D. A. Hashimoto, "Current applications of artificial intelligence for intraoperative decision support in surgery," *Front Med*, vol. 14, no. 4, pp. 369–381, Aug. 2020, doi: 10.1007/s11684-020-0784-7.

[32] L. F. A. M. Gomes, L. A. D. Rangel, and P. P. Fernandes, "Decision aiding in plastic surgery: a multicriteria analysis," *Pesquisa Operacional*, vol. 32, no. 2, pp. 371–387, Jun. 2012, doi: 10.1590/S0101-74382012005000011.

[33] E. D. Sears and K. C. Chung, "Decision Analysis in Plastic Surgery: A Primer," *Plast Reconstr Surg*, vol. 126, no. 4, pp. 1373–1380, Oct. 2010, doi: 10.1097/PRS.0b013e3181ead10a.

[34] C. J. Pannucci *et al.*, “Validation of the Caprini Risk Assessment Model in Plastic and Reconstructive Surgery Patients,” *J Am Coll Surg*, vol. 212, no. 1, pp. 105–112, Jan. 2011, doi: 10.1016/j.jamcollsurg.2010.08.018.

[35] F. Ziani *et al.*, “Development of a Digital Application Program Based on an Institutional Algorithm Sustaining the Decisional Process for Breast Reconstruction in Patients with Large and Ptotic Breasts: A Pilot Study,” *Cancers (Basel)*, vol. 17, no. 11, p. 1807, May 2025, doi: 10.3390/cancers17111807.

[36] J. E. Ospina-Gómez, J. M. Molano-Diaz, M. C. Rojas-Gómez, M. G. Latorre-Arévalo, and M. Sanchez-Vargas, “Advances and applications of artificial intelligence in breast reconstruction surgery: a systematic review,” *Eur J Plast Surg*, vol. 48, no. 1, p. 12, Jan. 2025, doi: 10.1007/s00238-025-02271-z.

[37] M. Mansoor and A. F. Ibrahim, “The Transformative Role of Artificial Intelligence in Plastic and Reconstructive Surgery: Challenges and Opportunities,” *J Clin Med*, vol. 14, no. 8, p. 2698, Apr. 2025, doi: 10.3390/jcm14082698.

[38] K. Y. Bilimoria *et al.*, “Development and Evaluation of the Universal ACS NSQIP Surgical Risk Calculator: A Decision Aid and Informed Consent Tool for Patients and Surgeons,” *J Am Coll Surg*, vol. 217, no. 5, pp. 833-842e3, Nov. 2013, doi: 10.1016/j.jamcollsurg.2013.07.385.

[39] S. A. M. Nashef, F. Roques, P. Michel, E. Gauducheau, S. Lemeshow, and R. Salamon, “European system for cardiac operative risk evaluation (EuroSCORE),” *European Journal of Cardio-Thoracic Surgery*, vol. 16, no. 1, pp. 9–13, Jul. 1999, doi: 10.1016/S1010-7940(99)00134-7.

[40] A. C. O’Neill, D. Yang, M. Roy, S. Sebastiampillai, S. O. P. Hofer, and W. Xu, “Correction to: Development and Evaluation of a Machine Learning Prediction Model for Flap Failure in Microvascular Breast Reconstruction,” *Ann Surg Oncol*, vol. 29, no. 6, pp. 3867–3867, Jun. 2022, doi: 10.1245/s10434-021-10979-y.

[41] T. F. Byrd IV and C. J. Tignanelli, “Artificial intelligence in surgery—a narrative review,”

J Med Artif Intell, vol. 7, pp. 29–29, Sep. 2024, doi: 10.21037/jmai-24-111.

[42] B. Shickel *et al.*, "Dynamic predictions of postoperative complications from explainable, uncertainty-aware, and multi-task deep neural networks," *Sci Rep*, vol. 13, no. 1, p. 1224, Jan. 2023, doi: 10.1038/s41598-023-27418-5.

[43] C. J. Atkinson *et al.*, "Artificial Intelligence Language Model Performance for Rapid Intraoperative Queries in Plastic Surgery: ChatGPT and the Deep Inferior Epigastric Perforator Flap," *J Clin Med*, vol. 13, no. 3, p. 900, Feb. 2024, doi: 10.3390/jcm13030900.

[44] S. Saglam, V. Uludag, Z. O. Karaduman, M. Arıcan, M. O. Yücel, and R. E. Dalaslan, "Comparative evaluation of artificial intelligence models GPT-4 and GPT-3.5 in clinical decision-making in sports surgery and physiotherapy: a cross-sectional study," *BMC Med Inform Decis Mak*, vol. 25, no. 1, p. 163, Apr. 2025, doi: 10.1186/s12911-025-02996-8.

[45] A. O. Wamalwa, T. Stasch, F. W. Nangole, and S. O. Khainga, "Surgical anatomy of reduction mammaplasty: a historical perspective and current concepts.," *S Afr J Surg*, vol. 55, no. 1, pp. 22–28, Mar. 2017.

[46] J. Penn, "Breast reduction," *Br J Plast Surg*, vol. 7, pp. 357–371, 1954, doi: 10.1016/S0007-1226(54)80046-4.

[47] H. A. Khan and A. Bayat, "A Geometric Method for Nipple Localization," *Canadian Journal of Plastic Surgery*, vol. 16, no. 1, pp. 45–47, Mar. 2008, doi: 10.1177/229255030801600101.

[48] D. Lazarus, "A New Template-Goniometer for Marking the Wise Keyhole Pattern of Reduction Mammaplasty," *Plast Reconstr Surg*, vol. 101, no. 1, pp. 171–173, Jan. 1998, doi: 10.1097/00006534-199801000-00029.

[49] F. R. Zuluaga, D. P. Ríos, A. M. Moreno, and N. M. Jiménez, "Tripedicle Breast Reduction Technique: Superomedial, Central and Inferior. A Modified McKissock Technique," *Aesthetic Plast Surg*, Jun. 2025, doi: 10.1007/s00266-025-04874-0.

[50] L. Alaniz *et al.*, "Developing Practice Guidelines on Chest Masculinization: Designing

Male Neo-nipple–Areolar Complex," *Plast Reconstr Surg Glob Open*, vol. 12, no. 12, p. e6376, Dec. 2024, doi: 10.1097/GOX.0000000000006376.

[51] S. Tanini and G. Lo Russo, "Shape, Position and Dimension of the Nipple Areola Complex in the Ideal Male Chest: A Quick and Simple Operating Room Technique," *Aesthetic Plast Surg*, vol. 42, no. 4, pp. 951–957, Aug. 2018, doi: 10.1007/s00266-018-1131-0.

[52] M. Maas, A. C. Howell, D. J. Gould, and E. C. Ray, "The Ideal Male Nipple-Areola Complex," *Ann Plast Surg*, vol. 84, no. 3, pp. 334–340, Mar. 2020, doi: 10.1097/SAP.0000000000002018.

[53] W. R. Lindsay, "Creation of a male chest in female transsexuals.," *Ann Plast Surg*, vol. 3, no. 1, pp. 39–46, Jul. 1979.

[54] S. Tanini and G. Lo Russo, "Shape, Position and Dimension of the Nipple Areola Complex in the Ideal Male Chest: A Quick and Simple Operating Room Technique," *Aesthetic Plast Surg*, vol. 42, no. 4, pp. 951–957, Aug. 2018, doi: 10.1007/s00266-018-1131-0.

[55] J. J. Hage and P. J. M. van Kesteren, "Chest-Wall Contouring in Female-to-Male Transsexuals," *Plast Reconstr Surg*, vol. 96, no. 2, pp. 386–391, Aug. 1995, doi: 10.1097/00006534-199508000-00019.

[56] O. Shulman, E. Badani, Y. Wolf, and D. J. Hauben, "Appropriate Location of the Nipple-Areola Complex in Males," *Plast Reconstr Surg*, vol. 108, no. 2, pp. 348–351, Aug. 2001, doi: 10.1097/00006534-200108000-00010.

[57] N. Kenig, J. Monton Echeverria, and L. De la Ossa, "Identification of Key Breast Features Using a Neural Network: Applications of Machine Learning in the Clinical Setting of Plastic Surgery," *Plast Reconstr Surg*, vol. 153, no. 2, pp. 273e–280e, Feb. 2024, doi: 10.1097/PRS.0000000000010603.

[58] J. K. Park, S. Baek, C. Y. Heo, J. H. Jeong, and Y. Myung, "A Novel, Deep Learning-Based, Automatic Photometric Analysis Software for Breast Aesthetic Scoring," *Arch Plast Surg*, vol. 51, no. 01, pp. 030–035, Jan. 2024, doi: 10.1055/a-2190-5781.

[59] C. Guo *et al.*, “A fully automatic framework for evaluating cosmetic results of breast conserving therapy,” *Machine Learning with Applications*, vol. 10, p. 100430, Dec. 2022, doi: 10.1016/j.mlwa.2022.100430.

[60] M. J. Cardoso *et al.*, “Turning subjective into objective: The BCCT.core software for evaluation of cosmetic results in breast cancer conservative treatment,” *The Breast*, vol. 16, no. 5, pp. 456–461, Oct. 2007, doi: 10.1016/j.breast.2007.05.002.

[61] M. Ghodratigohar, K. Cheung, N. Baddour, and H. Al Osman, “A Machine Learning Method for Nipple-Areola Complex Localization for Chest Masculinization Surgery,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2019, pp. 479–485. doi: 10.1007/978-3-030-18305-9_48.

[62] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch, “Plastic Surgery: An Obstacle for Deep Face Recognition?,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2020, pp. 3510–3517. doi: 10.1109/CVPRW50498.2020.00411.

[63] K. Keefer *et al.*, “The State of Data Sharing in Plastic Surgery: An Analysis of Journal Practices and Author Adherence,” *Plast Reconstr Surg Glob Open*, vol. 13, no. 5, p. e6761, May 2025, doi: 10.1097/GOX.0000000000006761.

[64] A. Singh, I. Seth, B. Lim, R. Cuomo, and W. M. Rozen, “Ethical issues of artificial intelligence in plastic surgery: a narrative review,” *Plast Aesthet Res*, Nov. 2024, doi: 10.20517/2347-9264.2024.108.

[65] E. Solomon *et al.*, “FastMRI Breast: A Publicly Available Radial k-Space Dataset of Breast Dynamic Contrast-enhanced MRI,” *Radiol Artif Intell*, vol. 7, no. 1, Jan. 2025, doi: 10.1148/ryai.240345.

[66] A. Cheong, “Computational modeling of breast shape using spherical harmonics,” PhD Dissertation, University of Houston, Houston, 2018.

[67] D. L. Spoer *et al.*, “A Systematic Review of Artificial Intelligence Applications in Plastic

Surgery: Looking to the Future," *Plast Reconstr Surg Glob Open*, vol. 10, no. 12, p. e4608, Dec. 2022, doi: 10.1097/GOX.0000000000004608.

[68] D. Patmon, H. Sandhu, J. Girotto, and R. Ford, "Legal Ramifications of Publishing Patient Photographs: A Review of Legal Cases," *Plast Reconstr Surg Glob Open*, vol. 11, no. 8, p. e5162, Aug. 2023, doi: 10.1097/GOX.0000000000005162.

[69] K. Nguyen, C. Fookes, R. Jillela, S. Sridharan, and A. Ross, "Long range iris recognition: A survey," *Pattern Recognit*, vol. 72, pp. 123–143, Dec. 2017, doi: 10.1016/j.patcog.2017.05.021.

[70] T. Schlett, C. Rathgeb, J. Tapia, and C. Busch, "Double Trouble? Impact and Detection of Duplicates in Face Image Datasets," in *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods*, SCITEPRESS - Science and Technology Publications, 2024, pp. 801–808. doi: 10.5220/0012422500003654.

[71] T. Van Duong, V. P. T. Vy, and T. N. K. Hung, "Artificial Intelligence in Plastic Surgery: Advancements, Applications, and Future," *Cosmetics*, vol. 11, no. 4, p. 109, Jun. 2024, doi: 10.3390/cosmetics11040109.

[72] A. Althnian *et al.*, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, vol. 11, no. 2, p. 796, Jan. 2021, doi: 10.3390/app11020796.

[73] J. Gütter, A. Kruspe, X. X. Zhu, and J. Niebling, "Impact of Training Set Size on the Ability of Deep Neural Networks to Deal with Omission Noise," *Frontiers in Remote Sensing*, vol. 3, Jul. 2022, doi: 10.3389/frsen.2022.932431.

[74] J. Buolamwini, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *," 2018.

[75] B. B. Ozmen, E. Pinsky, and G. S. Schwarz, "Future of outcomes research in plastic surgery: Artificial intelligence generated synthetic data and predictive models," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 94, pp. 38–39, Jul. 2024, doi: 10.1016/j.bjps.2024.05.014.

[76] A. Zargaran, S. Sousi, S. P. Glynou, H. Mortada, D. Zargaran, and A. Mosahebi, “A systematic review of generative adversarial networks (GANs) in plastic surgery,” *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 95, pp. 377–385, Aug. 2024, doi: 10.1016/j.bjps.2024.04.007.

[77] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets.”

[78] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN.”

[79] P. K. Chandaliya and N. Nain, “PlasticGAN: Holistic generative adversarial network on face plastic and aesthetic surgery,” *Multimed Tools Appl*, vol. 81, no. 22, pp. 32139–32160, Sep. 2022, doi: 10.1007/s11042-022-12865-5.

[80] J. P. Flynn, E. Cha, T. J. Flynn, J. D. Kriet, and C. D. Humphrey, “Machine Learning Generated Synthetic Faces for Use in Facial Aesthetic Research,” *Facial Plast Surg Aesthet Med*, vol. 23, no. 5, pp. 339–343, Oct. 2021, doi: 10.1089/fpsam.2020.0328.

[81] A. Hayajneh, E. Serpedin, M. Shaqfeh, G. Glass, and M. A. Stotland, “Adapting a style based generative adversarial network to create images depicting cleft lip deformity,” *Sci Rep*, vol. 15, no. 1, p. 3614, Jan. 2025, doi: 10.1038/s41598-025-86588-6.

[82] C. Chartier, A. Watt, O. Lin, A. Chandawarkar, J. Lee, and E. Hall-Findlay, “BreastGAN: Artificial Intelligence-Enabled Breast Augmentation Simulation,” *Aesthet Surg J Open Forum*, vol. 4, Jan. 2022, doi: 10.1093/asjof/ojab052.

[83] B. Lim *et al.*, “Using Generative Artificial Intelligence Tools in Cosmetic Surgery: A Study on Rhinoplasty, Facelifts, and Blepharoplasty Procedures,” *J Clin Med*, vol. 12, no. 20, p. 6524, Oct. 2023, doi: 10.3390/jcm12206524.

[84] T. C. van de Grift, E. Elaut, S. C. Cerwenka, P. T. Cohen-Kettenis, and B. P. C. Kreukels, “Surgical Satisfaction, Quality of Life, and Their Association After Gender-Affirming Surgery: A Follow-up Study,” *J Sex Marital Ther*, vol. 44, no. 2, pp. 138–148, Feb. 2018, doi: 10.1080/0092623X.2017.1326190.

[85] V. Gupta and V. K. Sharma, “Skin typing: Fitzpatrick grading and others,” *Clin Dermatol*,

vol. 37, no. 5, pp. 430–436, Sep. 2019, doi: 10.1016/j.clindermatol.2019.07.010.

[86] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” May 2016.

[87] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Routledge, 2017. doi: 10.1201/9781315139470.

[88] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[89] W.-Y. Loh, “REGRESSION TREES WITH UNBIASED VARIABLE SELECTION AND INTERACTION DETECTION,” 2002. [Online]. Available: http://lib.stat.cmu

[90] J. R. Quinlan, “Bagging, Boosting, and C4.5.”

[91] L. Breiman, “Random forests.,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[92] D. R. Cutler *et al.*, “RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY,” *Ecology*, vol. 88, no. 11, pp. 2783–2792, Nov. 2007, doi: 10.1890/07-0539.1.

[93] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[94] B. Ibragimov and G. Gusev, “Minimal Variance Sampling in Stochastic Gradient Boosting,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2019/file/5c8cb735a1ce65dac514233cbd5576d6-Paper.pdf

[95] A. Maulana, R. P. F. Afidh, N. B. Maulydia, G. M. Idroes, and S. Rahimah, "Predicting Obesity Levels with High Accuracy: Insights from a CatBoost Machine Learning Model," *Infolitika Journal of Data Science*, vol. 2, no. 1, pp. 17–27, May 2024, doi: 10.60084/ijds.v2i1.195.

[96] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychol Rev*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.

[97] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmos Environ*, vol. 32, no. 14–15, pp. 2627–2636, Aug. 1998, doi: 10.1016/S1352-2310(97)00447-0.

[98] G. E. Hinton, "CONNECTIONIST LEARNING PROCEDURES," in *Machine Learning*, Elsevier, 1990, pp. 555–610. doi: 10.1016/B978-0-08-051055-2.50029-8.

[99] A. R. Rout, "Ml-advantages and disadvantages of linear regression," 2020, *GeeksforGeeks. https://www. geeksforgeeks. org/ml-advantagesand~…*.

[100] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. Jordan, and T. Petsche, Eds., MIT Press, 1996. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf

[101] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat Comput*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[102] P. Mehta *et al.*, "A high-bias, low-variance introduction to Machine Learning for physicists," May 2019, doi: 10.1016/j.physrep.2019.03.001.

[103] W. M. Adams, "Free transplantation of the nipples and areolae," *Surgery*, vol. 15, no. 1, pp.

186–195, Jan. 1944, doi: 10.5555/uri:pii:S0039606044900589.

[104] A. Nimboriboonporn and S. Chuthapisith, “Nipple-areola complex reconstruction.,” *Gland Surg*, vol. 3, no. 1, pp. 35–42, May 2014, doi: 10.3978/j.issn.2227-684X.2014.02.06.

[105] H. Awad, K. Al Ghoul, H. Al Osman, and N. Baddour, “Enhancing Nipple Positioning Accuracy in Chest Reconstruction Surgery: An Automated Machine Learning Approach,” in *IEEE Medical Measurements & Applications (MeMeA)*, Chania: IEEE, May 2025. doi: http://dx.doi.org/10.1109/MeMeA65319.2025.11068037.

[106] E. R. Jefferson and E. Trucco, “The challenges of assembling, maintaining and making available large data sets of clinical data for research,” in *Computational Retinal Image Analysis: Tools, Applications and Perspectives*, Elsevier, 2019, pp. 429–444. doi: 10.1016/B978-0-08-102816-2.00021-6.

[107] E. Wood *et al.*, “Fake It Till You Make It: Face analysis in the wild using synthetic data alone,” May 2021, [Online]. Available: http://arxiv.org/abs/2109.15102

[108] A. Saviolo *et al.*, “Learning to Segment Human Body Parts with Synthetically Trained Deep Convolutional Networks,” 2022, pp. 696–712. doi: 10.1007/978-3-030-95892-3_52.

[109] G. Riegler, M. Urschler, M. Ruther, H. Bischof, and D. Stern, “Anatomical Landmark Detection in Medical Applications Driven by Synthetic Data,” in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., May 2016, pp. 85–89. doi: 10.1109/ICCVW.2015.21.

[110] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, pp. 321–331, May 2018, doi: 10.1016/j.neucom.2018.09.013.

[111] T. Ahmed and S. Choudhury, “An Integrated Approach to AI-Generated Content in e-health,” May 2025, [Online]. Available: http://arxiv.org/abs/2501.16348

[112] Z. Wang, Q. Ma, W. Wan, H. Li, K. Wang, and Y. Tian, “Is this Generated Person Existed

in Real-world? Fine-grained Detecting and Calibrating Abnormal Human-body," May 2024, [Online]. Available: http://arxiv.org/abs/2411.14205

[113] J. Betker *et al.*, "Improving Image Generation with Better Captions."

[114] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing Prompts for Text-to-Image Generation." [Online]. Available: https://github.com/microsoft/LMOps

[115] T. Cao, C. Wang, B. Liu, Z. Wu, J. Zhu, and J. Huang, "BeautifulPrompt: Towards Automatic Prompt Engineering for Text-to-Image Synthesis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 1–11. doi: 10.18653/v1/2023.emnlp-industry.1.

[116] S. Rosenman, V. Lal, and P. Howard, "NeuroPrompts: An Adaptive Framework to Optimize Prompts for Text-to-Image Generation," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2311.12229

[117] E.-Y. Eom and J.-M. Lee, "Experience and ethical awareness of using artificial intelligence platforms." [Online]. Available: www.bpasjournals.com

[118] B. Omurkulov, D. Ylhamov, M. K. Gokgol, and M. Gokgol, "Comparative Recognition Analysis of Image Accuracy: A Study of OpenAI and Gemini in Matching Original Visuals", doi: 10.31039/plic.2024.12.282.

[119] A. Borji, "Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2," May 2022, [Online]. Available: http://arxiv.org/abs/2210.00586