FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Deep Learning for the Aesthetic Evaluation of Breast Cancer Treatment

Fábio Almeida Teixeira

Mestrado em Engenharia Informática e Computação

Supervisor: Jaime dos Santos Cardoso

Second Supervisor: Maria Helena Sampaio de Mendonça Montenegro e Almeida

July 21, 2025

Deep Learning for the Aesthetic Evaluation of Breast Cancer Treatment

Fábio Almeida Teixeira

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Ricardo Pereira de Magalhães Cruz
Referee: Prof. Jaime dos Santos Cardoso
Referee: Prof. Wilson Santos Silva

July 21, 2025

Resumo

O resultado estético é uma variável determinante na qualidade de vida a longo prazo após cirurgia conservadora da mama ou reconstrução, mas continua a ser avaliado sobretudo por inspeção visual subjetiva. Esta dissertação investiga se os avanços recentes em *deep learning* podem proporcionar uma alternativa totalmente automática, ordinal e explicável.

Recorrendo a um conjunto de dados privado obtido no âmbito do projecto CINDERELLA (1,406 fotografias pós-operatórias anotadas com a escala de quatro níveis de Harvard), a avaliação estética é reformulada como um problema de ordenação de imagens. Avaliam-se três arquiteturas. R2-ResNeXt acrescenta uma cabeça de ordenação par-a-par a um *backbone* ResNet-50, permitindo uma regressão guiada pela ordenação. OrdinalCLIP ajusta um modelo de visão–linguagem de modo a que o espaço latente codifique a ordem semântica entre as classes *pobre*, *razoável*, *bom* e *excelente*. Partindo desta base, a proposta SiameseOrdinalCLIP mantém o *backbone* guiado por linguagem congelado e treina um comparador siamês leve com *losses* de hinge-ranking e de erro quadrático médio. O desbalanceamento de classes é mitigada através de morfismo sintético de casos *pobre* e pela seleção manual de imagens pré-operatórias como exemplos *excelente*. Um módulo *concept bottleneck* de simetria fornece explicabilidade ao nível de cada caso.

Num grupo de teste de 15%, a SiameseOrdinalCLIP com aumento de dados atinge em média 65,9% de acurácia e 87,7% de acurácia de ordenação, superando o padrão clínico BCCT.core em mais 8,7% e 30,2% respetivamente, reduzindo quase para metade o erro absoluto médio e quadrado. As ablações confirmam que o pré-treino visão–linguagem, a ordenação par-a-par e o aumento de dados são todos essenciais, enquanto a pseudo-labellung não traz, por agora, benefício. O *bottleneck* de simetria permite aos cirurgiões inspecionar ou desativar conceitos individuais com apenas uma queda marginal no desempenho.

O estudo demonstra, assim, que um enquadramento siamês orientado por linguagem e sensível à ordenação pode proporcionar uma avaliação objetiva, interpretável e de vanguarda dos resultados estéticos do tratamento do cancro da mama, representando mais um passo rumo a um seguimento clínico rotineiro e padronizado.

Palavras-chave: Aprendizagem Profunda, Cancro da Mama, Avaliação Pós-operatória, Avaliação Estética, Rede Siamesa, Classificação Ordinal, Aprendizagem Semi-supervisionada

Abstract

Aesthetic outcome is a key variable in long-term quality of life after breast-conserving surgery or reconstruction, yet today it is still judged largely by subjective visual analysis. This dissertation explores whether recent advances in deep learning can yield a fully automatic, ordinal and explainable alternative.

Using a private dataset obtained in the context of the CINDERELLA project (1,406 post-operative photographs labelled with the four-grade Harvard scale), aesthetic assessment is recast as an image-ranking problem. Three architectures are evaluated. R2-ResNeXt adds a pair-wise ranking head to a ResNet-50 backbone, enabling ranking-guided regression. OrdinalCLIP fine-tunes a vision–language model so the latent space encodes the semantic order between the classes *poor*, *fair*, *good*, and *excellent*. Building on this, the proposed SiameseOrdinalCLIP freezes the language-guided backbone and trains a lightweight Siamese comparator with hinge-ranking and mean-squared error losses. Class imbalance is mitigated through synthetic morphing of *poor* cases, and hand selection of preoperative images as *excellent* images. A symmetry concept-bottleneck module provides case-level explainability.

On an unseen 15% test split, SiameseOrdinalCLIP with balanced augmentation attains an average of 65.9% accuracy and 87.7% ranking accuracy, exceeding the clinical gold standard BCCT.core by 8.7% and 30.2% respectively, while nearly halving mean square and absolute error. Ablations confirm that vision–language pre-training, pair-wise ranking, and data augmentation are each essential, whereas pseudo-labelling currently offers no benefit. The symmetry bottleneck allows surgeons to inspect or disable individual concepts with only a marginal drop in performance.

The study therefore demonstrates that a ranking-aware, language-guided Siamese framework can deliver objective, interpretable and state-of-the-art assessment of breast-cancer aesthetic outcomes, representing another step in routine, standardised follow-up in clinical practice.

Keywords: Deep Learning, Breast Cancer, Post-operative Evaluation, Aesthetic Evaluation, Siamese Network, Ordinal Classification, Semi-Supervised Learning

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. This dissertation touches four of those goals in a direct, measurable way.

SDG 3 Good Health and Well-being

SDG 5 Gender Equality

SDG 9 Industry, Innovation and Infrastructure

SDG 10 Reduced Inequalities

SDG	Target	Contribution	Suggested indicators
3	3.8	Fast, objective reports boost access to quality care.	Proposed a novel architecture for objective aesthetic evaluation of breast cancer treatments Evaluated using RMSE, accuracy, MAE, Macro-MAE and ranking accuracy.
5	5.6 & 5.b	Enabling access to appropriate healthcare by empowering women to make informed decisions	Objective treatment assessment allows women to make informed decisions based on their personal needs. Model trained using a variety of different surgeries.
9	9.5	Novel vision–language model furthers research in medical imaging.	Implementation of a novel architecture. Contribution to two research papers.
10	10.2	Promote inclusivity in women's health by enabling access to tools that are agnostic to body type and ailment.	Training data includes a variety of body types.

Acknowledgements

First and foremost, I wish to express my profound gratitude to my supervisors, Jaime and Helena, whose guidance, patience and immense help turned every challenge into a learning opportunity. I am equally indebted to my friends and family for their unwavering support, constant encouragement and love throughout this journey.

I gratefully acknowledge INESC TEC for welcoming me into its research community and for supporting this work through the Research Initiation Grant (BII), reference 11598/BII-E_B4/2025, awarded in the scope of the HfPT – Health from Portugal agenda funded by IAPMEI. The resources and collaborative environment provided under this grant were essential for the successful completion of my Master's research.

Finally, my sincere thanks go to everyone who, in ways big or small, helped shape this dissertation. Your contributions are deeply appreciated and will not be forgotten.

Fábio Almeida Teixeira

"Whether you think you can or you can't, you are right."

Henry Ford

Contents

List of Figures

List of Tables

List of Acronyms

Adam	Adaptive Moment Estimation (optimiser)
BAD	Breast Area Difference
BCD	Breast Contour Difference
BCE	Breast Compliance Evaluation
BCCT.core	Breast Cancer Conservative Treatment.cosmetic results software
BOD	Breast Overlap Difference
BRA	Breast Retraction Assessment
BYOL	Bootstrap Your Own Latent (self-supervised model)
CBM	Concept Bottleneck Model
CLIP	Contrastive Language–Image Pre-training
CNN	Convolutional Neural Network
CORAL	Consistent Rank Logits (rank-consistent ordinal regression)
CORN	Conditional Ordinal Regression for Neural Networks
DNN	Deep Neural Network
LBC	Lower Breast Contour
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
ROI	Region of Interest
SGD	Stochastic Gradient Descent
UNR	Upward Nipple Retraction

Chapter 1

Introduction

1.1 Context & Motivation

Breast cancer treatment has seen remarkable advancements in recent years, resulting in improved survival rates and a wider range of surgical options for patients [4]. However, post-operative aesthetic outcomes remain a significant concern for many patients. Surgical interventions, such as breast-conserving surgery or mastectomy followed by reconstruction, often result in cosmetic challenges, including asymmetry, scarring, or deformities. These aesthetic outcomes can negatively affect a patient's quality of life, self-esteem, and overall satisfaction with their treatment.

The assessment of these post-operative outcomes has traditionally relied on subjective evaluations by clinicians and patients, typically based on simple rating scales. While these evaluations are helpful, they are prone to inter-observer variability and lack standardisation due to biases [41, 47]. To address this, machine learning techniques have been adopted in recent years to provide more objective, automated assessments of post-operative breast aesthetics. This shift towards automated image classification was a promising step in standardising evaluations and providing more consistent, reliable feedback to both clinicians and patients [5].

Recently, several successful attempts at developing objective systems for assessing post-operative breast aesthetics have been made [5, 16, 56]. While these models have achieved good results, they rely heavily on manual annotations, which may not always be feasible in real-world clinical settings. Additionally, most work done has focused almost entirely on symmetry, which does not capture other important aesthetic factors, such as skin colouration and scar visibility. Additionally, architectures that showcase promising potential for objective evaluation remain largely unexplored in this field. Therefore, it would be of great interest to develop a model that incorporates these additional characteristics to offer a more holistic evaluation of aesthetic outcomes, potentially reducing the need for extensive manual annotations while still providing meaningful assessments.

1.2 Problem

While aesthetic assessment methods for breast reconstruction surgery have advanced significantly, offering better objectivity and performance, some clear improvements are still possible. Existing methods for classifying aesthetic outcomes tend to overemphasise symmetry or rely heavily on manual intervention. However, these approaches are time-consuming and often fail to capture the full range of aesthetic concerns patients may have.

Recent developments, including the state-of-the-art deep learning-based system by Silva et al. (2022) [51], provide a more objective approach by automating the aesthetic assessment process. This method improves performance but is still limited by binary classifications (e.g., Excellent, Good vs. Fair, Poor). The key challenges include expanding the ordinal nature of the classification (Excellent, Good, Fair, Poor), exploring beyond symmetry to other aesthetic metrics, and improving the overall robustness of the neural network architecture.

This dissertation aims to address these gaps by developing a deep learning model capable of providing comprehensive, ordinal aesthetic evaluations without the need for manual annotations, and using a Siamese architecture to provide a more objective result. By leveraging recent increases in available data, as well as the exploration of semi-supervised learning techniques, the model will not only account for multiple aesthetic factors such as symmetry, scar visibility, and skin discolouration but also improve its performance and robustness. Solving this problem would not only standardise the aesthetic evaluation process but also improve patient satisfaction and clinical decision-making by offering more realistic outcomes.

1.3 Research Questions

To investigate the effectiveness of a deep learning model for post-operative breast aesthetic evaluation, the following research questions are posed:

1. How can a deep learning model be developed to automatically classify post-operative breast images into multiple aesthetic categories, considering not only symmetry but also scar visibility and skin discolouration?
2. Can Siamese Networks be used successfully for objective aesthetic assessment in breast reconstructive surgery?
3. How can semi-supervised learning techniques be effectively employed to enhance model performance using both labelled and unlabeled post-operative breast images?
4. Can a deep learning model fully eliminate the need for manual intervention in the process of aesthetic assessment for breast reconstructive surgery?

1.4 Main Contributions

The main contributions of this research are as follows:

- We perform an extensive survey over the main methods for aesthetic evaluation and prediction over many types of plastic and oncological surgeries. The work aims to review key developments, identify cross-domain similarities, highlight major limitations, and explore future research directions. The paper is currently being prepared for submission.
- We propose a novel aesthetic assessment architecture that leverages language and ranking to guide its understanding of ordinality and better match the human decision-making process. This architecture improves upon the state-of-the-art while being fully automatic and allowing built-in explainability. We submitted the article with this work to "Deep-Brea3th 2025: 2nd Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care", held as part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025).

Chapter 2

Background

2.1 Introduction

This chapter establishes the fundamental ideas necessary to comprehend the approaches utilised in this study. It starts with an examination of neural networks, going into detail about the basic components and learning methods, including artificial neurons, activation functions, forward and backward propagation, and loss functions.

Next, the attention turns to deep learning and its importance in managing intricate data representations. Convolutional and pooling layers are studied in Convolutional Neural Networks (CNNs), showcasing their effectiveness in analysing images. We also introduce the idea of explainability in deep architectures, which is often a requirement for application in medical domains.

Since aesthetic evaluation is an inherently subjective task where labels may lack consistency, this chapter also introduces weakly supervised learning, highlighting its importance in situations with restricted or inaccurate labelled data.

Additionally, given the ordinal nature of the problem at hand, the text addresses ordinal classification by explaining traditional approaches and the difficulties of arranging data with an inherent rank but without clear numerical distinctions. In conclusion, the combination of ranking models in deep learning frameworks and their particular uses in medical imaging are examined, preparing for the suggested method.

This chapter establishes the theoretical groundwork that provides a comprehensive understanding of the concepts that are required for this research aimed at improving the evaluation of breast cancer surgery outcomes through advanced AI methods.

2.2 Neural Networks

Neural networks are computational models inspired by the human brain's structure and functionality. They consist of interconnected layers of *artificial neurons* that process data by simulating the way biological neurons signal to one another. Neural networks are fundamental to many deep learning applications, providing the architecture for machines to learn complex patterns and make

intelligent decisions. The simplest neural networks are made up of layers of artificial neurons, which take in one or more input values and produce an output based on linear combinations of the inputs and activation functions, as explained in the following subsections. This chain of operations is known as forward propagation during training. The last layer generates the final prediction, which is evaluated by a loss function. This information is then used to tune the parameters of the network, so as to minimise the loss function. This training process occurs multiple times over the training data (epochs), enabling the neural network to gradually learn the relevant features and patterns in the data.

2.2.1 Artificial Neurons

Artificial neurons, also known as *perceptrons* (shown in figure 2.1), proposed by Frank Rosenblatt in 1958 [45], are the basic units of a neural network. They mimic the behaviour of biological neurons by receiving inputs, processing them, and producing an output.

An artificial neuron receives input signals $x_1, x_2, ..., x_n$ and associates each with a corresponding weight $w_1, w_2, ..., w_n$. The neuron computes a weighted sum of the inputs and passes this sum through an activation function to produce an output:

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{2.1}$$

Here, b represents the *bias* term, which allows the activation function to be shifted left or right on the graph, providing additional flexibility in the model.

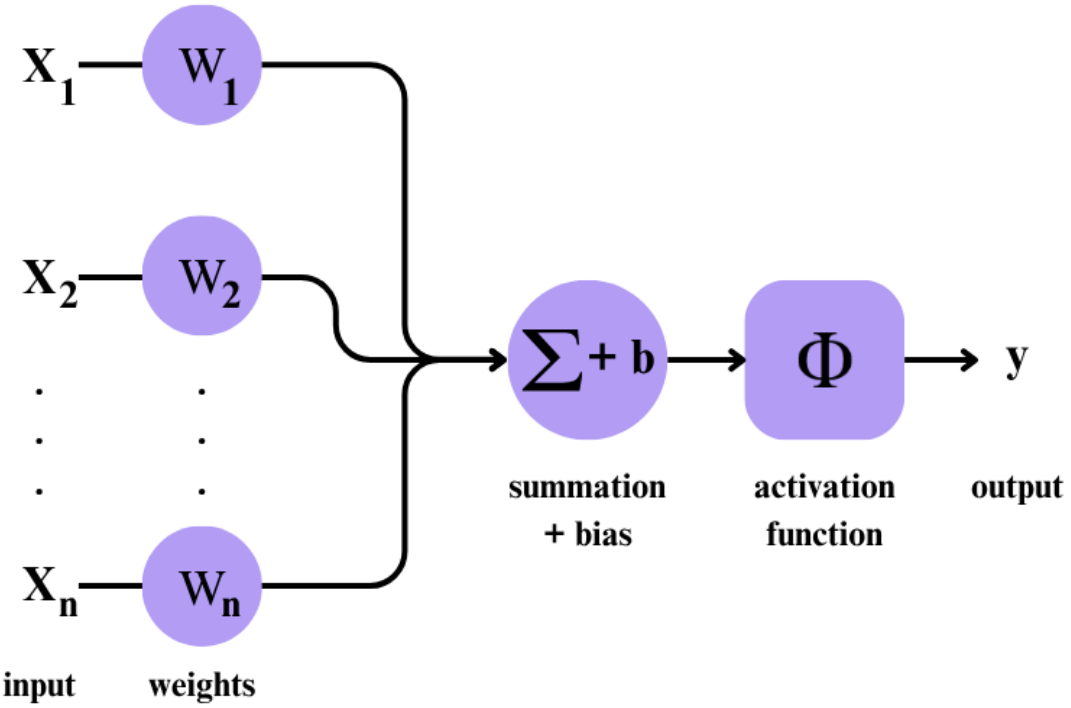


Figure 2.1: Artificial Neuron

2.2.2 Activation Functions

Activation functions introduce non-linearity into the neural network, enabling it to learn and model complex patterns. Without activation functions, the network would behave like a linear regression model, regardless of the number of layers.

Common activation functions include [27]:

- **Sigmoid Function**: Maps input values into the range (0, 1):

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

- **Hyperbolic Tangent (Tanh)**: Maps input values into the range (-1, 1), centred at zero.

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.3}$$

- **Rectified Linear Unit (ReLU)**: Introduces sparsity and mitigates the vanishing gradient problem, a phenomenon that occurs when the gradients used to update the network become extremely small, causing the training process to stagnate. ReLU avoids this by outputting zero for negative inputs and linear for positive inputs:

$$\text{ReLU}(x) = \max(0,x) \tag{2.4}$$

- **Softmax Function**: Used in the output layer for multi-class classification problems to produce a probability distribution.

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2.5}$$

2.2.3 Forward Propagation

Forward propagation is the process by which input data passes through the neural network layers to generate an output. In each layer, the neurons compute the weighted sum of inputs, apply the activation function, and pass the result to the next layer.

For a simple neural network with one hidden layer, the forward propagation can be expressed as:

1. Compute the weighted sum for the hidden layer neurons:

$$z^{(1)} = W^{(1)}x + b^{(1)} \tag{2.6}$$

2. Apply the activation function to the hidden layer outputs:

$$y^{(1)} = \phi^{(1)}\left(z^{(1)}\right) \tag{2.7}$$

3. Compute the weighted sum for the output layer:

$$z^{(2)} = W^{(2)} y^{(1)} + b^{(2)} \tag{2.8}$$

4. Apply the activation function to produce the final output:

$$y^{(2)} = \phi^{(2)} \left(z^{(2)} \right) \tag{2.9}$$

Here, $W^{(1)}$ and $W^{(2)}$ are the weight matrices for the hidden and output layers, respectively, $b^{(1)}$ and $b^{(1)}$ are the bias vectors, $\phi^{(1)}$ and $\phi^{(2)}$ represent the activation functions, and x is the input vector.

2.2.4 Loss Functions

Loss functions quantify the difference between the predicted outputs and the actual targets, guiding the optimization of the network during training. The choice of loss function depends on the type of problem being addressed.

Common loss functions include:

- **Mean Squared Error (MSE)**: Used for regression tasks, measuring the average squared difference between predicted and actual values.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.10}$$

  Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

- **Cross-Entropy Loss**: Used for classification tasks, measuring the dissimilarity between the predicted probability distribution and the actual distribution.

  - *Binary Cross-Entropy* (for binary classification):

$$\text{Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{2.11}$$

  - *Categorical Cross-Entropy* (for multi-class classification):

$$\text{Loss} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{i,k} \log(\hat{y}_{i,k}) \tag{2.12}$$

    Where K is the number of classes, $y_{i,k}$ is a binary indicator (0 or 1) if class label k is the correct classification, and $\hat{y}_{i,k}$ is the predicted probability that the observation is of class k.

2.2.5 Backpropagation and Learning

The backpropagation process enables a neural network to learn from the loss function by updating its weights and biases. This is done by computing the gradient of the loss concerning each parameter and then propagating the error backward through the network. The steps involved in backpropagation are:

1. **Compute the Loss**: The loss function should be computed with respect to the network's output and actual target values.
2. **Calculate Gradients**: Find the partial derivatives of the loss function with respect to each weight and bias.
3. **Update Weights and Biases**: Alter the weights and biases in the opposite direction of the gradient in order to minimise the loss function. This step is typically done using an optimisation algorithm like Gradient Descent.
4. **Iterate**: Repeat the back and forward propagation steps for multiple epochs until the network converges to a minimal loss.

Optimisation algorithms make backpropagation more efficient. Variants such as SGD, Adam, and RMSprop algorithms, dynamically adjust the learning rate and help escape local minima.

2.3 Deep Learning

Deep learning is a subset of machine learning that focuses on neural networks with multiple layers, known as *deep neural networks* (DNNs). These networks have the capacity to model complex patterns in data by learning hierarchical representations. Deep learning has revolutionised fields such as computer vision, natural language processing, and speech recognition by enabling models to learn features directly from complex data.

2.3.1 Deep Neural Networks

A deep neural network is an artificial neural network with multiple hidden layers between the input and output layers, as shown in image 2.2. The depth of the network allows it to capture intricate structures in data by learning multiple levels of abstraction.

The depth and complexity of DNNs enable them to approximate highly nonlinear functions, making them suitable for tasks like image and speech recognition.

2.3.2 Overfitting and Regularization

One of the challenges in training deep neural networks is *overfitting*, where the model learns the training data too well, including its noise and outliers, resulting in poor generalisation to new data. Overfitting occurs when the model is too complex relative to the amount of training data.

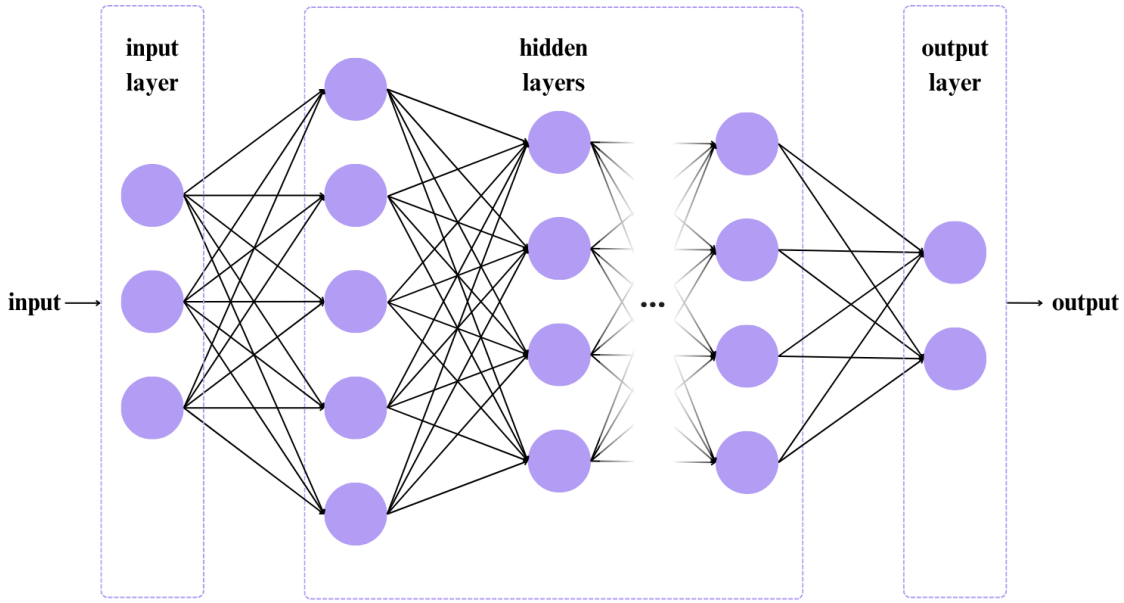


Figure 2.2: Basic fully connected DNN architecture.

Overfitting can be identified when the model's performance on the training data is significantly better than on validation or test data. This indicates that the model has not learned the underlying patterns but has memorised the training examples.

Regularisation methods are used to prevent overfitting by adding constraints to the learning process:

- **L1 and L2 Regularisation**: Add a penalty term to the loss function proportional to the magnitude of the weights.
  - *L2 Regularisation*:

$$\text{Loss}_{\text{total}} = \text{Loss} + \lambda \sum_i w_i^2 \tag{2.13}$$

  - *L1 Regularisation*:

$$\text{Loss}_{\text{total}} = \text{Loss} + \lambda \sum_i |w_i| \tag{2.14}$$

  Where λ is the regularisation parameter controlling the strength of the penalty.

- **Dropout**: Randomly sets a fraction of the neurons' outputs to zero during training, preventing units from overly adapting to the data.
- **Early Stopping**: Monitors the model's performance on a validation set and stops training when performance begins to degrade.
- **Data Augmentation**: Increases the diversity of training data by applying transformations such as rotation, scaling, and flipping to existing data (in the case of images).

- **Batch Normalisation**: Normalises the input of each layer to stabilise and accelerate training.

2.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep neural networks specifically designed to process data with a grid-like topology, such as images. They have achieved remarkable success in computer vision tasks due to their ability to capture spatial hierarchies of features through local connections and shared weights.

2.4.1 Convolutional Layers

The core component of a CNN is the *convolutional layer*, which applies convolution operations to the input data to extract local features.

The convolution operation involves sliding a small matrix, known as a *kernel* or *filter*, over the input data to produce a feature map. Figure 2.3 shows the first operation of the convolution layer. For a two-dimensional input I and a kernel K, the discrete convolution is defined as:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{2.15}$$

Here $S(i,j)$ represents the output feature map at position (i,j) and the sums are over the dimensions of the kernel.

Two important hyperparameters in convolutional layers are *padding* and *stride*:

- **Padding (P)**: Refers to adding zeros around the border of the input matrix. Padding controls the spatial size of the output. With appropriate padding, the spatial dimensions can be preserved after convolution.
- **Stride (S)**: Defines the number of pixels by which the kernel moves over the input matrix. A larger stride reduces the spatial dimensions of the output feature map.

An input image may have multiple channels (e.g., RGB channels). Convolutional layers can process multi-channel inputs and produce multiple output feature maps by applying a set of kernels.

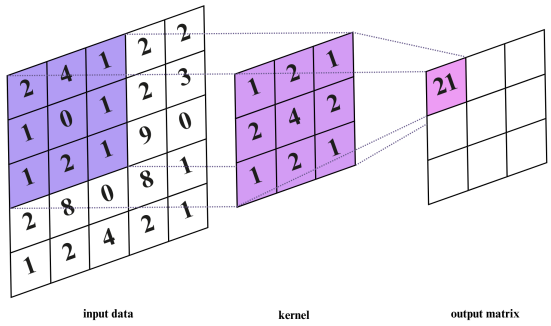


Figure 2.3: Convolution operation for single channel input and output

2.4.2 Pooling Layers

Pooling layers reduce the spatial dimensions of the input data, which decreases the number of parameters, and therefore computation, in the network, and helps control overfitting while still maintaining the most important information.

Common types of pooling include:

- **Max Pooling:** Selects the maximum value within a pooling window.

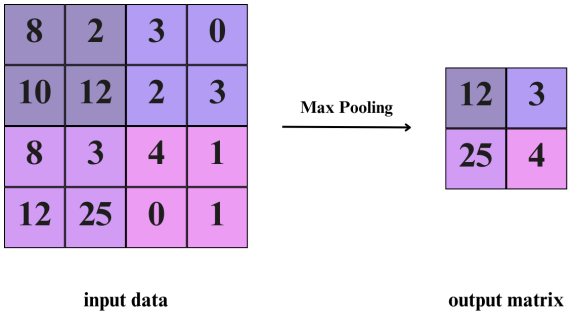


Figure 2.4: Max pooling operation with 2×2 stride and window size.

- **Average Pooling:** Computes the average of all values within a pooling window.

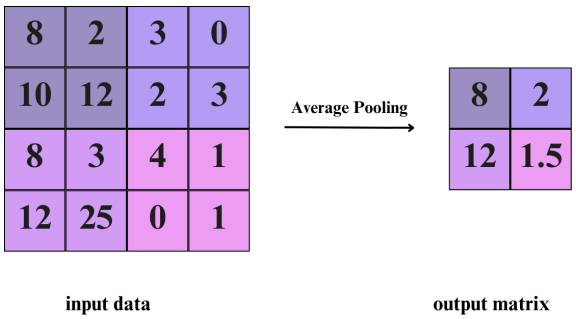


Figure 2.5: Average pooling operation with 2×2 stride and window size.

Pooling layers have hyperparameters similar to convolutional layers, such as window size and stride.

2.4.3 CNN Architectures

CNNs are constructed by stacking multiple layers of convolutions, activations, and pooling, often followed by fully connected layers.

Several landmark CNN architectures have significantly advanced the field:

- **AlexNet**: Achieved breakthrough performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, using a deeper architecture and ReLU activations [31].
- **VGGNet**: Explored the impact of the depth of the network on its performance, by using a larger number of layers with very small (3x3) convolution filters, leading to significantly improved performance [54].
- **GoogLeNet (Inception)**: Introduced the Inception module, which allows for parallel convolutional operations within the same layer [58].
- **ResNet**: Addressed the degradation problem in deep networks by introducing residual learning and skip connections [25].

2.4.4 Applications in Image Analysis

CNNs have become the de facto standard for image-related tasks due to their ability to automatically learn hierarchical feature representations. They are extensively used for tasks such as:

- Classification: CNNs classify images by learning discriminative features that distinguish between different classes. They have achieved state-of-the-art performance on datasets such as ImageNet [13].
- Object Detection: Combining CNNs with region proposal methods enables models to detect and localise multiple objects within an image. Examples include Region-based CNNs (R-CNN) [19] and its variants [43].
- Segmentation: Semantic segmentation assigns a class label to each pixel in an image. Fully Convolutional Networks (FCNs) adapt CNNs for pixel-wise predictions.

2.4.4.1 Medical Imaging

In medical imaging, CNNs assist in tasks such as:

- **Disease Diagnosis**: Identifying pathological regions in medical scans (e.g., tumours in MRI images).
- **Image Segmentation**: Delineating anatomical structures or lesions for treatment planning.
- **Image Enhancement**: Improving image quality through denoising or super-resolution techniques.

2.5 Siamese Networks

Siamese networks are a unique architecture of neural networks designed to measure similarity between two inputs by learning a shared feature representation. Introduced by Bromley et al. (1993) [2] for signature verification, they have proven effective across various domains, including face recognition, image comparison, and metric learning tasks.

2.5.1 Architecture of Siamese Networks

A Siamese network consists of two or more identical subnetworks with shared weights, parameters, and architecture. Each subnetwork processes one of the inputs independently, and the outputs are combined using a distance metric (e.g., Euclidean distance, cosine similarity) to determine the similarity or dissimilarity of the inputs. This shared structure ensures that the same features are extracted and compared from both inputs.

Given two input images, x_1 and x_2, the subnetworks map each input to a feature vector, $f(x_1)$ and $f(x_2)$, respectively. A distance metric, such as the Euclidean distance or cosine similarity, is then applied to these feature vectors to quantify their similarity or dissimilarity. The network is usually trained to minimise a contrastive loss function, which encourages similar inputs to have embeddings that are closer together and dissimilar inputs to have embeddings that are farther apart.

2.5.2 Contrastive Loss Function

The contrastive loss function is a key component in training Siamese networks. It is defined as:

$$\mathscr{L}(y, D_W) = (1-y)\frac{1}{2}D_W^2 + y\frac{1}{2}\{\max(0, m - D_W)\}^2, \tag{2.16}$$

where:

- $y \in \{0,1\}$ is the label indicating whether the input pair is similar ($y = 0$) or dissimilar ($y = 1$).
- $D_W = \|f(x_1) - f(x_2)\|_2$ is the Euclidean distance between the feature embeddings; however, other distance metrics may be used.
- $m > 0$ is a margin parameter that defines the minimum acceptable distance for dissimilar pairs.

This loss function penalises similar pairs that are far apart in the embedding space and dissimilar pairs that are too close, fostering a clear separation in the learned representation.

2.5.3 Advantages of Siamese Networks

Siamese networks have several advantages, making them particularly suitable for image-based tasks in medical imaging. In the context of breast reconstructive surgery, Siamese networks offer a powerful approach for evaluating aesthetic outcomes, since ranking post-operative images is a

task inherently rooted in similarity assessment rather than direct classification. The use of Siamese networks enables:

- **Pairwise Analysis**: Directly comparing post-operative results with reference images, such as ideal outcomes or patient-specific baselines.
- **Objective Aesthetic Evaluation**: Reducing subjectivity by learning an embedding space where visually similar outcomes are closer together, while dissimilar results are farther apart.
- **Handling Ordinal Labels**: By learning continuous embeddings, the network can naturally handle ordinal or graded aesthetic labels without relying solely on discrete categories.

For instance, if the goal is to determine whether a reconstructed breast closely resembles its contralateral counterpart, Siamese networks can effectively quantify this similarity, providing actionable insights for both surgeons and patients.

In conclusion, the versatility and pairwise comparison capabilities of Siamese networks make them an ideal choice for tasks requiring nuanced evaluations, making them ideal for the scenario described in this work.

2.6 Explainability in Deep Neural Networks

Deep neural networks have achieved remarkable performance in tasks like medical image analysis, often matching or exceeding expert accuracy. However, their "black-box" nature, an opaque mapping from input to output, presents a serious challenge for trust and adoption in critical domains. In clinical settings, especially, the lack of interpretability is viewed as a barrier to using AI for diagnosis because users (e.g. physicians) need to understand why a model made a prediction before acting on it. This challenge reflects a major problem between model accuracy and interpretability: the most accurate models (deep or ensemble models) are typically the least transparent, forcing a trade-off between performance and explainability. In response, the field of Explainable AI (XAI) has emerged with techniques to make AI decisions more understandable. The goal is to build trustworthy AI systems by providing human-interpretable reasons for predictions. This transparency is crucial not only for user trust, but also to satisfy ethical and legal requirements (e.g. a "right to an explanation" in decisions affecting individuals). Explainability methods, therefore, aim to bridge the gap between complex models and human understanding, ensuring that AI decisions can be interpreted, validated, and appropriately questioned before deployment.

2.6.1 Concept Bottleneck Models

Concept Bottleneck Models (CBMs)[30] are an approach to deep learning that embeds interpretability directly into the model's structure. In a CBM, the neural network is divided into two stages: first it predicts a set of human-defined concepts from the input, and then uses those concepts to predict the final output label. These intermediate concepts are high-level, semantically

meaningful features (for example, for breast aesthetic analysis, concepts might include the presence of scars, or discolouration). The layer of concept predictions forms a "bottleneck" through which all information must pass, ensuring that the model's decisions are grounded in interpretable concepts rather than abstract latent features. The purpose of CBMs is to make the model **inherently explainable** and interactive. Because predictions are explicitly based on understandable concepts, one can ask why the model made a certain decision in terms of those concepts. For instance, using a CBM for breast aesthetic analysis, a doctor could be told that "the model predicts *poor* aesthetic value because it detected scars and discolouration". Moreover, CBMs support counterfactual concept-level inquiries: one can question "if the model did not think there was a scar, would it still predict *poor*?" – a question that standard end-to-end networks cannot easily answer. By construction, it is also possible to intervene on the model's reasoning: if a concept prediction is suspected to be wrong, a user can correct that concept and propagate the change to see how the final prediction updates. This property enables a richer human–model interaction loop than conventional models. Importantly, recent research has demonstrated that CBMs can achieve accuracy comparable to standard end-to-end deep models while offering these transparency benefits [30]. Koh et al. (2020) showed that on tasks like medical chest X-ray grading and bird species identification, a concept bottleneck model reached competitive performance to a black-box model, yet its predictions could be explained in terms of high-level concepts (e.g. clinical findings or bird attributes). Because each prediction comes with a vector of concept outputs, a practitioner can inspect which concepts the model relied on, providing a clear rationale for the decision. An additional advantage is that if humans are able to correct any mispredicted concepts at test time, the final accuracy improves significantly. This demonstrates how CBMs facilitate a form of human-AI collaboration, where the human expert and model can communicate via the concept interface. CBMs thus offer interpretable-by-design models: the entire decision pipeline is transparent. However, it is worth noting that they require the availability of concept annotations for training data, which can be labour-intensive to obtain in some domains, and they impose constraints on the model's representation. If not properly implemented, a CBM might allow extra, uninterpretable information to slip through the bottleneck (a phenomenon known as concept leakage where the concept layer carries latent information not truly captured by the defined concepts).

2.6.2 Post-hoc Explainability Methods vs. CBMs

In contrast to CBMs' built-in interpretability, most deep learning models today are explained post hoc, i.e. after the model is trained, using various heuristic techniques. These post-hoc explainability methods do not alter the model's internal structure; instead, they analyse an already-trained black-box to provide insights into its predictions. Below, we review some popular approaches:

- Saliency Maps: Saliency methods produce a heatmap over the input (e.g. an image) indicating which regions most influenced the model's prediction. Techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) use the gradient flowing into the last convolutional layer to compute a coarse localisation of important image regions for a given class.

Saliency maps are widely used in medical image analysis as a way to verify that a neural network is "looking" at the relevant anatomy or pathology; they visualise whether the model focuses on the same areas a clinician would consider important [53].

- Local Interpretable Model-agnostic Explanations (LIME): LIME is a technique that explains an individual prediction by approximating the complex model with a simpler interpretable model in the vicinity of the input. In practice, LIME generates many perturbations of the input and observes the black-box's outputs, then fits a small linear model (or another simple model) to mimic the black-box's behaviour on those perturbed samples. The coefficients of this local surrogate model serve as an explanation for which features influence the prediction [44].

- Shapley Additive Explanations(SHAP): SHAP is another model-agnostic explanation framework based on cooperative game theory. It assigns each feature a Shapley value, which represents that feature's contribution to the difference between the actual prediction and a baseline expectation. In essence, SHAP explains a single prediction by distributing the prediction score among the input features in a way that satisfies desirable properties (such as consistency with feature importance and additivity). One can interpret SHAP values as: "if we start from a baseline prediction and add this feature, how does it change the prediction (on average)?". Summing all feature contributions yields the actual prediction [36].

In summary, CBMs provide built-in, concept-level transparency that aligns well with clinical reasoning, whereas post-hoc tools supply complementary but sometimes less trustworthy insights.

2.7 Weakly Supervised Learning

Weakly supervised learning is a branch of machine learning that deals with training models using data that is not fully annotated or labelled with high precision. In many real-world scenarios, obtaining fully labelled datasets is impractical due to the cost, time, and expertise required for accurate labelling. Weakly supervised learning addresses this issue by utilizing incomplete or inaccurate labels to build effective predictive models.

This section explores the various forms of weak supervision, the challenges associated with learning from weakly labelled data, and the methodologies developed to overcome these challenges. We will also discuss the applications of weakly supervised learning in computer vision and medical imaging, highlighting its relevance in the context of breast reconstruction surgery assessment.

2.7.1 Forms of Weak Supervision

Weak supervision can be classified into three primary categories based on the nature of the labelling imperfections:

1. **Incomplete Supervision**: Only a subset of the training data is labelled, while the rest remains unlabeled. This scenario is common when labelling every instance is too resource-intensive.

2. **Inexact Supervision**: The available labels are coarse or imprecise. For example, having image-level labels without specific annotations for objects within the image.

3. **Inaccurate Supervision**: The labels provided may contain errors or noise due to human mistakes or automated labelling processes.

In this work, only incomplete and inaccurate supervision will be addressed.

2.7.2 Challenges in Weakly Supervised Learning

Learning through weak supervision introduces several challenges, such as inexact or coarse labels, making it difficult for models to learn fine-grained patterns, as the true correspondence between inputs and outputs cannot be precisely defined. Inaccurate labels can also mislead the learning process, causing models to learn incorrect associations. In addition, incomplete supervision can result in overfitting or underfitting due to the scarcity of labelled data. Addressing these challenges is essential for the success of weakly supervised learning methods.

2.7.3 Techniques in Weakly Supervised Learning

Several methodologies have been developed to mitigate the challenges posed by weak supervision:

2.7.3.1 Semi-Supervised Learning

Semi-supervised learning combines a small amount of labelled data with a large amount of unlabeled data during training. It assumes that the data distribution can be captured by the unlabeled data, aiding the learning process.

Approaches

- **Generative Models**: Semi-supervised methods can use generative models to model the joint distribution of features and labels, allowing label inference for unlabeled data. For instance, deep generative approaches based on variational inference have been shown to effectively generalise from small labelled sets to large unlabeled ones [29].

- **Consistency Regularisation**: This approach encourages a model to produce stable, consistent predictions when its inputs are perturbed or augmented. Techniques like Temporal Ensembling and Mean Teacher enforce that the network's output for an example remains the same under different noise or augmentation conditions [59].

- **Pseudo-Labelling**: Pseudo-labelling assigns artificial labels to unlabeled examples using the model's current predictions. In each training iteration, the class with the highest predicted probability for an unlabeled sample is treated as its label, and the model is then trained on these pseudo-labelled examples along with the true labelled data [32].

2.7.3.2 Self-Supervised Learning

Self-supervised learning creates surrogate tasks where the supervision signal is derived from the data itself, eliminating the need for external labels.

Examples

- **Image Inpainting**: In this self-supervised task, a model learns to fill in missing parts of an image. For example, context encoder networks are trained to generate the content of a missing region of an image given its surroundings [40]. By trying to predict the missing pixels, the model is forced to learn high-level features about objects and scenes, which can later be fine-tuned for downstream tasks. This inpainting task provides a supervision signal (the known missing patch) derived purely from the data itself, requiring no manual labels.
- **Colourisation**: The model learns to colourise a grayscale image, using the original colour image as self-supervision. Approaches like Zhang et al. (2016) [60] treat colourisation as a pretext task where the network predicts plausible colours for a given black-and-white input. The model must infer semantic context (e.g. sky vs. grass) to assign realistic colours, thereby learning useful visual representations.

These tasks help the model learn useful representations that can be fine-tuned for the main task.

2.7.4 Evaluation Metrics

Evaluating models trained with weak supervision requires careful consideration. Metrics like accuracy, precision, recall, and F1-score are used but must be interpreted in the context of weak labels.

Techniques

- **Cross-Validation**: Using multiple data splits to ensure consistency.
- **Confidence Intervals**: Providing statistical significance of the results.

2.8 Ordinal Classification

Ordinal classification, also known as ordinal regression, is a type of predictive modelling where the target variable is categorical with a natural order or ranking among the categories, but the

intervals between the categories are not necessarily equal or known. This is distinct from nominal classification, where there is no inherent order among categories, and from metric regression, where the target variable is continuous.

In many real-world applications, data often comes in the form of ratings or rankings, such as customer satisfaction levels, stages of disease progression, or quality assessments. In the context of breast reconstruction surgery outcomes, ordinal classification can be used to predict or assess the quality of surgical results, which may be categorised into ordered levels such as *poor*, *fair*, *good*, and *excellent*.

2.8.1 Characteristics of Ordinal Data

Ordinal data possess a natural order among categories, but the exact differences between the categories are not quantifiable. For example, in a satisfaction survey with responses like "Very Dissatisfied," "Dissatisfied," "Neutral," "Satisfied," and "Very Satisfied," each response is ranked relative to the others, but the difference between "Satisfied" and "Very Satisfied" is not necessarily the same as between "Neutral" and "Satisfied."

Key characteristics of ordinal data include:

- **Order Relation**: There exists a total ordering among the categories.
- **Unequal Intervals**: The distances between adjacent categories are not assumed to be equal.
- **Non-Numeric Labels**: Categories may be represented by labels rather than numeric values.

These properties necessitate specialised techniques for modelling ordinal data that respect its inherent order.

2.8.2 Challenges in Ordinal Classification

Ordinal classification presents unique challenges:

- **Loss of Order Information**: Treating ordinal classes as nominal ignores the ordering, potentially reducing model performance.
- **Assumption Violations**: Applying regression techniques assumes equal intervals between categories, which may not hold.
- **Imbalanced Classes**: Ordinal datasets often have imbalanced class distributions, complicating model training.
- **Evaluation Metrics**: Selecting appropriate metrics that account for ordering is critical for model assessment.

Addressing these challenges requires models and loss functions that explicitly incorporate the ordinal nature of the data. Custom loss functions can be designed to penalise misclassifications according to the severity of the error (e.g., predicting class 1 when the true class is 3 is penalised more than predicting class 2).

2.8.3 Evaluation Metrics for Ordinal Classification

Evaluating ordinal classification models requires metrics that consider the ordering of classes.

- **Mean Absolute Error**: Treating ordinal classes as nominal ignores the ordering, potentially reducing model performance. MAE measures the average absolute difference between the predicted and true class labels:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.17}$$

- **Macro-Averaged MAE (Macro-MAE)**: Computes MAE separately for each ordinal class and then averages across classes to mitigate class imbalance. This ensures that each class contributes equally to the final metric, regardless of its frequency. Let K be the number of classes and N_k the number of instances with true label k. Then:

$$\text{Macro-MAE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{i:\ y_i=k}|y_i - \hat{y}_i|\right) \tag{2.18}$$

- **Root Mean Squared Error (RMSE)**: Computes the square root of the mean of squared errors, penalising larger deviations more heavily. RMSE is more sensitive to large errors than MAE, making it useful when large deviations are particularly undesirable.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2.19}$$

- **Quadratic Weighted Kappa (QWK)**: A metric that measures the agreement between two raters (true vs. predicted) while accounting for chance agreement and weighting disagreements according to their squared distance. Let O_{ij} be the observed joint histogram and E_{ij} the expected histogram under chance; define weight matrix $w_{ij} = \frac{(i-j)^2}{(K-1)^2}$. Then:

$$\kappa_{\text{QWK}} = 1 - \frac{\sum_{i,j=1}^{K} w_{ij} O_{ij}}{\sum_{i,j=1}^{K} w_{ij} E_{ij}} \tag{2.20}$$

- **Spearman's Rank Correlation Coefficient (ρ)**: Measures the correlation between the rankings of the true and predicted labels. Let $d_i = \text{rank}(y_i) - \text{rank}(\hat{y}_i)$ be the difference in ranks for the i^{th} instance. Then:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{2.21}$$

2.8.4 Recent Deep-Learning Approaches to Ordinal Classification

Early neural solutions often treated ordinal labels as a series of independent binary tasks, an approach that can yield rank-inconsistent outputs. To address this, CORAL – COnsistent RAnk Logits [3] adds a dedicated "ordinal layer" on top of any backbone network. The layer shares its weights across all rank thresholds and applies a monotonic constraint during training, guaranteeing that the predicted probability of belonging to class k + 1 is never higher than that of class k. CORAL is architecture-agnostic and has become a strong baseline in age estimation and medical staging problems, but weight sharing can limit representational flexibility when class boundaries are highly non-linear.

CORN – Conditional Ordinal Regression for Neural Networks ([50]) removes CORAL's weight-sharing constraint by modelling the ordinal target as a probability chain. The network learns a sequence of conditional classifiers—"is the label ≥ 1?", "≥ 2?" and so on, each conditioned on the previous answer. Because each boundary decision has its own parameters, CORN usually recovers stronger nonlinear decision functions while still enforcing overall rank consistency, and it consistently outperforms CORAL on standard benchmarks.

Recently, OrdinalCLIP [33] has shown that large vision–language models can be repurposed for ordinal tasks. Each rank is encoded as a trainable text prompt (e.g. "age estimation: the age of the person is age"), which combined with a CLIP image encoder maps the images to a continuous semantic space where the embeddings of the learned rank are ordered.

Chapter 3

Literature Review

3.1 Introduction

This literature review aims to provide an in-depth exploration of the available methodologies and techniques relevant to the aesthetic assessment of outcomes following the treatment of breast cancer.

Central to the work discussed in this dissertation lies the area of medical imaging, deep learning, and aesthetic evaluation. The review begins with an exploration of recent studies of aesthetic assessments in reconstructive breast surgery. Next, it delves into the development and application of advanced neural network architectures, such as Siamese networks, which are particularly suited for tasks that require nuanced similarity evaluations.

3.2 Bibliographic search

A thorough literature review has been conducted for this work, covering relevant domains comprehensively, using a wide range of search terms and filtering criteria with the intention of providing representation to diverse perspectives and approaches with regards to the aesthetic evaluation of breast cancer reconstructive surgery, deep learning techniques and architectures, and weakly supervised learning methodologies. Specific details about the search process, source, keywords, and filtering criteria can be seen in the sections below.

3.2.1 Sources and Databases Searched

The bibliographic search was carried out over a set of academically and medically relevant databases, which host large collections of peer-reviewed articles, conference papers, and technical reports. Examples include:

- **PubMed:** For studies related to medical imaging, clinical outcomes, and patient evaluations.

- **IEEE Xplore, ArXiv, and ACM digital library:** Research on deep learning, machine learning, and their applications in medical imaging.
- **Google Scholar:** Supplemental articles and grey literature for comprehensiveness.

3.3 Keywords and Search Queries

Search strings were formulated to target specific domains and their intersections. Boolean operators and domain-specific keywords ensured precision in retrieving relevant articles. Below are the primary queries categorised by domain, though more were used for stricter search:

Medical and Aesthetic Evaluation

("breast cancer treatment" OR "post-mastectomy" OR "breast reconstruction") AND ("aesthetic evaluation" OR "cosmetic outcome" OR "aesthetic assessment") AND ("patient satisfaction" OR "clinician assessment" OR "treatment results")

Objective and Image-based Assessments

("objective evaluation" OR "automated assessment" OR "image-based evaluation") AND ("breast symmetry" OR "breast asymmetry" OR "breast reconstruction") AND ("aesthetic outcomes" OR "cosmetic outcomes" OR "visual outcomes")

AI and Machine Learning Techniques

("deep learning" OR "machine learning" OR "AI" OR "artificial intelligence") AND ("breast cancer" OR "mastectomy" OR "breast reconstruction") AND ("aesthetic evaluation" OR "image classification" OR "ordinal classification")

Weak Supervision

("weakly supervised learning") AND ("imaging" OR "image classification" OR "aesthetic evaluation")

3.4 Filtering Criteria

A set of filtering criteria allowed for reducing the amount of found works into a realistic value, as well as assuring that the included references were of quality and relevance.

Inclusion Criteria

- Articles published in peer-reviewed journals or conferences within the last 6 years. This constraint was ignored for ground-breaking introductory references.

- Grey literature was included for cutting-edge topics.
- All work related to the evaluation of post-operative aesthetic classification was kept.
- The final collection was made through manual selection by full-text reading.

3.5 Literature Review: Aesthetic Evaluation in Reconstructive Surgery

The assessment of aesthetic results related to the treatment of breast cancer has become an integral part of modern oncological practice and reflects the progress in the efficacy of treatment and improvement in survival statistics. As breast cancer is one of the most common malignancies in women and given the increasingly higher survival rate seen in recent years, improvement in prognosis has shifted attention to quality of life after treatment, with emphasis on both physical and psychological aspects [8].

The indications for breast-conserving treatments have become standard practice for treating localised tumours, offering comparable outcomes to more radical surgeries while preserving as much of the appearance of the breast as possible. However, the aesthetic outcomes of these procedures vary widely due to factors such as tumour location, surgical techniques, and patient-specific anatomical differences. This variability reinforces the importance of consistent and reliable methods for aesthetic evaluation, benefiting both patients and healthcare providers.

Traditionally, aesthetic results have been evaluated subjectively based on patient feedback and expert panel assessments. While pleasing the patient is the goal, subjective assessment is influenced by individual bias, cultural norms, and observer experience. This variability challenges reproducibility and comparability across studies and institutions.

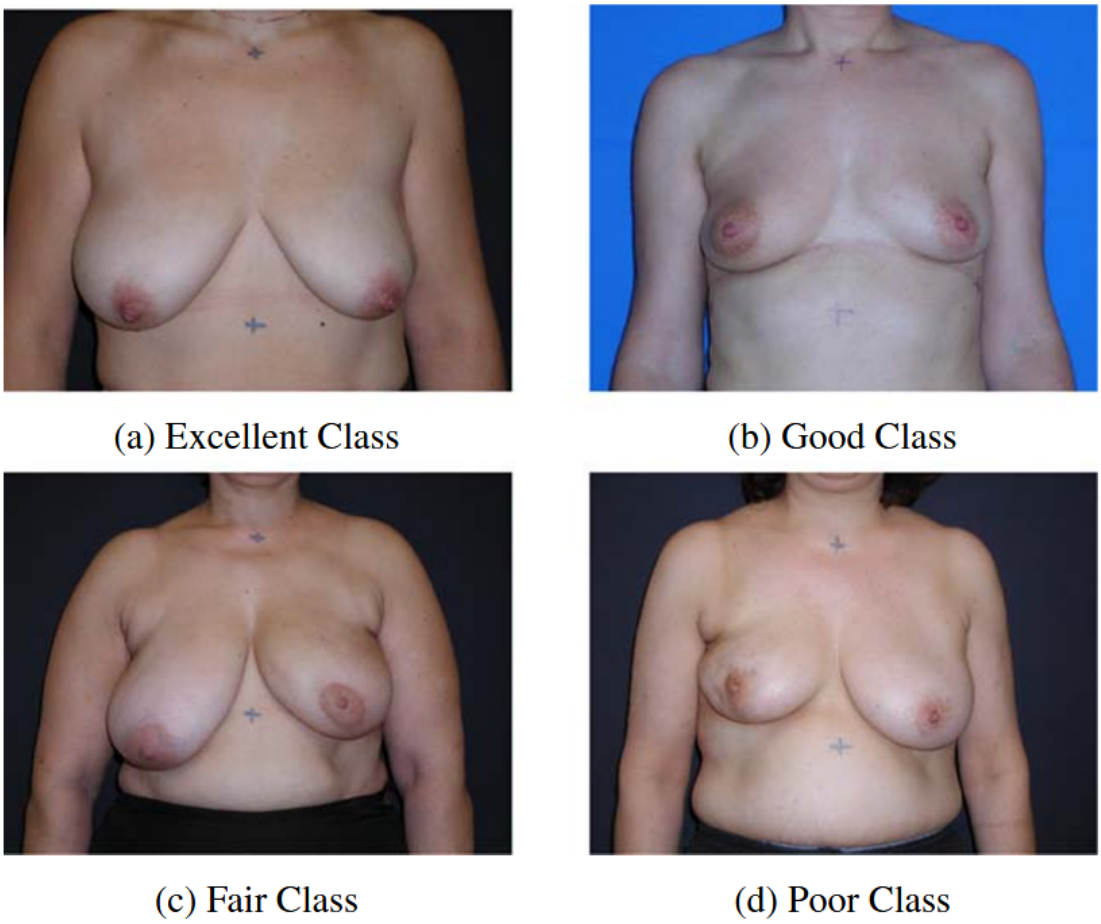
Due to these limitations, objective methods of assessment have been proposed based on measurable parameters: symmetry of breasts, their contour, colour, and skin quality. The purpose of all these strategies is to produce a standardised system of aesthetic analysis and hence allow for comparative studies and continued betterment in clinical practice.

While much progress has been made, the research is still held back by a lack of globally agreed-upon gold-standard for aesthetic assessment. More recent technological development in imaging and artificial intelligence holds a lot of promise to improve the accuracy and consistency of aesthetic assessment.

3.5.1 Subjective Aesthetic Evaluation

Historically, the most commonly used method for the evaluation of aesthetic outcomes after breast cancer treatment is subjective assessment. It is most often based on visual inspection by patients, health professionals, or expert panels assessing the treated and untreated breasts. One of the major frameworks for subjective assessment is the Harvard Scale, which was introduced by Harris et al. (1979) [24]. The scale classifies results into four categories: Excellent (breasts that are almost identical), Good (breasts that show minimal variations), Fair (breasts that are clearly different

but not grossly distorted), and Poor (breasts that are significantly distorted). Figure 3.1 shows examples of each class. Its simplicity and intuitive classification methodology make it one of the most widely used scales in clinical settings.



(a) Excellent Class (b) Good Class

(c) Fair Class (d) Poor Class

Figure 3.1: Harvard scale examples. [20]

The advantage of subjective assessment is that it focuses on each patient as a unique individual, mostly in dimensions such as satisfaction and self-perception of body image. Conducted by healthcare professionals or expert panels, subjective assessments can also apprehend insights that may be too subtle to be captured by strict quantitative measures.

Despite its utility, subjective evaluation brings many challenges. One major limitation of the technique is the dependence on the previous experience of the observer, cultural views, and inherent biases, resulting in great variability. Even the Harvard Scale itself has been criticised for its simplistic approach, since the results may significantly vary between observers or organisations. Besides, the individual characteristics of a patient, such as age, expectations, and emotional status, often affect subjective assessments, making them rather hard to compare. These constraints highlight the necessity for more impartial methodologies in the assessment of aesthetics.

To account for the inherent variability of subjective methods, objective assessment tools have

been developed and emphasise quantifiable parameters when evaluating outcomes. These normally include the geometric measurements — breast symmetry, contour, colour, texture, and volume differences — derived from photographs. Most objective evaluations also make use of the Harvard Scale where metrics are used to score outcomes into the same four categories. Objective assessment has the major advantages of standardisation and reproducibility. Measurements, such as the Breast Retraction Assessment (BRA), which quantifies nipple displacement [41], and other asymmetry indices, allow for the generation of consistent and comparable results.

Objective methods, however, are not free of challenges. Many of the solutions found before the use of Deep Neural Networks (DNN) require very careful marking and calibration, which can be both resource-intensive and time-consuming, as well as introduce variability. There is also the risk of focusing solely on tangible characteristics to the detriment of broader psychological and cultural aspects of aesthetic outcomes. Besides, changing imaging conditions—for example, variations in lighting and position—can affect the accuracy of such ratings. Finally, the absence of a commonly agreed-upon gold standard is one of the hindrances in this field, as different institutes and research efforts can opt for different methods, making research in this domain more difficult.

3.5.2 Objective Aesthetic Evaluation

Objective Features

Several objective features have been developed to assess aesthetic outcomes, with a particular focus on breast symmetry and proportion. The first prominent metric introduced by Pezner et al. (1985) [41] was the BRA, which calculates the displacement of the nipple to quantify asymmetry. Other features introduced by Cardoso et al. (2007) [5] include:

- Lower Breast Contour (LBC): Measures the difference in the levels of the lower contours of the breasts as shown on the left side of figure 3.2.
- Upward Nipple Retraction (UNR): Evaluates the vertical displacement between the nipples.
- Breast Compliance Evaluation (BCE): Assesses the difference in distances from the inframammary fold to the nipple for both breasts.
- Breast Contour Difference (BCD): Compares the lengths of the breast contours.
- Breast Area Difference (BAD): Quantifies discrepancies in the areas of the treated and untreated breasts.
- Breast Overlap Difference (BOD): Measures the non-overlapping area when one breast is flipped and aligned with the other for comparison, as shown on the right side of figure 3.2.

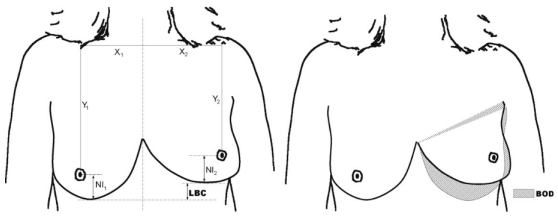


Figure 3.2: Illustration showing measurements for LBC and BOD. [5]

These features can be dimensionless and normalised relative to anatomical landmarks such as the sternal notch, which simplifies measurement by reducing the need for precise scale calibration. Dimensionless indices also improve robustness against variations in patient positioning or imaging conditions [5].

Objective Software

To overcome the shortcomings of subjective assessment, attempts at objective and automatic systems for evaluating breast cosmesis have been made, leading to the development of the tools described below.

- BAT: The Breast Analysing Tool (BAT) [17] is used to measure breast symmetry index (BSI) by calculating the difference in size and shape of both breasts. Although this study demonstrated positive results, the tool is only capable of binary classification (good or bad) and requires extensive manual annotation of markers of interest such as the nipples and the circumference of the breasts. In addition, only breast size and shape are used for the evaluation, leaving out valuable attributes such as colouration and scar visibility.

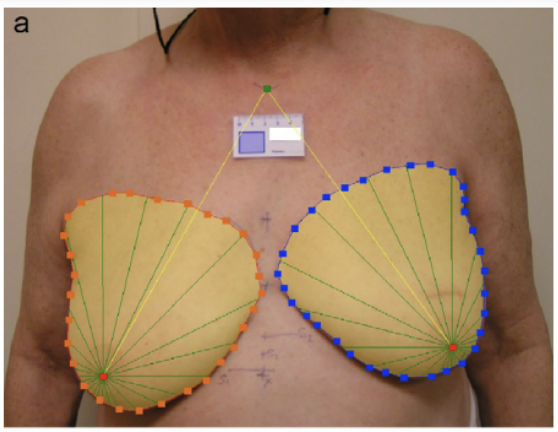


Figure 3.3: Digital picture of a patient in frontal view, which has been imported into the BAT software. The pictures show the markings of the breast to mathematically analyse the frontal BSI. [17]

- BCCT.core: Another tool that attempts to increase objectivity in post-operative breast surgery is the breast cancer conservative treatment cosmetic results (BCCT.core) developed in 2007

[5, 9]. As opposed to BAT, this software takes into account a diverse array of attributes, including symmetry, scar visibility, and colour differences, and rates the images according to the Harris scoring system using a support vector machine (SVM) to make the decision. This effort resulted in a performance comparable to that of a panel of experts. Still, the tool requires manual annotation, which may make it impractical as well as introduce variability in the system.

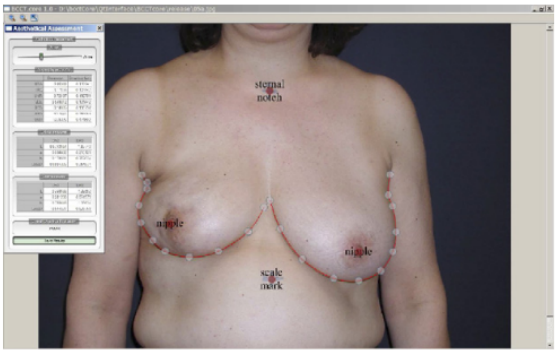


Figure 3.4: Graphical user interface of the BCCT.core software. [5]

- kOBCS: The kOBCS software, introduced more recently in 2020, is designed to facilitate the Objective Breast Cosmesis Scale (OBCS) [55]. The OBCS calculates the geometric asymmetry between treated and untreated breasts following breast-conserving therapy. It uses measurements derived from two virtual triangles formed by key anatomical landmarks - suprasternal notch, both nipples, and the intersection points of the lateral and lower breast contours bilaterally - as shown in figure 3.5. While this software has shown good results, there are some downsides to this software, including only using breast shape and size as discriminatory features, and requiring manual annotation, although minimal. The software is also unable to produce valid results in cases where both breasts have undergone surgery, and can only perform binary classification. A free tool based on this one was later introduced as Breast Aesthetic Scale Calculator (BAS-Calc) by Monton et al. (2021) [39].

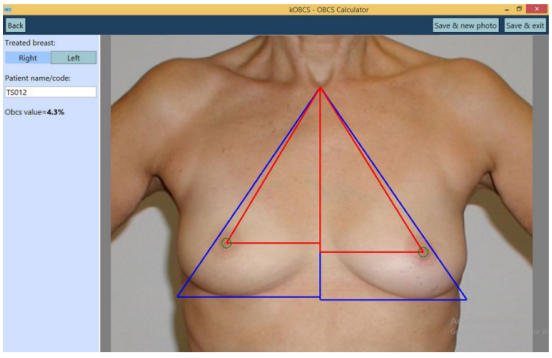


Figure 3.5: The kOBCS program with the different measured dimensions and the angle points at SSN, nipples, and the point of intersection between the projection of the lateral and the lower breast contours bilaterally. [55]

3.5.3 Feature Extraction

Feature extraction is an important phase in the objective and automatic assessment of the aesthetic outcome of breast surgery, forming the foundation for reproducible metrics. With feature extraction, specific anatomical and geometric attributes can be identified and measured for each breast, enabling standardised comparison across the treated and untreated ones. Below, key techniques found in the literature and used in feature extraction are explored.

Nipple Detection

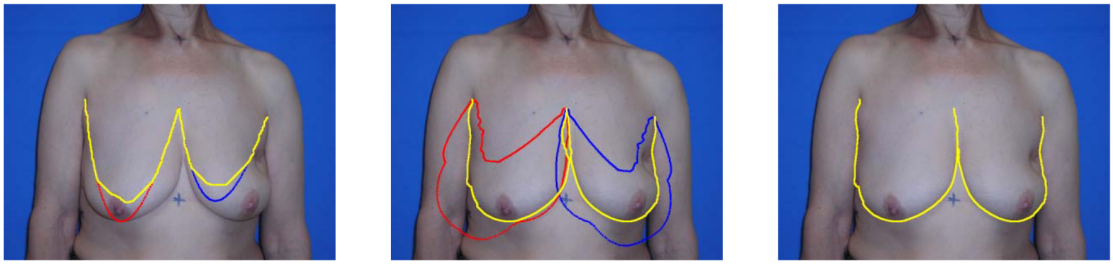
Cardoso et al. (2015) [7] introduced a structured approach to nipple detection. They developed a method that restricts the search for the nipple to the detected areola region, substantially narrowing the detection area and improving accuracy. This technique involves training a Support Vector Machine using features such as contour shape, diameter, and corner intensity, to find the most likely pair of nipple and aureola candidates.

Contour Detection

The delineation of breast contours plays a critical role in evaluating shape and symmetry. The first method for breast contour detection was introduced by Cardoso and Cardoso in 2007 [6] and relied on graph-based shortest path calculations, where the contour was identified between manually marked endpoints. This technique, while reliable, required manual input, which limited its efficiency and scalability.

The previous solution was later improved with the introduction of shape priors, which incorporate anatomical expectations into the detection process [57]. For instance, probabilistic models, including parabolic or elliptical priors, guide the identification of breast contours, reducing variability and improving accuracy.

Recently, another approach based on deep learning was introduced [18]. RINDNet is an edge detection network, which is capable of automatically identifying edge types specific to breast contours while suppressing other irrelevant background edges. However, the network suffers from sensitivity to overfitting and poor generalisation on diverse datasets due to a lack of training data. As an alternative, SobelU-Net integrates a traditional Sobel operator inside U-Net's convolutional architecture. The hybrid model efficiently combines both Sobel edge maps and raw image features through convolutional layers, producing refined edge maps in the decoding process. SobelU-Net outperforms traditional methods and RINDNet with better robustness and generalisation under different imaging conditions.



(a) Parabola prior. (b) Mask prior. (c) Unimodal prior.

Figure 3.6: Results of contour detection with different prior models. [57]

Feature Extraction with Deep Learning

Recent artificial intelligence advancements, particularly deep learning, have made it possible to automatically extract features from images without any need for manual intervention; however, its application for post-operative breast aesthetic assessment is still largely underexplored. Silva et al. (2019) [52] proposed a deep learning model specifically designed to detect key anatomical points crucial for the aesthetic evaluation of breast cancer surgery outcomes. This model integrates the detection of endpoints, nipples, and contours into a unified framework, aiming to improve both accuracy and reproducibility in aesthetic assessments. The model introduces a dual-module architecture using convolutional neural networks for keypoint detection in breast images. The first module focuses on heatmap regression and refinement, using a modified U-Net architecture to generate fuzzy heatmaps indicating the probable locations of keypoints. This intermediate representation helps by regularising the network, reducing overfitting due to small datasets, and improving accuracy. The second module performs keypoint regression, combining the refined heatmaps with the original image to predict the precise coordinates of points of interest. This module incorporates VGG16 pre-trained on ImageNet, followed by additional convolutional and dense layers to refine the predictions. A hybrid method was also introduced, which utilised endpoints detected by the deep learning model while employing traditional shortest-path algorithms for contour delineation. The hybrid method outperformed both the deep learning model and traditional approaches in breast contour detection.

3.5.4 AI for objective evaluation

With the introduction of deep learning, many improvements have been made to postoperative breast aesthetic evaluation, offering automated, objective, and reproducible solutions that address the limitations of both subjective assessments and traditional semi-automated objective systems. Recent AI developments have introduced fully automatic frameworks that integrate feature detection, classification, and even content-based retrieval of patient cases to optimise the evaluation process.

Guo et al. (2022) [22] developed a fully automatic framework, shown in figure 3.7, that eliminates the need for manual input while assessing cosmetic outcomes of breast-conserving

therapy. The framework begins with a YOLO-based deep learning detector to segment the breast and locate key features, including the areola and nipple. The identified points of interest are used to generate a feature vector with 298 dimensions, which encapsulate key metrics such as symmetry, nipple placement, scar visibility, and skin colour. These features are analysed to compute an aesthetic score based on the Harvard scale. This approach significantly improves the accuracy and reliability of the evaluation, addressing limitations of earlier systems like BCCT.core, which required manual point marking. Guo et al.'s framework demonstrates superior performance in predicting cosmesis scores, achieving state-of-the-art accuracy while ensuring full automation.

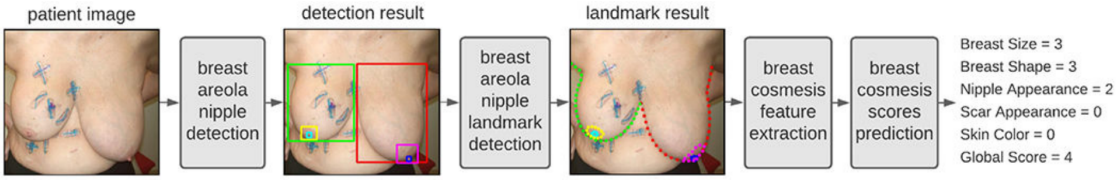


Figure 3.7: Overview of the proposed Breast Cosmesis evaluation system. [22]

Silva et al. (2022) [51] proposed a deep neural network for binary classification of aesthetic outcomes and retrieval of similar cases from a database. This model improves upon traditional systems by introducing intermediate supervision during training, which optimises the detection of keypoints such as nipples and breast contours. These keypoints are used to compute asymmetry measures, including BRA, LBC, and UNR. The model combines these measures with high-level features extracted from convolutional layers to produce a robust and interpretable aesthetic classification. One of the unique aspects of Silva et al.'s approach is the integration of content-based image retrieval. Using features extracted in the high-semantic space before classification, the network retrieves the top three most visually similar cases from a dataset of past patients as shown in figure 3.8. This functionality not only aids in assessment but also helps manage patient expectations by presenting realistic postoperative outcomes from patients with similar anatomical and clinical characteristics.

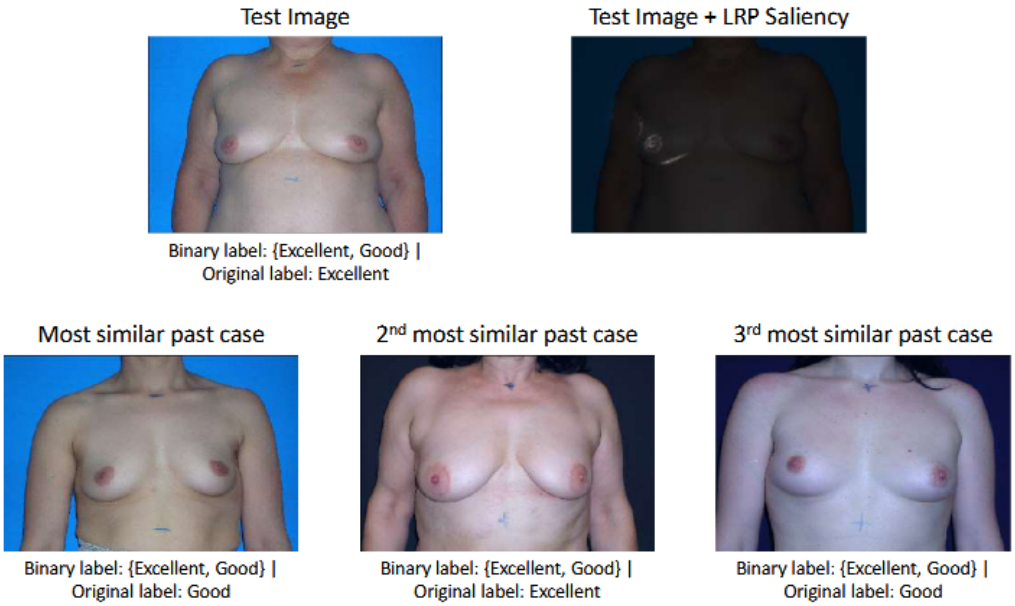


Figure 3.8: Example of query and retrieved images for an Excellent aesthetic outcome. Binary label means class belongs to set Excellent, Good. Original label is the ordinal label previous to binarisation (Excellent, Good, Fair, or Poor). LRP saliency map is also shown for the test image. [51]

3.5.5 Summary

The aesthetic assessment of breast cancer treatment has evolved significantly with the advancement of surgical techniques and technology. Early on, there were mainly subjective assessments by use of guidelines such as the Harvard Scale. These methods were patient-centred, though not reproducible due to observer variability and biases. Objective evaluations introduced measurable features related to symmetry, contour, and skin quality, with metrics like Breast Retraction Assessment and others. tools like BAT, BCCT.core, and kOBCS are more consistent, but still rely heavily on manual annotation and are therefore resource-intensive and less scalable.

More recently, breakthroughs in AI and, more specifically, deep learning have revolutionised this field. Fully automated systems, such as that by Guo et al., use AI to analyse features, symmetry, and scars for state-of-the-art accuracy without the intervention of a human operator. Equally, Silva et al.'s deep neural network coupled keypoint detection with a content-based image retrieval strategy to improve classification and manage patient expectations regarding treatment outcomes.

Challenges still persist in the form of limited data, which hinders the training process of deep learning approaches. The integration of AI into this field has also not been fully explored, with architectures such as siamese networks offering great potential due to their inherent capacity for ranking.

3.6 Literature Review: Siamese Networks for Image Ranking

Siamese networks are a special class of neural networks designed especially to compare and distinguish relationships between pairs of inputs. Originally introduced by Bromley et al. (1993) [2] for signature verification, Siamese networks usually adopt a two-branch configuration in which the two branches are identical with respect to sharing the same weights and structure properties. This ensures that the network always extracts the same features from input pairs. These branches then output their results, which are combined using some similarity metric like Euclidean distance or cosine similarity, measuring how similar or dissimilar the inputs are.

The main strength of Siamese networks is their ability to learn from paired examples rather than from isolated cases. Such a pairwise learning mechanism has been pivotal for tasks like signature verification, face recognition, and image matching [2, 12, 37, 49]. By constructing a feature embedding space wherein the similar inputs are located close to each other while the dissimilar inputs are spread out helps Siamese networks achieve superior performance in contexts where ranking or comparative evaluations are required.

The training process of Siamese networks typically relies on a contrastive loss function, which adjusts the embeddings based on the similarity or dissimilarity of input pairs. For similar pairs, the network minimises the distance between their embeddings, while for dissimilar pairs, it ensures that the distance exceeds a predefined margin [23]. The contrastive loss function is defined in section 2.5.2. This approach allows the network to develop a robust representation of the features present in the data that best discriminate different classes.

Siamese networks have proven to be a powerful tool in image ranking, allowing for detailed comparisons and order classifications [10, 35]. The pairwise nature of the training process fits well with tasks that rank images based on certain qualities, such as their aesthetic qualities, their medical importance, or how similar they are to a reference image. This makes them an ideal candidate for exploration in areas like aesthetic assessment in breast reconstructive surgery, where assessment through comparison may provide a more objective result when compared to direct classification.

The following sections will explore the progress in Siamese network methods, how they are used in ranking and aesthetic assessment, and how they work with semi-supervised learning techniques to solve problems related to limited data.

3.6.1 Advancements in Siamese Networks

Since their inception, crucial improvements have been made to Siamese networks, solving difficult problems and enabling them to be used in different areas. The main focus of these improvements is on changes in design and ways of learning.

Architectural Innovations

Architectural refinements in Siamese networks aim to improve their capacity for feature extraction and similarity assessment. One notable variation is the pseudo-Siamese network, which differs from the traditional Siamese model by allowing its branches to have independent weights. This flexibility is particularly advantageous for processing inputs with distinct characteristics, such as cross-modal data. For example, pseudo-Siamese networks are often used in scenarios where inputs like text and images need to be compared, requiring feature extraction methods tailored to each modality [14, 15].

Another significant and well-known extension is the triplet network, which introduces a third branch into the architecture. This branch facilitates the simultaneous comparison of an anchor input, a positive sample, and a negative sample. The triplet network is particularly effective in ranking tasks because it ensures that the anchor is closer to the positive sample than to the negative sample in the embedding space. Hoffer and Ailon (2015) [26] demonstrated the effectiveness of this architecture in metric learning, showcasing its ability to improve the discrimination of embeddings in applications such as facial recognition and image retrieval.

Hybrid architectures are yet another area of innovation, combining Siamese networks with other deep learning models to enhance their applicability. For instance, convolutional Siamese networks have been used in visual tracking, where the spatial features of inputs are vital. Bertinetto et al. (2016) [1] proposed a fully convolutional Siamese network that utilises convolutional layers to process visual data efficiently, allowing the network to excel in tasks requiring high spatial resolution and rapid processing, such as real-time object tracking. Similarly, recurrent Siamese networks have been developed to handle sequential data, such as video frames or time-series information, making them suitable for applications in video analysis and speech processing. For example, Roy et al. (2019) [46] propose a DNN for detecting collision-prone vehicle behaviour at intersections, which leverages a Siamese Interaction Long Short-Term Memory network (SILSTM) to capture temporal dependencies in sequential data.

Learning Mechanisms

Innovations in training methodologies have significantly improved the efficiency and accuracy of Siamese networks. Besides contrastive loss, many other options have been introduced in recent literature.

Triplet loss, popularised by Schroff et al. (2015) [49], extends this concept by comparing an anchor sample with both a similar (positive) and a dissimilar (negative) example. This loss function, used in the previously mentioned triplet networks, encourages the model to learn embeddings that respect relative similarity within the data, making it highly effective for tasks like facial recognition and image retrieval.

Ranking loss functions focus on preserving the relative order of samples in the embedding space. Lin et al. (2019) [35] demonstrated the value of ranking loss in aesthetic evaluations,

showing how this type of loss can be used to teach the network the order of the data according to its features in regression tasks.

Self-supervised learning techniques have further broadened the applicability of Siamese networks. Approaches like SimCLR [11] and BYOL [21], shown in figure 3.9, leverage pretext tasks such as instance discrimination to train on unlabeled data. These methods have significantly reduced the reliance on annotated datasets, enabling the extraction of robust feature representations from large-scale unlabeled data.

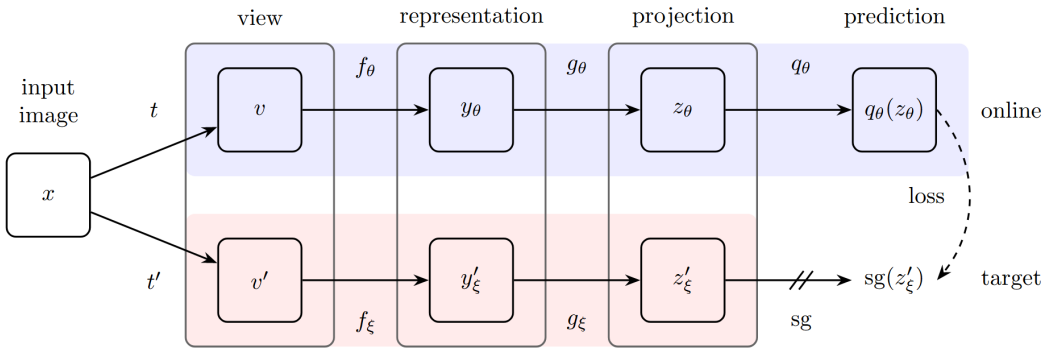


Figure 3.9: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $sg(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ, and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation [21].

3.6.2 Aesthetic Ranking

The inherent ability of Siamese networks to compare inputs has made them particularly well-suited for ranking tasks, where the objective is to predict or classify based on a defined order among classes. This characteristic can be exploited in domains such as beauty prediction and aesthetic assessment, under the assumption that in such subjective tasks, the ordering between data points tends to be more objective.

One notable study is the work by Chandakkar et al. (2017) [10], who developed a Siamese network to rank image aesthetics using a relative ranking loss function. Their model analysed pairs of images to decide which one was more visually appealing. By focusing on the relative differences between images, the network avoided the pitfalls of traditional binary classification systems, which only categorise images as beautiful or not. This approach allowed the model to pick up on subtle relationships between images, achieving greater accuracy and improved nuance.

Lin et al. (2018) [34] expanded upon these ideas in the domain of facial beauty prediction through the development of the R^2-ResNeXt architecture as shown in figure 3.10. This model integrated a regression subnetwork for absolute beauty score prediction with a ranking subnetwork for relative comparisons. By combining a regression loss and a pairwise ranking loss into an aggregated loss function, the network was able to learn both absolute and relative aesthetic

representations. This approach enhanced the robustness of the learned embeddings, resulting in state-of-the-art performance on benchmark datasets.

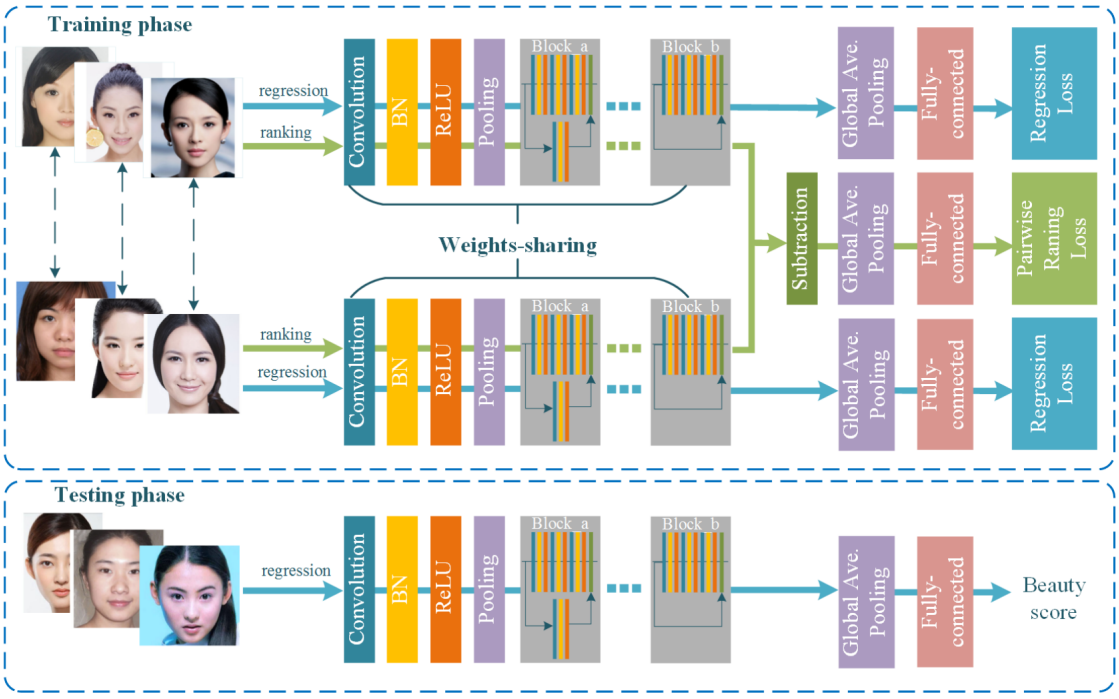


Figure 3.10: The architecture of proposed R2-ResNeXt, where the parameters of feature extractors are shared regression subnetworks (shown in blue arrow) and two-branches ranking subnetwork (shown in green arrow) [34]

Using relative ranking offers several clear advantages in aesthetic evaluations. First, it mirrors how humans often judge beauty—through comparisons rather than absolute ratings. Second, it acts as a regularising force, helping the network generalise better by learning features that highlight differences between data points. Finally, it is robust to limited data, as pairwise comparisons generate more training samples and provide richer learning opportunities than isolated annotations.

3.6.3 Semi-Supervised Learning with Siamese Networks

Semi-supervised learning bridges the gap between supervised and unsupervised learning by utilising both labelled and unlabeled data during training. Siamese networks are inherently well-suited for SSL tasks due to their reliance on pairwise comparisons, which can be extended to unlabeled data using techniques such as pseudo-labelling and self-supervised pretraining [11].

In semi-supervised settings, labelled data is often used to train the initial embedding space, while unlabeled data is integrated into the training process through strategies such as contrastive learning. For example, Sahito et al. (2019) [48] propose a self-training network for semi-supervised learning where initially, the network is trained on labelled data to generate an initial embedding space. Then, using k-nearest neighbours, pseudo-labels are generated for unlabeled embeddings.

Finally, unlabeled instances are selected based on a confidence measure and added with their pseudo-labels for further training.

The application of semi-supervised learning in aesthetic evaluations, particularly in medical imaging, is sometimes necessary, as labelled datasets in this domain are often scarce and expensive to annotate. By leveraging the more abundant unlabeled post-operative images, semi-supervised learning approaches have the potential to heavily improve the training of Siamese networks capable of providing better aesthetic assessments.

3.6.4 Summary

Siamese networks have emerged as an effective tool in tasks requiring comparative evaluations, including image ranking and aesthetic assessments. Their ability to learn from paired examples using shared weights ensures consistent and robust feature extraction. From their foundational application of signature verification to more recent domains such as facial beauty prediction, these networks have demonstrated remarkable adaptability.

Key advancements, such as the introduction of pseudo-Siamese networks, triplet networks, and hybrid architectures, have expanded their scope to handle diverse data modalities and complex relationships. Learning mechanisms, including contrastive loss, triplet loss, and ranking loss, further enhance their capabilities, allowing for effective embedding of high-dimensional data for specific tasks.

In aesthetic ranking, studies like those by Chandakkar et al. (2017) [10] and Lin et al. (2018) [34] have paved the way for using Siamese networks to evaluate subjective qualities objectively.

As research progresses, Siamese networks show promise for more general applications, including medical imaging and semi-supervised learning. Their inherent design, combined with constant innovations, ensures their relevance in solving some of the most challenging problems in machine learning and artificial intelligence.

Chapter 4

Preliminary Experiments

Given a post-surgical image of a patient's breasts, we aim to develop a model that automatically classifies its aesthetic outcome into one of four classes: *Poor*, *Fair*, *Good*, or *Excellent*, according to the Harvard scale. To achieve this goal, certain prerequisites must be met.

First, understanding the available data is a necessity for any data-driven task. This allows us to understand the limitations of the available information and gives insight into how to overcome them if possible. Sections 4.1 and 4.2 delineate the data analysis and preprocessing process.

After the data is analysed and preprocessed, we develop a baseline with which to compare future experiments.

4.1 Dataset Analysis

The dataset used in this work is a private dataset obtained from the Portuguese Champalimaud Foundation in the context of the CINDERELLA project [28]. This dataset is composed of over 5,000 breast images, including pre- and postsurgery, various geometric keypoints and metrics, as well as different views of the breasts. A subset of the data also contains subjective aesthetic labels ranging the entire Harvard scale, assigned by three oncoplastic surgeons, and objective labels extracted using objective metrics calculated through the aforementioned keypoints with the BCCT.core software. For this research, our interest lies in post-surgery images with subjective ratings, which lowers the amount of data available for supervised training to 1,406. It is important to note that within these samples, there are cases where different experts rate the same image, and these ratings may not match. These samples also reveal disparities in class representations, where extreme classes (*Excellent* and especially *Poor*) are less represented, as well as differing data distribution between experts as seen in figures 4.1 and 4.2.

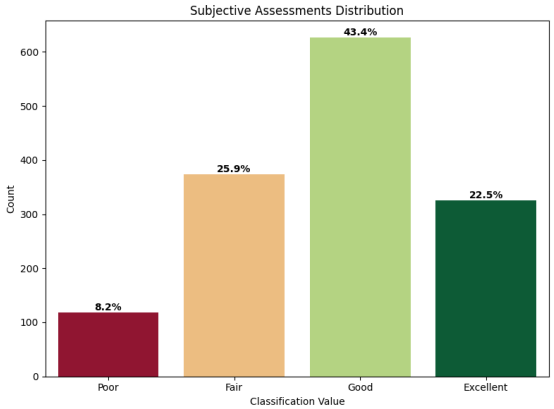


Figure 4.1: Overall class distribution for subjective ratings in the CINDERELLA dataset.

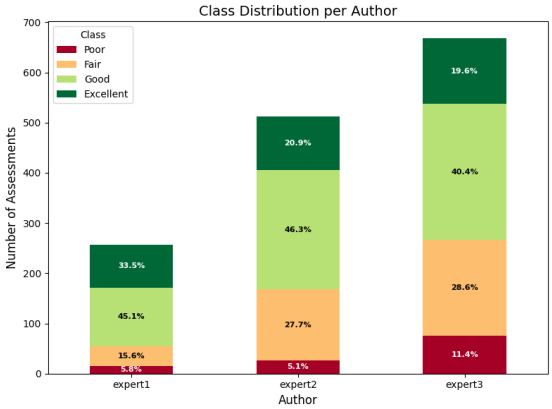


Figure 4.2: Class distribution per author for subjective ratings in the CINDERELLA dataset.

4.2 Data Preprocessing

The available data was subjected to an initial preprocessing pipeline that addresses the most immediate issues.

4.2.1 Duplicate Data

While the dataset was fairly clean, there were some rare cases where the same author would rate the same image differently at different time points. After discussing with the experts responsible for the labelling process, it was decided that only the latest label assigned to each image from each author is to be considered.

4.2.2 Data Balancing

Ideally, the model should be trained on balanced data in order to achieve good performance even in extreme classes, regardless of the real data distribution. To achieve this, weighted sampling was used. This technique assigns a weight to each data sample that is inversely proportional to the label's representation in the data. This allows oversampling of minority classes, balancing the data distribution for training.

4.2.3 Data Augmentation

To improve data variety, a series of transformations is applied to each image when it is sampled.

- Horizontal Flip: 50% chance of flipping the image horizontally
- Colour Jitter: A slight jitter is applied to the image's contrast, saturation, brightness, and hue
- Rotation: The image is rotated by up to 15º.

These transformations are then followed by normalisation. Later on more involved techniques were used in an attempt to further improve data variety and availability as discussed in section 5.4.

Across all experiments the data was divided into a 70% training split, 15% validation split, and 15% test split.

4.3 Baseline Architecture

To compare future architectures, baselines are required. To this end, a simple ResNet-50 network trained for regression was developed. We also use the classic BCCT.core evaluation results for comparison with the experiments.

4.3.1 ResNet-50

ResNet is a deep convolutional neural network (CNN) architecture, known for its use of residual blocks, which address the vanishing gradient problem, allowing for very deep networks. Using a pre-trained ResNet-50 model and adding a linear regression head, we can fine-tune it for our particular task of breast cancer treatment aesthetic evaluation. To train this model, we use data augmentation and split it as described in 4.2.3. We then set the learning rate to $1e^{-4}$, use a batch size of 32, and let it run for 100 epochs with early stopping. Table 4.1 and figure 4.4 show the resulting performance on the test set. Ranking accuracy is defined as the fraction of image pairs where the model's relative ranking matches the ground truth relative ranking for all image pairs with differing ground truth labels. The model achieves a relatively high ranking accuracy, however it is unable to consistently differentiate between the top 3 classes as demonstrated by the scattering on the bottom-right of the confusion matrix. This model severely misclassified (difference between predicted and true class ≥ 2) 12 data points.

Table 4.1: ResNet50 performance on the test set.

Model	Accuracy ↑	Ranking Accuracy ↑	MAE ↓	Macro-MAE ↓	RMSE ↓
ResNet50	48.34%	76.35%	0.6246	0.6029	0.7993

4.3.2 BCCT.core

BCCT.core [9] is a software introduced by Cardoso et. al (2007), which requires manual annotation and uses an SVM to objectively rate post-cancer treatment breast images according to features such as symmetry and colour. It was found that the algorithm used achieved a performance comparable to that of a group of experts.

As the dataset used includes objective labels obtained through this software, we can calculate relevant metrics for our test split. The results of this evaluation are presented in table 4.2 and in figure 4.3. As seen the algorithm performs poorly on the extreme classes, overly focusing on the middle classes, especially the prevalent *Good* class, obtaining relatively good accuracy but a fairly low ranking accuracy and high Macro-MAE and RMSE.

The model also severely misclassified 6 data points.

Table 4.2: BCCT.core performance on the test set.

Model	Accuracy ↑	Ranking Accuracy ↑	MAE ↓	Macro-MAE ↓	RMSE ↓
BCCT.core	57.21%	56.49%	0.4558	0.6057	0.7153

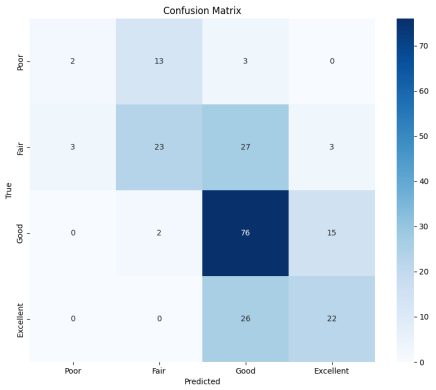


Figure 4.3: Confusion Matrix for BCCT.core on the test split

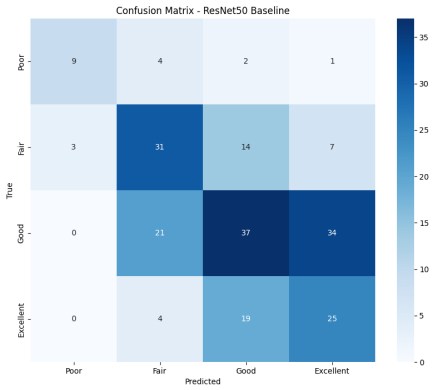


Figure 4.4: Confusion Matrix for ResNet50 on the test split

Chapter 5

Methodology

This chapter introduces the various methods and techniques used to assign an ordinal class of *poor*, *fair*, *good*, or *excellent* to a post-operative breast image. We first introduce the architectures that inspired the final network. Then, we present the proposed Siamese OrdinalCLIP architecture. Next, we attempt to overcome the scarcity of data through various techniques. Finally, we tackle the problem of decision opacity prevalent in deep architectures by introducing explainability into the network.

5.1 Ranking Guided Aesthetic Classification

Research in facial beauty has focused extensively on the advantages that Siamese networks bring to aesthetic evaluation by enabling models to learn relative orderings in addition to absolute scores. When applied to breast-surgery outcomes, an analogous formulation allows the network to reason about subtle inter-patient differences that are difficult to capture with point predictions alone. The methodology presented in this section extends these ideas to ordinal breast-aesthetic assessment.

5.1.1 Architecture

Lin et al. (2018) [34] introduced *R2-ResNeXt*, a ResNeXt-50–based model that couples a regression head with a Siamese ranking head, trained jointly through an aggregated loss. Inspired by their design, the network proposed here adopts the same guiding principles while being adapted to the specific constraints of breast-aesthetic evaluation.

5.1.1.1 Overview

An input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, where H and W represent the image's height and width, respectively, is first processed by a ResNeXt-50 backbone that outputs a d-dimensional embedding $\mathbf{f}$. After the backbone, a jointly trained Multilayer Perceptron (MLP), and two complementary heads follow:

Shared embedding MLP A two-layer perceptron $\phi : \mathbb{R}^d \to \mathbb{R}^{128}$ produces a compact representation

$$\mathbf{e} = \phi(\mathbf{f}) = \mathrm{ReLU}\big(W_2\,\mathrm{Dropout}\big(\mathrm{ReLU}\big(W_1\mathbf{f}\big)\big)\big), \quad W_1 \in \mathbb{R}^{256\times d},\ W_2 \in \mathbb{R}^{128\times 256}.$$

All subsequent heads see exactly the same $\mathbf{e}$, so gradients from every loss term update the same MLP.

1. **Regression head** $\psi_{\mathrm{reg}} : \mathbb{R}^{128} \to \mathbb{R}$ is a *single* linear layer $w_{\mathrm{reg}} \in \mathbb{R}^{1\times 128}$ applied to $\mathbf{e}$. A squashing transformation rescales the raw output to the 4-level aesthetic interval:

$$\hat{y} = 0.5 + 4\,\sigma\big(\psi_{\mathrm{reg}}(\mathbf{f})\big) \in [0.5, 4.5] \tag{5.1}$$

   Where σ represents the sigmoid operation. The reason we rescale the output to $[0.5, 4.5]$ instead of $[1, 4]$ is simply to ensure all classes have an equal output space, if the rescale function were limited to $[1, 4]$ we would be effectively halving the amount of potential outputs that fall within the extreme classes. To allow the model to explore the extremes within the [0.5, 1] and [4,4.5] ranges, a random amount $r \in [0, 0.5]$ was subtracted from each *Poor* label, and added to each *Excellent* label.

2. **Ranking head** $\psi_{\mathrm{rank}} : \mathbb{R}^{256} \to \mathbb{R}$, implemented analogously but receiving the concatenation of two feature vectors, as opposed to the difference between the embeddings as proposed in the original paper. This allows the ranking head to receive more information. Given a pair $(\mathbf{f}_i, \mathbf{f}_j)$, the head predicts a signed score s_{ij} whose sign encodes the expected order.

5.1.1.2 Aggregated Loss

Let $\mathscr{B} = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^{N}$ denote a batch with N data points, ground-truth scores $y_i \in [0.5, 4.5]$ and author identifiers a_i. After a forward pass we obtain features $\mathbf{f}_i$ and predictions $\hat{y}_i$. Two complementary objectives are minimised jointly, namely mean squared error loss L_{reg} for the regression head, and hinge loss L_{rank} for the ranking head:

$$\mathscr{L}_{\mathrm{reg}} = \frac{1}{N}\sum_{i=1}^{N}\big(\hat{y}_i - y_i\big)^2, \tag{5.2}$$

$$\mathscr{L}_{\mathrm{rank}} = \frac{1}{|\mathscr{P}|}\sum_{(i,j)\in\mathscr{P}}\max\big(0,\, \alpha - \eta_{ij}\, s_{ij}\big), \tag{5.3}$$

where $\eta_{ij} = +1$ if $y_i > y_j$ and $\eta_{ij} = -1$ otherwise. Pairs are formed *within the same annotator* ($a_i = a_j$) to account for inter-surgeon bias, and only if $\lceil y_i \rceil \neq \lceil y_j \rceil$. $\mathscr{P}$ represents the total number of valid pairs in the batch $\mathscr{B}$. The margin α is scaled linearly with $|y_i - y_j|$, encouraging larger separations when ground-truth differences are pronounced, as opposed to the original paper, where a fixed margin is used.

The total loss is

$$\mathscr{L} = \mathscr{L}_{\text{reg}} + \lambda \mathscr{L}_{\text{rank}} \tag{5.4}$$

with $\lambda = 0.5$ throughout unless stated otherwise.

5.1.2 Experiments

Input Pipeline Data is split according to subsection 4.2. The images are resized to 224 x 224 and standardised per channel. On-the-fly augmentations include horizontal flips, random rotation, and colour jitter as explained in 4.2.3.

Pair construction Within each batch of 32 images, we sample all annotator-consistent, label-discordant pairs (i, j). These pairs are used for the ranking calculations.

Optimizer and schedule AdamW is used with an initial learning rate of 1×10^{-4}, weight decay 10^{-2}, and cosine annealing over $T_{\max} = 100$ epochs. The backbone stays frozen; only the heads and MLP are updated. Early stopping triggers if validation MAE fails to improve for 15 epochs.

5.2 Using Natural Language for Aesthetic Evaluation

OrdinalCLIP [33] utilizes the rich semantic structure of the CLIP [42] vision-language latent space to impose *ordinal* constraints via natural-language prompts. In this section we describe how the original framework was adapted to breast-surgery outcome assessment, summarise the training protocol used for our dataset, and present the resulting performance.

5.2.1 Architecture

CLIP trains an image encoder to map images into the ordinal space of a pre-trained text encoder by approximating the image's latent representation to that of the respective textual class. OrdinalCLIP inherits CLIP's dual-encoder design, but adds a *rank-prompt learner* that injects ordinal structure into the textual side of the model. Figure 5.1 gives a schematic view of this design as presented in the original work.

5.2.1.1 Language prototypes in CLIP space

OrdinalCLIP replaces the linear head with language prototypes obtained from the CLIP text encoder. Let $p_j \in \mathbb{R}^d$ be the ℓ_2-normalised prototype for rank r_j. Given ℓ_2-normalised image embeddings $I = \{I_i\}_{i=0}^{B-1}$, the cross-modal similarity score is

$$a_{i,j} = I_i \cdot p_j^{\top}, \quad 0 \le i < B-1, 0 \le j < C-1. \tag{5.5}$$

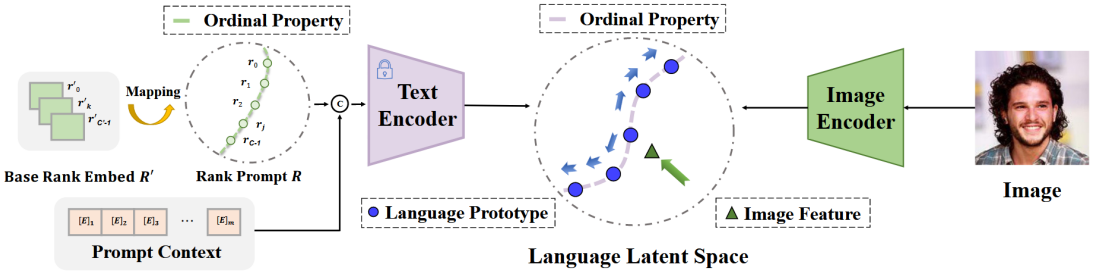


Figure 5.1: Diagram of the OrdinalCLIP framework. Image taken from the original paper [33]

Because both vectors are unit-normalised, $a_{i,j}$ represents their cosine similarity. Values close to $+1$ indicate a strong semantic match between image i and rank j, whereas values near -1 signal opposite semantics. The $B \times C$ matrix $A = [a_{i,j}]$ therefore gathers, for every image in the batch, its affinity to *all* ordinal prototypes. The network is optimised using image-to-text and text-to-image contrastive loss functions to minimise the distance between the image embeddings and the embeddings of the language prototypes. Before applying the loss function, the cross-modality similarity matrix must first be normalised, with $A' = [a'_{i,j}]$ and $A'' = [a''_{i,j}]$:

Image-to-text normalisation. Row-wise normalisation with temperature T gives

$$a'_{i,j} = \frac{\exp(a_{i,j}/T)}{\sum_{j=0}^{C-1} \exp(a_{i,j}/T)}. \tag{5.6}$$

Text-to-image normalisation. Column-wise normalisation yields

$$a''_{i,j} = \frac{\exp(a_{i,j}/T)}{\sum_{i=0}^{B-1} \exp(a_{i,j}/T)}. \tag{5.7}$$

The image-to-text normalisation turns each row of A into a categorical distribution that answers the question *which rank best describes this image?*". The text-to-image normalisation asks *which image best instantiates this rank?*". Encouraging consistency in *both* directions via the symmetric InfoNCE-style objective below has been shown to yield markedly stronger representations in CLIP-like models [42].

Let Y be the ground-truth, and let Y'' be Y with every non-zero column normalised to sum to 1. The bidirectional contrastive objective is

$$\mathscr{L} = \frac{1}{2}\left[\frac{1}{B}\sum_{i=0}^{B-1} \mathrm{KL}(Y_{i,\cdot}, A'_{i,\cdot}) + \frac{1}{C}\sum_{j=0}^{C-1} \mathrm{KL}(Y''_{\cdot,j}, A''_{\cdot,j})\right]. \tag{5.8}$$

Equations (5.5)–(5.8) follow the original derivation in [33].

5.2.1.2 Differentiable rank prompts

A prompt for rank r_j concatenates m learnable context tokens $[E]_0, \ldots, [E]_{m-1}$ with a learnable *rank embedding*r_j. Instead of treating each r_j independently, OrdinalCLIP interpolates from a small set of C' base embeddings $R' = [, r'_0, \ldots, r'_{C'-1}]$ $(C' \ll C)$:

$$r_j = \sum_{k=0}^{C'-1} w'_{j,k}, r'_k, \tag{5.9}$$

where the weights

$$w'_{j,k} = \frac{\mathscr{I}\left(|j - \frac{C-1}{C'-1}k|\right)}{\sum_{t=0}^{C'-1} \mathscr{I}\left(|j - \frac{C-1}{C'-1}k|\right)}, \qquad \mathscr{I}(u) = \begin{cases} 1 - \dfrac{u}{C-1}, & \text{(linear)} \\ \dfrac{1}{u+\varepsilon}, & \text{(inverse-proportion)} \end{cases} \tag{5.10}$$

explicitly enforce numerical continuity between neighbouring ranks. In this work $\mathscr{I}(u)$ follows the linear approach.

Motivation for interpolation. Learning a free vector for every rank risks scattering the prototypes arbitrarily on the embedding space and destroying their natural order. By rewriting each r_j as a combination of a handful of *base* embeddings (Eq.(5.9)), we constrain consecutive ranks to remain geometrically close and approximately collinear. This simple bias improves the "ordinality score" of the language prototypes, i.e. the fraction of prototype pairs whose cosine similarities respect the numerical ordering, as reported in the original paper [33].

During training, we optimise:

- the image encoder (initialised from ResNet-50),
- the context tokens $[E]_i$,
- the base rank embeddings R'.

5.2.1.3 Inference as nearest-prototype classification

After training, the heavy text encoder can be discarded. The model stores only the C normalised prototypes p_j. Given a new image x, the image encoder outputs an embedding $I(x) \in \mathbb{R}^d$, normalised to unit length. Prediction reduces to

$$\hat{r}(x) = \arg\max_j I(x) \cdot p_j^\top \tag{5.11}$$

i.e. selecting the prototype with highest cosine similarity. Consequently, OrdinalCLIP shares the *same* memory footprint and computational cost at deployment as a conventional linear classifier, while benefiting from the language priors distilled during training.

5.2.2 Experiments

Input pipeline Data is split according to subsection 4.2. Images are resized to 224×224; training augmentations comprise random resized crops and horizontal flips, followed by CLIP normalisation.

Optimiser RADAM is used with weight decay 0 and a learning rate of 1×10^{-4}, as suggested in the public OrdinalCLIP implementation.

Label ambiguity The Harvard Scale rates the aesthetic outcome of breast cancer treatment in four distinct classes: *poor*, *fair*, *good*, and *excellent*. While these class names were chosen with the delicate nature of the task in mind, to a language model the words used to represent the labels may hinder its performance if they are ambiguous or if their semantic meaning does not sufficiently differentiate the classes. The bottom two classes in the scale, *poor* and *fair*, are particularly problematic as they attenuate the meaning of the underlying class; as such, for the CLIP text encoder only, the labels for these two classes were changed to "*horrible*" and "*bad*", respectively.

Adaptation to breast-surgery assessment. For our four-class task, we instantiate C=4, and the prompt context is initialised as:

$$''\texttt{Aesthetic assessment: As a professional surgeon, I would rate the aesthetic outcome of this breast surgery as } \{r_j\}'' \quad (5.12)$$

.

5.3 Siamese OrdinalCLIP

In this section, we propose a new method for aesthetic ranking by joining the previously presented OrdinalCLIP and R2-ResNeXt.

5.3.1 Architecture

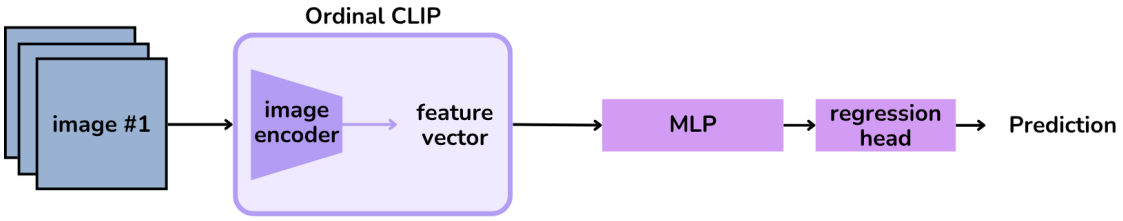


Figure 5.2: Diagram for Siamese OrdinalCLIP inference.

The guiding idea behind *Siamese OrdinalCLIP* is to apply the ranking–regression heads of the R^2-ResNeXt family onto an OrdinalCLIP image encoder that has already learned, through

natural–language prompts, that *poor* < *fair* < *good* < *excellent*. Figure 5.2 shows a diagram of the final architecture after training. During Stage 1 (section 5.2) the CLIP backbone acquired a semantically ordered latent space; Stage 2 keeps this backbone frozen and trains a shared MLP and two lightweight heads:

1. *Shared MLP* - Small 2-layer network outputs a compact representation of the feature vector obtained from OrdinalCLIP.

2. *Regression head*— A single layer of 128 units maps the incoming feature vector $\mathbf{f}$ to a scalar logit z. Sigmoid rescaling produces a continuous score in the Harvard four-level range as described in equation 5.1.

3. *Ranking head* - Receives the concatenation $[\mathbf{f}_i \,\|\, \mathbf{f}_j]$ of two images' embeddings and outputs a comparison score s_{ij}. Positive scores indicate that image i should rank higher than image j. This concatenation, rather than the difference used in the original R^2-ResNeXt, maximises the amount of information the ranking head receives.

The training objective is described in subsection 5.1.1.2.

5.3.2 Experiments

The hyper-parameters for the experiments done with the full architecture mirror those described in sections 5.1.2 and 5.2.2.

The training of the network happens in two stages as follows:

1. **OrdinalCLIP** - OrdinalCLIP is trained independently, producing an image encoder guided by the rich language prototype generated by the CLIP text encoder. The resulting image encoder is used as the backbone for our architecture, transforming the input images into feature vectors. Figure 5.3 shows a diagram of this training stage.

2. **Siamese Network** - Once the backbone is ready, the OrdinalCLIP image encoder is frozen, and the second training stage begins. In this stage, we train the Siamese portion of the network, where the model improves its ranking knowledge through direct comparison between images in the ranking head, which guides the regression objective of the main head due to the shared weights. This portion of the network learns ordinality through a more direct approach by learning to rank pairs of images. This is crucial since the low amount of data is not enough to train the complex OrdinalCLIP architecture to a desirable performance. Figure 5.4 shows the final stage of training.

5.4 Overcoming Data Scarcity

Computer vision in medical environments, particularly for surgical procedures, presents many challenges, one of which is the relatively low volume of labelled data. Labelling medical images

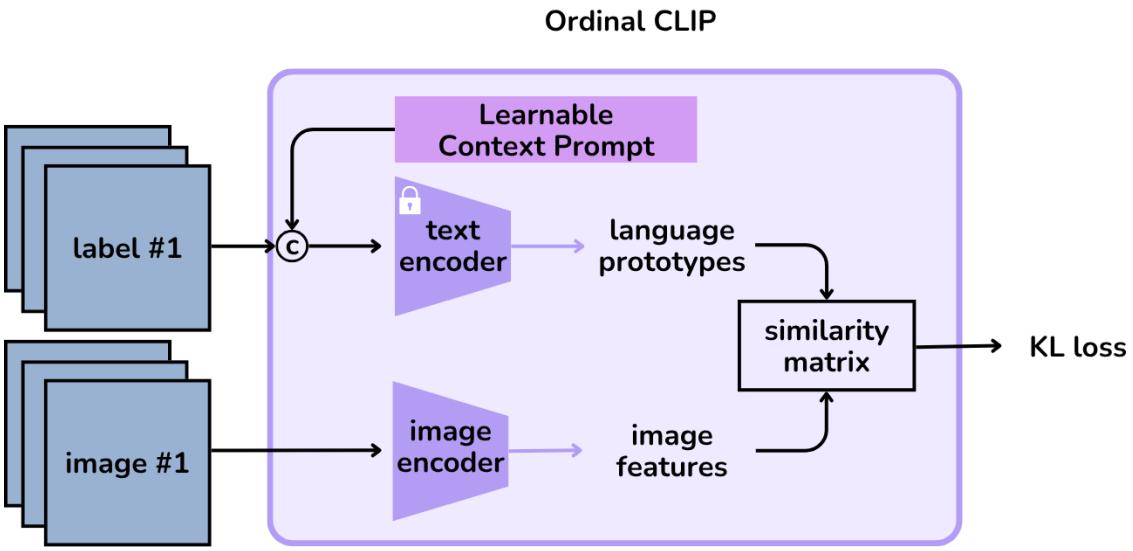


Figure 5.3: Diagram of the first training stage for the Siamese OrdinalCLIP architecture.

requires expert annotators and tends to be very time-consuming. To alleviate this issue, many techniques can be used to increase the availability and diversity of data.

In this work we found that the extreme classes, namely *poor* and *excellent*, were harder for the models to detect due to their scarcity. This section presents the methods and experiments developed to improve the detection of these classes.

5.4.1 Morphing: synthesising the *Poor* class

Starting from an unlabelled anterior image, we compute a dense set of contour key-points and pass the photograph, together with its key-points, to a morphing function, recently introduced by Montenegro et al. (2025) [38]. Internally, the function:

1. crops a padded ROI around the breast boundary;
2. feeds the cropped image and a binary breast mask into a *Double–U-Net* in-painting network that was trained to hallucinate plausible tissue under shape deformations;
3. concatenates the network output with the untouched contralateral side, producing an asymmetrically warped full-resolution image.

The deformation strength is controlled by a multiplicative factor $\in (0.5, 0.85]$; lower values push the lower breast contour further downwards, exaggerating ptosis. A random choice of *left* or *right* breast avoids mode collapse. This pipeline is applied to 40 post-surgery photos that contain key point data but not an aesthetic label assigned by an expert. Figure 5.5 displays the some of the morphed images.

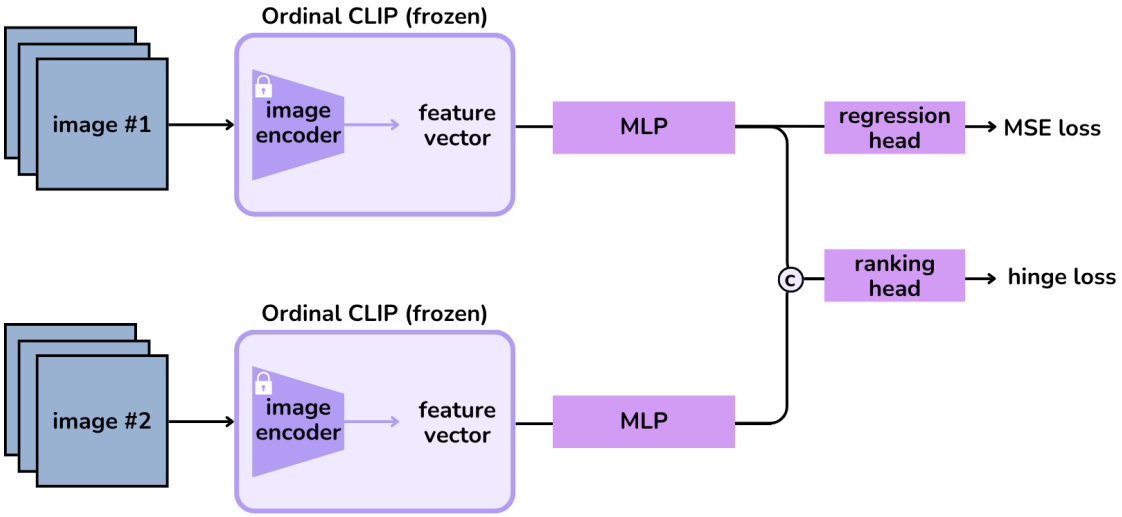


Figure 5.4: Diagram of the second training stage for the Siamese OrdinalCLIP architecture.

5.4.2 Augmentation improvements for the *Excellent* class

The *excellent* class is inherently more difficult to tackle. As the generated labels are not assigned by experts, it is important that the synthetic images are very strong *excellents* so as to not leave doubt on the boundary with the *good* class. Two approaches were tested:

Perfect mirroring From a pre-operative photograph, we draw the centre-line that joins the sternal notch to the nipple midpoint, rotate the image so that the centre-line becomes vertical, flip a random side, and rotate back. The result is a perfectly symmetrical rendering. Manual review showed that the generated images are too unrealistic, and would make the model even more conservative in assigning *excellent* labels; therefore, mirrored samples were *not* used for training but retained for qualitative demonstrations (Figure 5.6).

Pre-surgery photographs While the amount of labelled images in the dataset is relatively low, there is a wide variety of non-labelled images. We tested using raw pre-operative images to supplement the *excellent* class by sampling 40 random pre-operative images. Through manual inspection, it was found that many of the images contained peculiarities that would not be found in excellent reconstructions, including pronounced natural asymmetries, scars from previous surgeries, and differences in colouration due to bruising and other factors. Despite these problems, pre-operative images remained our best solution for natural *excellent* class supplementation, leading to the manual selection of 20 images within the pre-operative set that would serve to increase the variety of data in this class. Figure 5.7 shows a subset of the chosen images.

5.4.3 Pseudo-labelling of unannotated data

Pseudo-labelling consists of training a model on labelled data, passing unlabeled data through the model, and keeping the most confident predictions to then retrain the model with more data.

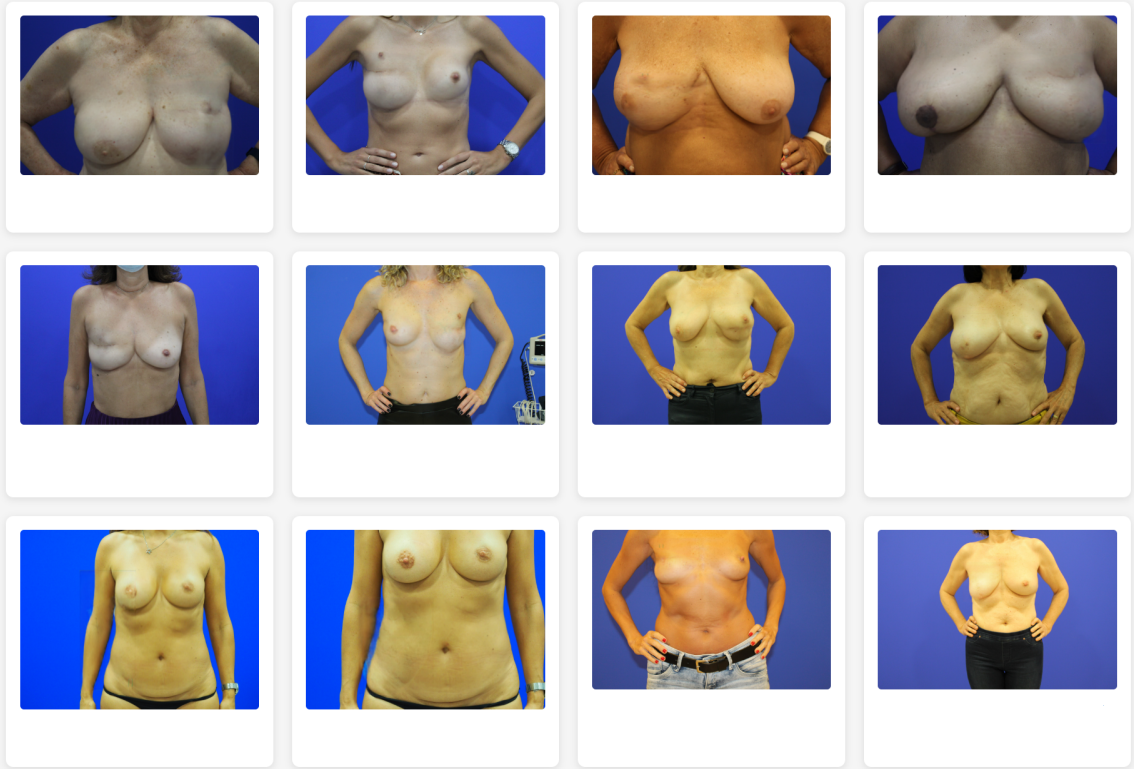


Figure 5.5: Examples of synthetic images of class "poor" generated through morphing

While this is a compelling technique, the model must be very robust in predicting all classes, as pseudo-labels can easily reinforce the model's biases.

To implement pseudo-labelling, we start by selecting all post-operative data that was not used for poor morphing, totalling 2,539 front-view post-surgery images. These images are then fed into the Siamese OrdinalCLIP model which predicts a continuous score rounds it to the nearest integer, and keeps the prediction only if the confidence $c = 1 - |y - \hat{y}|$ exceeds 0.72.

However, it was noted that the model was having trouble assigning *excellent* labels with sufficient confidence. To combat this, the first approach was to generate as many excellent pseudo-labels as the bottleneck class, allowing the most confident *excellents* even if the confidence value was under the threshold. This means that if the number of predictions above the threshold for each class are 0 *excellent*, n_1 *good*, n_2 *bad*, and n_3 *poor*, then the number of pseudo-labels generated for the *excellent* class will be $w = \min\{n_1, n_2, n_3\}$. In fact, to avoid exacerbating the class imbalance present in the dataset, the amount of pseudo labels generated for each class is limited to w, and the final amount of generated labels will be $4w$.

After balancing, 416 pseudo-labelled samples remained.

5.5 Explainability

Deep networks often achieve high predictive accuracy at the expense of interpretability. In medicine this opacity is a serious obstacle: a clinician must understand *why* a system recommends a cer-

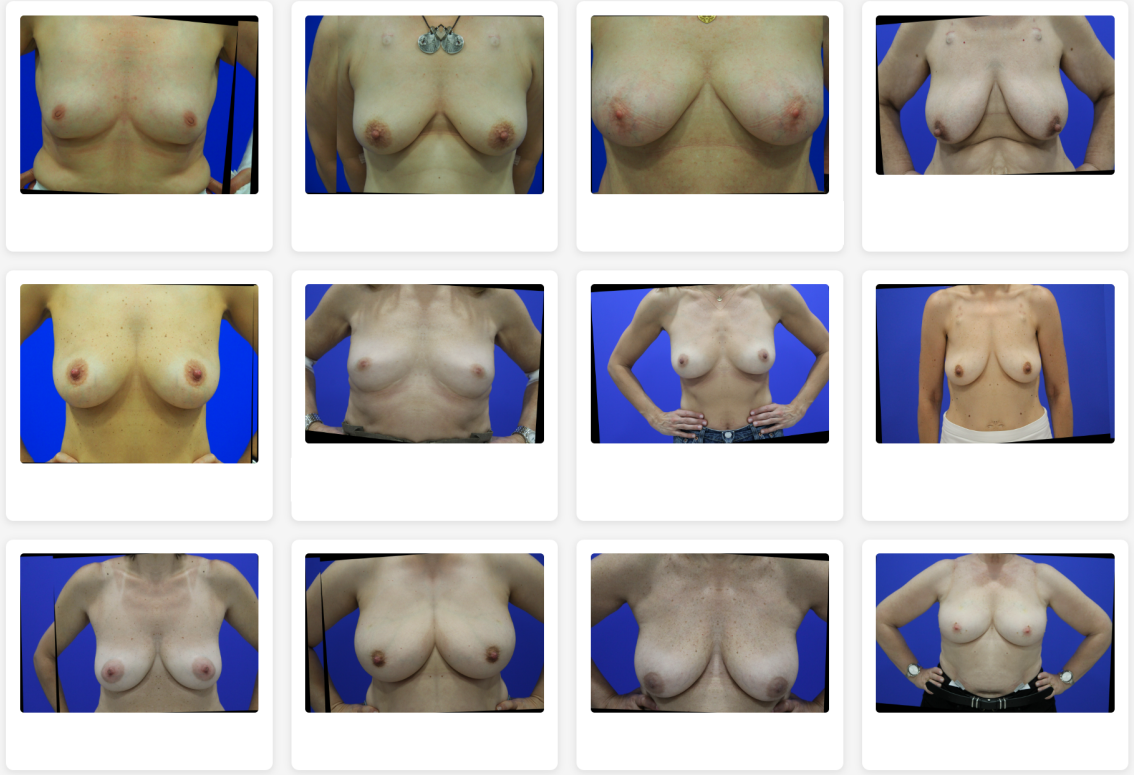


Figure 5.6: Examples of synthetic images of class *excellent* generated through mirroring

tain treatment or, in our case, assigns a particular aesthetic grade. We address this requirement by augmenting the Siamese OrdinalCLIP model with a *concept bottleneck* (CB) layer that forces the network to predict explicit, clinically meaningful attributes (e.g. breast symmetry), before computing the final aesthetic score.

5.5.1 Methodology

5.5.1.1 Clinical motivation

Surgeons routinely judge breast-surgery outcomes by decomposing the visual information into a handful of independent factors such as symmetry, scar visibility, and colour differences. Replicating this workflow inside the network makes the prediction pipeline more transparent and allows practitioners to cross-check each intermediate concept against their own assessment, increasing trust and facilitating error analysis.

5.5.1.2 Architecture

Figure 5.8 situates the CB layer in the full pipeline. After the frozen OrdinalCLIP image encoder produces a 1024-D feature vector $\mathbf{f}$, the vector is passed to a collection of concept modules. Each module is a small MLP that outputs a scalar in the range $[0, 1]$ representing one clinical concept. In

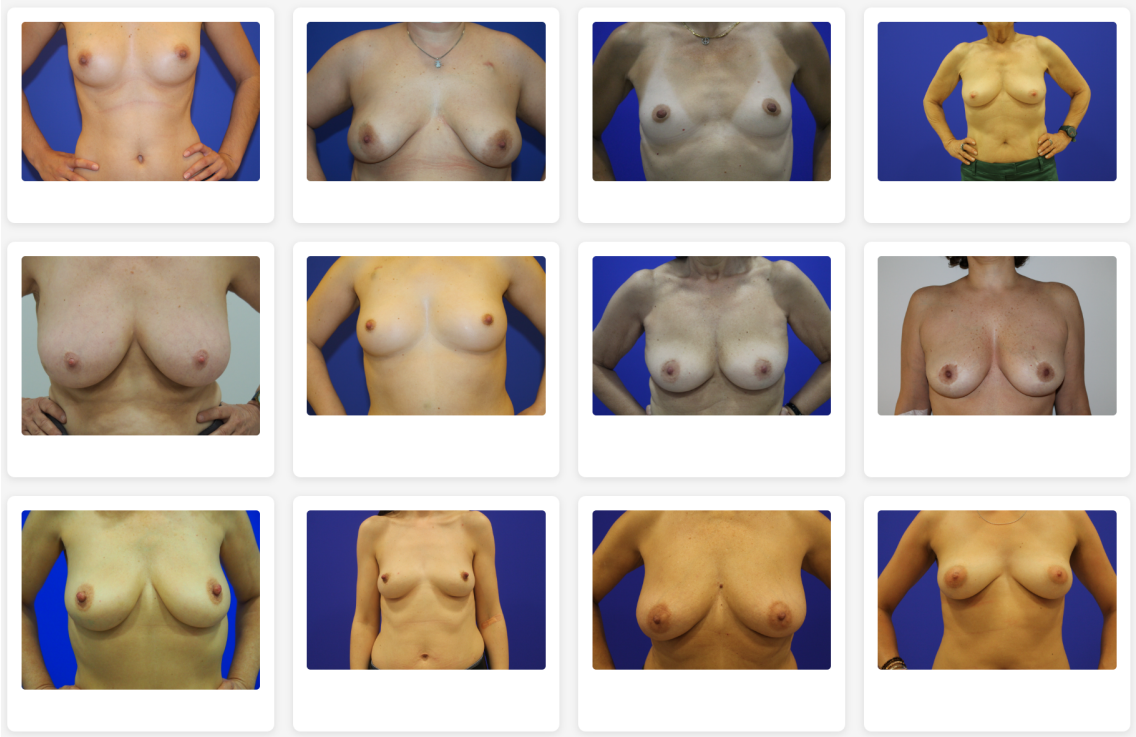


Figure 5.7: Examples of image samples from pre-operative data to serve as supplementation for the *excellent* class.

the current implementation we support a single concept, **symmetry**, but the framework is designed for extension.

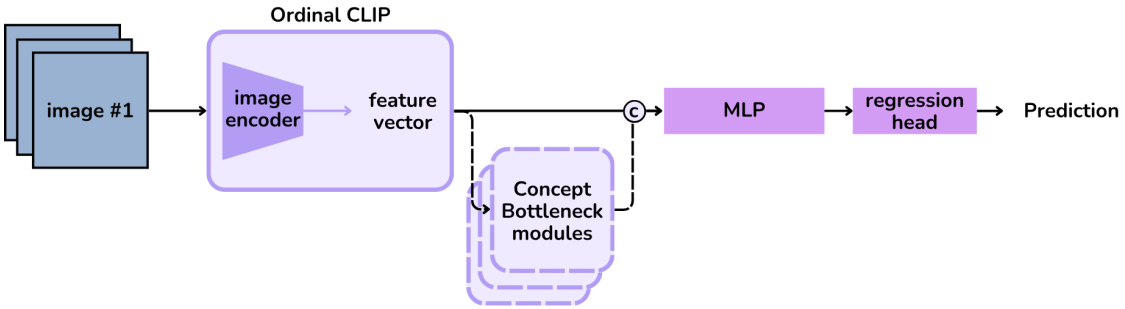


Figure 5.8: Inference diagram with concept bottleneck modules.

Symmetry bottleneck The symmetry module maps $\mathbf{f}$ through a $1024 \to 256 \to 128 \to 1$ network with sigmoid activation. Its output c_{sym} is extended to length 512 and concatenated to the original feature vector, yielding a *concept-augmented* feature $\tilde{\mathbf{f}} = [\mathbf{f} \,\|\, c_{\text{sym}}]$. Because each concept produces an output that is half the dimension of the original feature vector, the new output size is $d(1+0.5k)$ for k active concepts. The siamese network described in Section 5.3.1 operates on $\tilde{\mathbf{f}}$ with no other changes besides the input dimension.

5.5.1.3 Training objective

During training we supervise three quantities:

$$\mathscr{L} = \underbrace{\mathscr{L}_{\text{MSE}}(\hat{y}, y)}_{\text{regression}} + \lambda \underbrace{\mathscr{L}_{\text{hinge}}(s_{ij})}_{\text{ranking}} + \gamma \underbrace{\mathscr{L}_{\text{MSE}}(c_{\text{sym}}, t_{\text{sym}})}_{\text{concept}},$$

where t_{sym} is the ground-truth symmetry score derived from the Lower Breast Contour (pLBC) measurements recorded in the metadata of the dataset. The weights follow the code defaults $\lambda = 0.5$ and $\gamma = 0.1$. The training diagram for stage 2 is extended as shown in figure 5.9

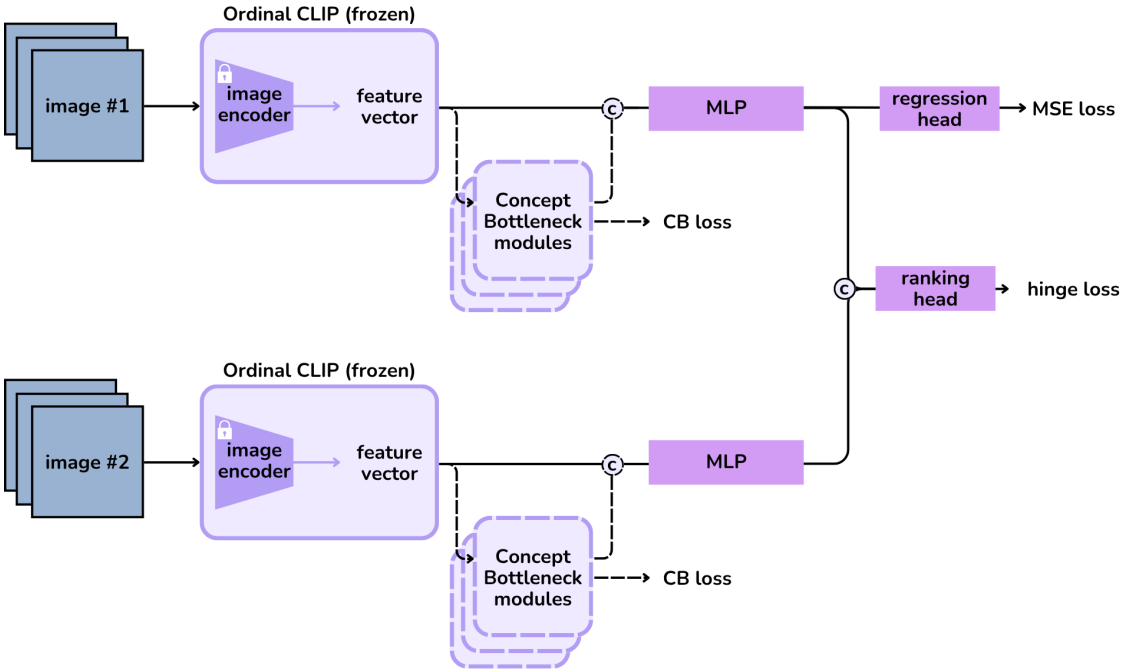


Figure 5.9: Diagram for the second stage of training, with the concept bottleneck modules.

5.5.1.4 Inference and toggling

At inference time, the model can be queried in two modes: (i) *CB-on*, the default, returns an aesthetic score together with concept predictions; (ii) *CB-off* deactivates one or more concepts. This *what-if* analysis is invaluable when a surgeon disagrees with a score and wishes to identify the offending factor.

5.5.2 Experiments

Concept labels Symmetry ground truth can be extracted from many of the objective symmetry-focused metrics described in the literature. For this work, both BRA and LBC were tested, specifically their sternal notch distance normalised versions.

5.5.3 Concept Bottleneck Explainability Experiment

To quantify how the learned *symmetry* concept impacts our aesthetic predictions, we perform a counterfactual experiment over a held-out set of four breast images. For each image we record:

1. The original aesthetic score $\hat{y}_{\text{orig}}$ and original symmetry bottleneck value c_{orig}.
2. The score when forcing symmetry to 0 (fully asymmetric): $\hat{y}_{c=0}$.
3. The score when forcing symmetry to 1 (perfectly symmetric): $\hat{y}_{c=1}$.

From these, we compute the two impacts

$$\Delta_0 = \hat{y}_{c=0} - \hat{y}_{\text{orig}} \qquad \text{and} \qquad \Delta_1 = \hat{y}_{c=1} - \hat{y}_{\text{orig}},$$

as well as the total prediction range $|\hat{y}_{c=1} - \hat{y}_{c=0}|$.

Protocol. We use the SiameseOrdinalCLIP model with concept bottleneck trained as explained in section 5.5. During inference, we override the scalar symmetry concept in the bottleneck and check the influence of symmetry for each image.

5.5.4 Discussion

Clinical relevance

Explainability is not a luxury in oncology follow-up; it is a pre-condition for clinical adoption. The concept bottleneck exposes *why* the model penalises a particular photograph, enabling surgeons to analyse model outputs. Furthermore, this gives patients insight into specific characteristics of their surgery. Explicit concept scores can also be charted over successive visits to monitor a patient's progression, a task for which a single aggregate grade is not well suited.

Limitations and future work

The current study covers only one concept; other factors, such as scar visibility or discolouration, remain embedded in the black box. A broader bottleneck would require additional annotations and careful balancing of loss weights to prevent over-regularisation.

Chapter 6

Results and Discussion

In this chapter, we present the results from all experiments.

6.1 Results

Evaluation metrics Since the task is ordinal and class-imbalanced, we report (i) Root Mean Squared Error (RMSE), (ii) Mean Average Error (MAE), (iii) Macro-Averaged MAE (M-MAE), (iv) Accuracy (Acc), and (v) Ranking Accuracy (Rank-Acc).

Included are various ablation studies to:

- Measure the impact of the alteration of the fair/poor classes to bad/horrible on OrdinalCLIP.
- Measure the impact of the pairwise ranking loss by altering λ on R^2-ResNext.
- Measure the impact of the augmentation of the poor and excellent classes (+ Aug) on the proposed SiameseOrdinalCLIP.
- Measure the impact of pseudo-labelling on the final model.
- Measure the impact of the concept bottleneck module (+ CB) on the proposed network.
- Measure the impact of different symmetry measures as labels for the concept bottleneck module.

Table 6.1 summarises the performance of all the models.

Table 6.1: Aesthetic evaluation results

Model	Acc ↑	Rank-Acc ↑	RMSE ↓	MAE ↓	M-MAE ↓
BCCT.core	0.5721	0.5649	0.7153	0.4558	0.6057
ResNet-50	0.4834	0.7635	0.7994	0.6246	0.6029
R^2-ResNeXt $\lambda = 0.1$	0.4787	0.6890	0.8254	0.6611	0.7600
R^2-ResNeXt $\lambda = 0.5$	0.5071	0.6963	0.8137	0.6527	0.7317
R^2-ResNeXt $\lambda = 0.9$	0.4976	0.6887	0.8183	0.6529	0.7471
OrdinalCLIP: fair/poor	0.4834	0.4902	0.8290	0.5735	0.6852
OrdinalCLIP: bad/horrible	0.5166	0.5360	0.7910	0.5308	0.6017
Siam.Ord.CLIP	0.6114	0.8616	0.6300	0.4856	0.3394
Siam.Ord.CLIP + Aug	**0.6588**	**0.8773**	**0.5641**	**0.4510**	**0.3104**
Siam.Ord.CLIP + Aug + CB (pLBC)	0.6161	0.8463	0.6142	0.4974	0.3838
Siam.Ord.CLIP + Aug + CB (pBRA)	0.5545	0.8271	0.6516	0.5212	0.5388
Siam.Ord.CLIP + Pseudo	0.4882	0.7275	0.8103	0.6430	0.5771

6.2 Discussion

This section discusses the results obtained for each architecture in detail.

6.2.1 R^2-ResNeXt

The R^2-ResNeXt inspired architecture achieves the best results when $\lambda = 0.5$ with 51% accuracy and 70% ranking accuracy. While it surpasses BCCT.core in ranking accuracy and ResNet-50 in accuracy, all other metrics are worse than in the baselines. Given the fact that the Resnet-50 backbone is frozen on the siamese model (contrary to the baseline, where it remains unfrozen), and the very specific domain of the data (post-operative breast images) differs significantly from ImageNet, we conclude that the relatively few tunable parameters of the siamese architecture along with a general backbone are not enough to surpass simple regression on a fine-tuned backbone.

6.2.2 OrdinalCLIP

The language-guided architecture achieves 51% accuracy and 54% ranking accuracy, with a slight improvement to RMSE, MAE, and Macro-MAE over the siamese model. Still, these results are unsatisfactory when compared to the BCCT.core baseline. These results likely stem from the complexity of the model relative to the amount of available data, as well as the difficult nature of the ordinality concept in our data. Fig. 6.1 shows that the validation loss increases as the training loss decreases, a clear sign of overfitting. The results also show the importance of choosing appropriate label names for the task at hand, with the changed labels (*bad* and *horrible*) achieving significantly better metrics across the board.

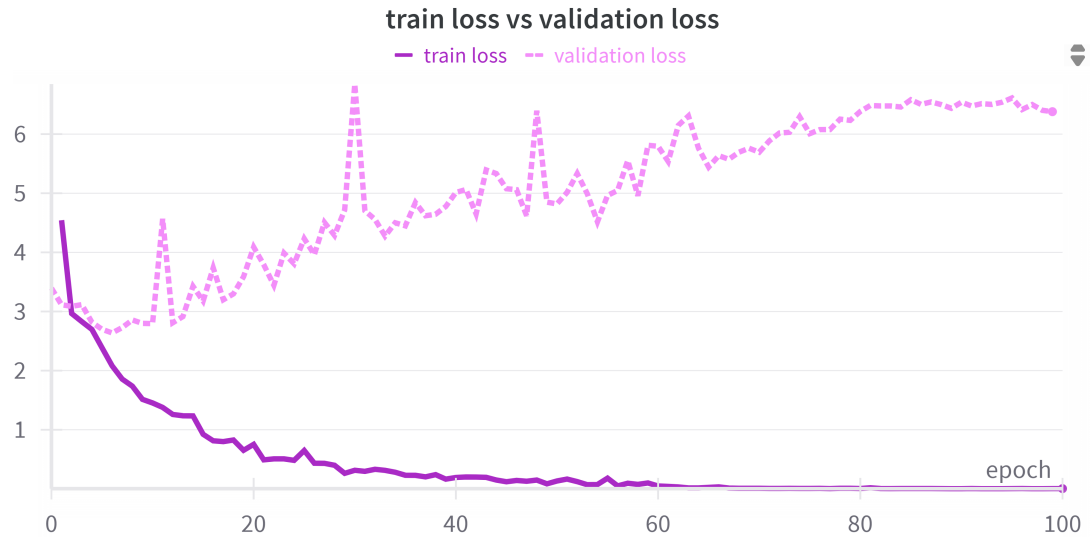


Figure 6.1: Training vs Validation split for the OrdinalCLIP model (w/ label change)

6.2.3 SiameseOrdinalCLIP

The proposed SiameseOrdinalCLIP architecture surpasses the state-of-the-art aesthetic evaluation model BCCT.core in all metrics, achieving 61.14% accuracy and 86.16% ranking accuracy. Data augmentation further improves these results, bringing the average improvement over BCCT.core to over 8% in accuracy and over 30% in ranking accuracy showing the classification process is much more consistent with the experts' opinions. The new architecture also lowers the number of severe misclassifications from 6 to an average of 1.67. This improvement suggests a synergistic effect between the OrdinalCLIP backbone and the siamese network; on one hand, OrdinalCLIP benefits from further optimisation from the direct pairwise ranking of images; on the other hand, the siamese portion benefits from the feature vector produced by OrdinalCLIP, which contains rich ordinal information. The symmetry concept bottleneck module introduces explainability into the network at the cost of a slight decrease in performance. For the symmetry measure, pLBC performs significantly better than pBRA, but a more involved strategy for symmetry labelling could further boost the model's performance. Finally, pseudo-labelling proved ineffective in this work, likely due to confirmation bias. Since the model often confuses *excellent* with *good* images, using pseudo-labels further reinforces this behaviour and leads to more confusion. We expect that this technique's viability increases as more data is collected. Fig. 6.2 shows the average confusion matrix over the three training runs done on the SiameseOrdinalCLIP + Aug architecture, where the improvement over the baseline architectures becomes apparent. The model is able to easily detect *poor* images and achieves solid performance on all other classes, with slightly more difficulty in distinguishing *excellent* from *good*. This contrasts with BCCT.core, where the predictions are overly concentrated on the majority class (*good*), which inflates accuracy.

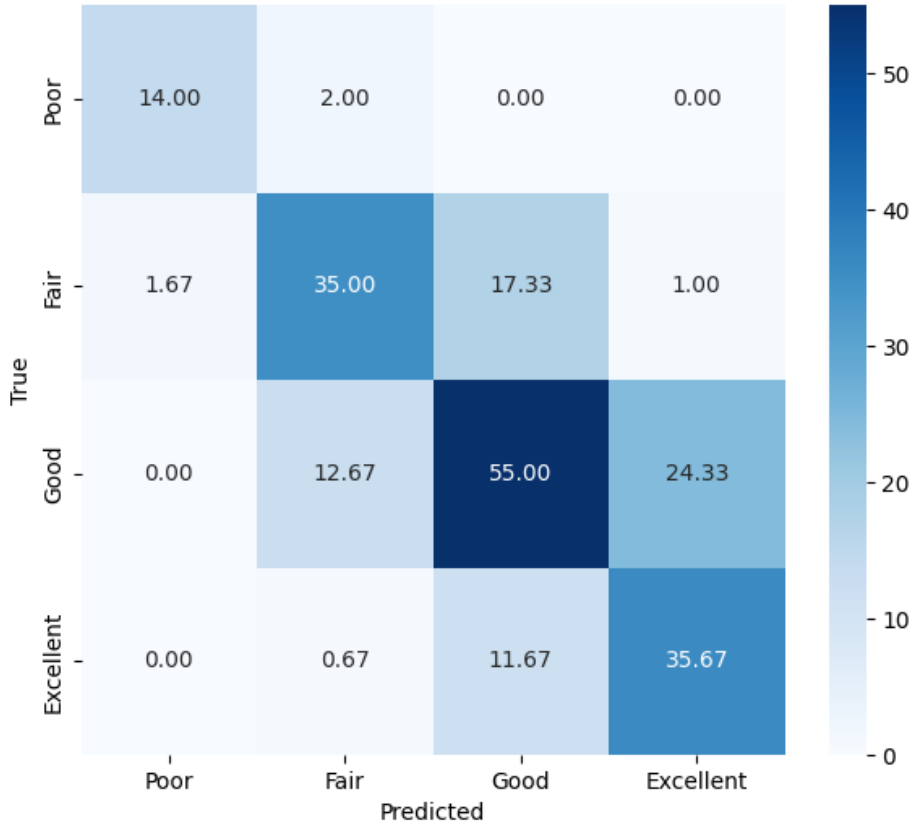


Figure 6.2: Averaged confusion matrix for SiameseOrdinalCLIP + Aug

6.2.4 Cross-Validation Experiment

To verify the robustness and generalisability of our base *SiameseOrdinalCLIP* architecture, we conducted 5-fold cross-validation. Each fold was created via stratified sampling on the four Harvard classes, with an internal 80/20 split for training and validation, and a held-out test set. We report the mean $\pm$ standard deviation of the key metrics across the five test folds in Table 6.2.

As shown, the model maintains performance comparable to the single-split evaluation, with low variance across folds. This confirms that our training and sampling strategy generalises well to unseen data, and that the conclusions drawn in the main experiments are not dependent on a particular data split.

6.2.5 Counterfactual Explainability Results

Table 6.3 summarises the per-image impacts, and Figure 6.3 visualises the impact of the original vs. manipulated scores on the final classification.

Table 6.2: Cross-validation results (mean ± std across 5 folds).

Metric	Mean ± Std
Accuracy ↑	0.6498 ± 0.0152
Ranking Accuracy ↑	0.8733 ± 0.0149
RMSE ↓	0.5803 ± 0.0172
MAE ↓	0.4644 ± 0.0141
Macro-MAE ↓	0.3501 ± 0.0193

6.2.5.1 Summary of Results

- **Average impact of setting symmetry to 0:** $\overline{\Delta_0} = -0.703$ points.
- **Average impact of setting symmetry to 1:** $\overline{\Delta_1} = +1.756$ points.
- **Average prediction range:** $\overline{|\hat{y}_{c=1} - \hat{y}_{c=0}|} = 2.459$ points.

These results confirm that the model's symmetry concept has a very strong positive effect on predicted aesthetic scores, especially for images with low natural symmetry. On the held-out set used for this experiment, perfect symmetry translates to the attribution of rank *excellent*, regardless of original classification, while perfect asymmetry always results in lower scores, even though these don't necessarily mean lower rank attribution.

6.2.5.2 Findings

Counterfactual manipulation of the symmetry bottleneck reveals that:

- *Increasing symmetry* consistently boosts aesthetic predictions (up to $\Delta_1 \approx +2.82$). This is exacerbated due to the training data consistently having lower symmetry values, since even *excellent* images have symmetry scores in the 0.3 to 0.4 range.
- *Decreasing symmetry* causes smaller drops ($\Delta_0 \approx -0.70$).
- Cases with originally low symmetry benefit most from forcing perfect symmetry.

These findings empirically validate the interpretability benefits of our Concept-Bottleneck approach and demonstrate its utility for actionable, case-level feedback.

Table 6.3: Individual counterfactual results for symmetry.

Image (ground truth score)	Orig. predicted Score	Orig. Symmetry	$\hat{y}_{c=0}$	$\hat{y}_{c=1}$	Δ_0	Δ_1	Range
Image 1 (1 / Poor)	1.33 (Poor)	0.128	1.29	4.15	-0.04	+2.82	2.86
Image 2 (2 / Fair)	2.86 (Good)	0.205	2.07	4.24	-0.79	+1.3	2.16
Image 3 (3/ Good)	2.49 (Fair)	0.156	1.89	4.20	-0.60	+1.71	2.30
Image 4 (4/ Excellent)	3.12 (Good)	0.295	1.73	4.24	-1.39	+1.12	2.51

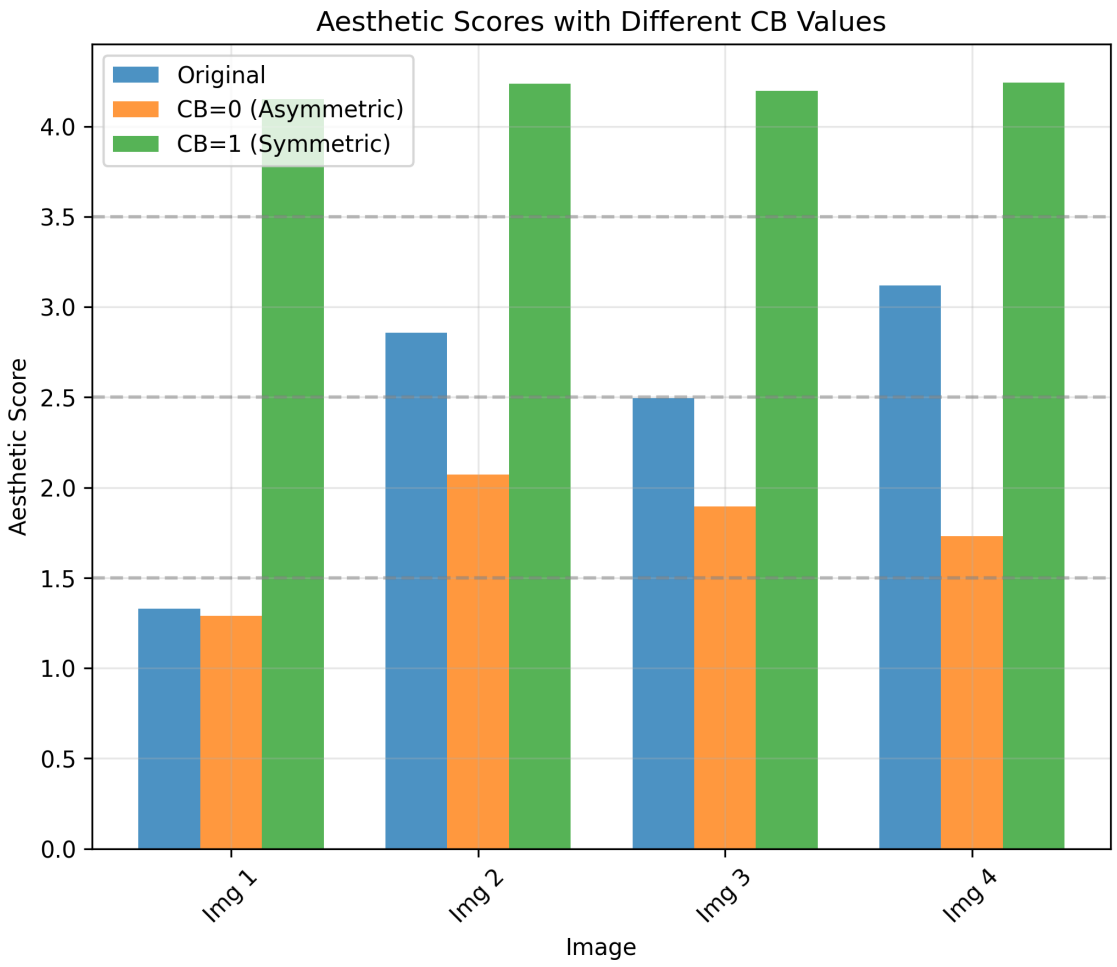


Figure 6.3: Original vs. CB-manipulated aesthetic scores (CB=0 asymmetric, CB=1 symmetric)

Chapter 7

Conclusions

Artificial-intelligence methods for postoperative breast-aesthetic assessment are still passing through an early, exploratory phase in which scarce, noisy data and a lack of agreed-upon metrics slow research progress. Within this landscape, this dissertation shows that deep, language–vision models trained to respect order rather than just class membership can already outperform the current state-of-the-art solutions while being fully automatic, handling different types of surgeries, and maintaining some degree of explainability.

This work first confirmed that coupling a ResNet-50 backbone with a joint regression-and-ranking head (R2-ResNeXt) improves ordinal discrimination, yet remains constrained by limited learnable parameters and ImageNet bias. It then introduced OrdinalCLIP, teaching the backbone the semantic notion that *poor* < *fair* < *good* < *excellent*. Finally, freezing that backbone and adding a light Siamese comparator produced SiameseOrdinalCLIP, which achieves an average of 65.9% accuracy and 87.7% ranking accuracy, around eight and thirty percentage points better than BCCT.core, respectively. A symmetry concept bottleneck lets clinicians analyse the score and even toggle the influence of different concepts at inference, with only a marginal loss in performance.

Despite these gains, several weaknesses persist. Data remain both small (only 1,406 labelled images) and skewed toward the middle Harvard classes, forcing oversampling, synthetic morphing, and still leaving the extreme classes fragile. Inter-observer disagreement injects label noise that is difficult to handle. The network also only analyses a single anterior view, and was trained on data acquired from a single centre, its fairness across skin tones and the different aesthetic expectations of different cultures, as well as its performance on different datasets, remains to be tested.

These findings suggest three priorities for future research. First, multi-centre data collection with concordant annotation guidelines is indispensable for a truly universal aesthetic assessment system. Second, broadening the concept bottleneck to scar visibility, contour regularity, skin discolouration, etc., can transform the tool from a black-box grader into an interactive assistant that clarifies why it penalises an image and how a score might shift under hypothetical improvements. Third, exploring whether explicitly incorporating clinical (e.g., type of surgery and more asymmetry measurements) and demographic (e.g., cup size) concepts improves the network's performance.

In short, while objective aesthetic assessment of breast cancer treatment procedures remains a difficult problem to tackle, this dissertation demonstrates that a ranking-aware, language-guided Siamese framework is a promising solution. Continued investment in richer data, broader explainability, and prospective validation will be essential to translate these results into clinical viability.

References

[1] Luca Bertinetto et al. *Fully-Convolutional Siamese Networks for Object Tracking*. arXiv:1606.09549 [cs]. Dec. 2021. DOI: 10.48550/arXiv.1606.09549. URL: http://arxiv.org/abs/1606.09549.

[2] Jane Bromley et al. "Signature Verification using a "Siamese" Time Delay Neural Network". In: *Advances in Neural Information Processing Systems*. Vol. 6. Morgan-Kaufmann, 1993.

[3] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. "Rank consistent ordinal regression for neural networks with application to age estimation". In: *Pattern Recognition Letters* 140 (Dec. 2020). arXiv:1901.07884 [cs], pp. 325–331. ISSN: 01678655. DOI: 10.1016/j.patrec.2020.11.008. URL: http://arxiv.org/abs/1901.07884.

[4] J.S. Cardoso, W. Silva, and M.J. Cardoso. "Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment". In: *Breast* 49 (2020), pp. 123–130. DOI: 10.1016/j.breast.2019.11.006.

[5] Jaime S Cardoso and Maria J Cardoso. "Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment". en. In: *Artif. Intell. Med.* 40.2 (June 2007), pp. 115–126.

[6] Jaime S. Cardoso and Maria J. Cardoso. "Breast Contour Detection for the Aesthetic Evaluation of Breast Cancer Conservative Treatment". en. In: *Computer Recognition Systems 2*. Ed. by Janusz Kacprzyk et al. Vol. 45. Series Title: Advances in Soft Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 518–525. ISBN: 978-3-540-75174-8 978-3-540-75175-5. DOI: 10.1007/978-3-540-75175-5_65. URL: http://link.springer.com/10.1007/978-3-540-75175-5_65.

[7] Jaime S. Cardoso, Inês Domingues, and Hélder P. Oliveira. "Closed Shortest Path in the Original Coordinates with an Application to Breast Cancer". In: *International Journal of Pattern Recognition and Artificial Intelligence* 29.01 (Feb. 2015). Publisher: World Scientific Publishing Co., p. 1555002. ISSN: 0218-0014. DOI: 10.1142/S0218001415550022.

[8] Jaime S. Cardoso, Wilson Silva, and Maria J. Cardoso. "Evolution, current challenges, and future possibilities in the objective assessment of aesthetic outcome of breast cancer locoregional treatment". In: *The Breast* 49 (Feb. 2020), pp. 123–130. ISSN: 0960-9776. DOI: 10.1016/j.breast.2019.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0960977619310987.

[9] Maria João Cardoso et al. "Turning subjective into objective: the BCCT.core software for evaluation of cosmetic results in breast cancer conservative treatment". eng. In: *Breast (Edinburgh, Scotland)* 16.5 (Oct. 2007), pp. 456–461. ISSN: 0960-9776. DOI: 10.1016/j.breast.2007.05.002.

[10] Parag S. Chandakkar, Vijetha Gattupalli, and Baoxin Li. *A Computational Approach to Relative Aesthetics*. arXiv:1704.01248 [cs]. Apr. 2017. DOI: 10.48550/arXiv.1704.01248. URL: http://arxiv.org/abs/1704.01248.

[11] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. arXiv:2002.05709 [cs]. July 2020. DOI: 10.48550/arXiv.2002.05709. URL: http://arxiv.org/abs/2002.05709.

[12] S. Chopra, R. Hadsell, and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. ISSN: 1063-6919. June 2005, 539–546 vol. 1. DOI: 10.1109/CVPR.2005.202. URL: https://ieeexplore.ieee.org/document/1467314.

[13] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. URL: https://ieeexplore.ieee.org/document/5206848.

[14] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. "Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2019, pp. 4095–4099. DOI: 10.1109/ICASSP.2019.8682632. URL: https://ieeexplore.ieee.org/document/8682632.

[15] Ryan Eloff, Herman A. Engelbrecht, and Herman Kamper. *Multimodal One-Shot Learning of Speech and Images*. arXiv:1811.03875 [cs]. Apr. 2019. DOI: 10.48550/arXiv.1811.03875. URL: http://arxiv.org/abs/1811.03875.

[16] F. Fitzal et al. "The use of a breast symmetry index for objective evaluation of breast cosmesis". In: *The Breast* 16.4 (2007), pp. 429–435. ISSN: 0960-9776. DOI: https://doi.org/10.1016/j.breast.2007.01.013. URL: https://www.sciencedirect.com/science/article/pii/S0960977607000422.

[17] F. Fitzal et al. "The use of a breast symmetry index for objective evaluation of breast cosmesis". In: *The Breast* 16.4 (Aug. 2007), pp. 429–435. ISSN: 0960-9776. DOI: 10.1016/j.breast.2007.01.013. URL: https://www.sciencedirect.com/science/article/pii/S0960977607000422.

[18] Nuno Freitas et al. "Deep Edge Detection Methods for the Automatic Calculation of the Breast Contour". en. In: *Bioengineering* 10.4 (Apr. 2023). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 401. ISSN: 2306-5354.

[19] Ross Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. ISSN: 1063-6919. June 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81. URL: https://ieeexplore.ieee.org/document/6909475.

[20] Tiago Gonçalves et al. "A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes". en. In: *Health and Technology* 10.4 (July 2020), pp. 891–903. ISSN: 2190-7196. DOI: 10.1007/s12553-020-00423-8. URL: https://doi.org/10.1007/s12553-020-00423-8.

[21] Jean-Bastien Grill et al. *Bootstrap your own latent: A new approach to self-supervised Learning*. arXiv:2006.07733 [cs]. Sept. 2020. DOI: 10.48550/arXiv.2006.07733. URL: http://arxiv.org/abs/2006.07733.

[22] Chenqi Guo et al. “A fully automatic framework for evaluating cosmetic results of breast conserving therapy”. In: *Machine Learning with Applications* 10 (Dec. 2022), p. 100430. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2022.100430. URL: https://www.sciencedirect.com/science/article/pii/S2666827022001050.

[23] R. Hadsell, S. Chopra, and Y. LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. ISSN: 1063-6919. June 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100. URL: https://ieeexplore.ieee.org/document/1640964.

[24] J. R. Harris et al. “Analysis of cosmetic results following primary radiation therapy for stages I and II carcinoma of the breast”. eng. In: *International Journal of Radiation Oncology, Biology, Physics* 5.2 (Feb. 1979), pp. 257–261. ISSN: 0360-3016. DOI: 10.1016/0360-3016(79)90729-6.

[25] Kaiming He et al. *Deep Residual Learning for Image Recognition*. arXiv:1512.03385. Dec. 2015. DOI: 10.48550/arXiv.1512.03385. URL: http://arxiv.org/abs/1512.03385.

[26] Elad Hoffer and Nir Ailon. *Deep metric learning using Triplet network*. arXiv:1412.6622 [cs]. Dec. 2018. DOI: 10.48550/arXiv.1412.6622. URL: http://arxiv.org/abs/1412.6622.

[27] Shruti Jadon. *Introduction to Different Activation Functions for Deep Learning*. en. Feb. 2022. URL: https://medium.com/@shrutijadon/survey-on-activation-functions-for-deep-learning-9689331ba092.

[28] Orit Kaidar-Person et al. “Evaluating the ability of an artificial-intelligence cloud-based platform designed to provide information prior to locoregional therapy for breast cancer in improving patient's satisfaction with therapy: The CINDERELLA trial”. en. In: (). DOI: 10.1371/journal.pone.0289365. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0289365.

[29] Diederik P. Kingma et al. *Semi-Supervised Learning with Deep Generative Models*. arXiv:1406.5298 [cs]. Oct. 2014. DOI: 10.48550/arXiv.1406.5298. URL: http://arxiv.org/abs/1406.5298 (visited on 07/08/2025).

[30] Pang Wei Koh et al. “Concept Bottleneck Models”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 5338–5348. URL: https://proceedings.mlr.press/v119/koh20a.html.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[32] Dong-Hyun Lee. “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. In: *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)* (July 2013).

[33] Wanhua Li et al. *OrdinalCLIP: Learning Rank Prompts for Language-Guided Ordinal Regression*. arXiv:2206.02338 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2206.02338. URL: http://arxiv.org/abs/2206.02338.

[34] Luojun Lin, Lingyu Liang, and Lianwen Jin. "R2-ResNeXt: A ResNeXt-Based Regression Model with Relative Ranking for Facial Beauty Prediction". In: *2018 24th International Conference on Pattern Recognition (ICPR)*. ISSN: 1051-4651. Aug. 2018, pp. 85–90. DOI: 10.1109/ICPR.2018.8545164. URL: https://ieeexplore.ieee.org/document/8545164.

[35] Luojun Lin, Lingyu Liang, and Lianwen Jin. "Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction". In: *IEEE Transactions on Affective Computing* PP (Aug. 2019), pp. 1–1. DOI: 10.1109/TAFFC.2019.2933523.

[36] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. arXiv:1705.07874 [cs]. Nov. 2017. DOI: 10.48550/arXiv.1705.07874. URL: http://arxiv.org/abs/1705.07874.

[37] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. *Image Patch Matching Using Convolutional Descriptors with Euclidean Distance*. arXiv:1710.11359 [cs]. Oct. 2017. DOI: 10.48550/arXiv.1710.11359. URL: http://arxiv.org/abs/1710.11359.

[38] Helena Montenegro, Maria J. Cardoso, and Jaime S. Cardoso. *An inpainting approach to manipulate asymmetry in pre-operative breast images*. arXiv:2502.05652 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2502.05652. URL: http://arxiv.org/abs/2502.05652.

[39] Javier Monton et al. "A Free Tool for Breast Aesthetic Scale Computation". eng. In: *Annals of Plastic Surgery* 86.4 (Apr. 2021), pp. 458–462. ISSN: 1536-3708. DOI: 10.1097/SAP.0000000000002440.

[40] Deepak Pathak et al. *Context Encoders: Feature Learning by Inpainting*. arXiv:1604.07379 [cs]. Nov. 2016. DOI: 10.48550/arXiv.1604.07379. URL: http://arxiv.org/abs/1604.07379 (visited on 07/08/2025).

[41] R D Pezner et al. "Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer". en. In: *Int. J. Radiat. Oncol. Biol. Phys.* 11.3 (Mar. 1985), pp. 575–578.

[42] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020 [cs]. Feb. 2021. DOI: 10.48550/arXiv.2103.00020. URL: http://arxiv.org/abs/2103.00020.

[43] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. arXiv:1506.01497 [cs]. Jan. 2016. DOI: 10.48550/arXiv.1506.01497. URL: http://arxiv.org/abs/1506.01497.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv:1602.04938 [cs]. Aug. 2016. DOI: 10.48550/arXiv.1602.04938. URL: http://arxiv.org/abs/1602.04938.

[45] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain." en. In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519. URL: https://doi.apa.org/doi/10.1037/h0042519.

[46] Debaditya Roy et al. *Detection of Collision-Prone Vehicle Behavior at Intersections using Siamese Interaction LSTM*. arXiv:1912.04801 [cs]. Dec. 2019. DOI: 10.48550/arXiv.1912.04801. URL: http://arxiv.org/abs/1912.04801.

[47] V Sacchini et al. “Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer”. en. In: *Eur. J. Cancer Clin. Oncol.* 27.11 (Nov. 1991), pp. 1395–1400.

[48] Attaullah Sahito, Eibe Frank, and Bernhard Pfahringer. “Semi-Supervised Learning using Siamese Networks”. In: vol. 11919. arXiv:2109.00794 [cs]. 2019, pp. 586–597. DOI: 10.1007/978-3-030-35288-2_47. URL: http://arxiv.org/abs/2109.00794.

[49] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1503.03832 [cs]. June 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682. URL: http://arxiv.org/abs/1503.03832.

[50] Xintong Shi, Wenzhi Cao, and Sebastian Raschka. “Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities”. In: *Pattern Analysis and Applications* 26.3 (Aug. 2023). arXiv:2111.08851 [cs], pp. 941–955. ISSN: 1433-7541, 1433-755X. DOI: 10.1007/s10044-023-01181-9. URL: http://arxiv.org/abs/2111.08851.

[51] Wilson Silva et al. *Deep Aesthetic Assessment and Retrieval of Breast Cancer Treatment Outcomes*. arXiv:2205.12611. May 2022. DOI: 10.48550/arXiv.2205.12611. URL: http://arxiv.org/abs/2205.12611.

[52] Wilson Silva et al. “Deep Keypoint Detection for the Aesthetic Evaluation of Breast Cancer Surgery Outcomes”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ISSN: 1945-8452. Apr. 2019, pp. 1082–1086. DOI: 10.1109/ISBI.2019.8759331. URL: https://ieeexplore.ieee.org/document/8759331.

[53] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv:1312.6034 [cs]. Apr. 2014. DOI: 10.48550/arXiv.1312.6034. URL: http://arxiv.org/abs/1312.6034.

[54] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556. Apr. 2015. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556.

[55] Tamer Soror et al. “kOBCS©: a novel software calculator program of the Objective Breast Cosmesis Scale (OBCS)”. en. In: *Breast Cancer* 27.2 (Mar. 2020), pp. 179–185. ISSN: 1880-4233. DOI: 10.1007/s12282-019-01006-w. URL: https://doi.org/10.1007/s12282-019-01006-w.

[56] Tamer Soror et al. “New objective method in reporting the breast cosmesis after breast-conservative treatment based on nonstandardized photographs: The Objective Breast Cosmesis Scale”. In: *Brachytherapy* 15.5 (2016), pp. 631–636. ISSN: 1538-4721. DOI: https://doi.org/10.1016/j.brachy.2016.06.008. URL: https://www.sciencedirect.com/science/article/pii/S1538472116304998.

[57] Ricardo Sousa et al. “Breast contour detection with shape priors”. In: *2008 15th IEEE International Conference on Image Processing*. ISSN: 2381-8549. Oct. 2008, pp. 1440–1443. DOI: 10.1109/ICIP.2008.4712036. URL: https://ieeexplore.ieee.org/document/4712036.

[58] Christian Szegedy et al. *Going Deeper with Convolutions*. arXiv:1409.4842. Sept. 2014. DOI: 10.48550/arXiv.1409.4842. URL: http://arxiv.org/abs/1409.4842.

[59] Antti Tarvainen and Harri Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. arXiv:1703.01780 [cs]. Apr. 2018. DOI: 10.48550/arXiv.1703.01780. URL: http://arxiv.org/abs/1703.01780 (visited on 07/08/2025).

[60] Richard Zhang, Phillip Isola, and Alexei A. Efros. *Colorful Image Colorization*. arXiv:1603.08511 [cs]. Oct. 2016. DOI: 10.48550/arXiv.1603.08511. URL: http://arxiv.org/abs/1603.08511 (visited on 07/08/2025).

Appendix A

Further Use of AI Tools

AI-assisted writing software was employed in a supporting role during the preparation of this thesis. Specifically, I used **OpenAI ChatGPT** and **Grammarly** to:

- suggest alternative phrasings and improve paragraph flow,
- detect and correct grammar, spelling, and punctuation errors, and
- flag potential stylistic inconsistencies.

All AI-generated suggestions—whether lexical, syntactic, or stylistic—were individually reviewed and, where appropriate, adapted by me before incorporation. The software was *not* used to create original research content, formulate arguments, design experiments, analyse data, or draw conclusions.

I bear sole responsibility for the originality, accuracy, and academic integrity of the work presented herein.