Breaking the Curse of the Imbalanced Dataset in a Breast Cancer Detection Model

Author
Marharyta Kurban

Advisor
Professor Ribeiro

Minerva Schools at KGI

College of Computational Sciences

2020

Abstract

Even though breast cancer detection is a heavily researched discipline, thermography-based methods have long received insufficient attention. This can be mainly attributed to experimental design flaws, such as too small sample sizes, especially for positive cases, mixed data acquisition methods with unclear pre-processing steps, and omission of controversial cases, leading to non-generalizable and non-reproducible results. In this paper, we quantify and compare human evaluation methods for thermography interpretation available in the literature. We also demonstrate that modern feature engineering and data augmentation techniques, combined with a sufficient number of samples, help to overcome the problem of the imbalanced dataset and prove that thermography can be used as a reliable method for breast cancer screening and lead to much earlier detection, improving cancer patient survival.

This paper was written in LaTex and rendered by overleaf.com. Figures and tables were created in Python. While full supplementary materials cannot be provided, some code and example outputs can be found in the GitHub repository.

Contents

Glossary

Angiogenesis The development of new blood vessels.

Breast Imaging-Reporting and Data System (BI-RADS) Risk assessment and quality assurance tool that provides a widely accepted lexicon and reporting schema for breast imaging.

Contralateral The side of the body opposite to that on which a particular structure or condition occurs.

Delta-T The temperature difference between the hottest point in the asymmetrical vascular region and its corresponding location on the opposite breast.

Lesion Any damage or abnormal change in the tissue of an organism, usually caused by disease or trauma.

Mammography The process of using low-energy X-rays to examine the human breast for diagnosis and screening.

Thermography A procedure that uses an infrared camera to detect heat patterns and blood flow in body tissues.

Vascular Pattern The pattern of blood vessels (arteries and veins).

Acronyms

HOG Histogram of Oriented Gradients.

ICA Independent Component Analysis.

LBP Local Binary Patterns.

LPP Locality Preserving Projections.

MRI Magnetic Resonance Imaging.

PCA Principal Component Analysis.

SHAP Shapley Additive Values.

SMOTE Synthetic Minority Oversampling Technique.

T-SNE T-Distributed Stochastic Neighbor Embedding.

1 Introduction

Breast cancer is prevalent all around the world. Despite a high success rate of modern treatments with 95% survival upon early detection, the disease remains one of the leading causes of death in women due to the difficulty of accurately identifying early-stage tumors (Kennedy et al., 2009). Thus, early detection is the key to reducing breast cancer mortality rate. Mammography is an X-ray of the breast that has long been considered the gold standard for breast cancer detection. While mammography produces images of high quality, it is also very damaging to the breast tissue, increases the risk of radiation-induced breast cancer, and is not very efficient at identification, especially in women with dense breasts. The harmful effects of radiation limit the use of the method to older women and only infrequently, which increases the risk of overlooking aggressive tumors. The potential of infrared thermography as an alternative breast cancer detection tool has been studied for decades. Research suggests that tumors form distinct thermal patterns that can be more consistently captured than mammogram density patterns. The work presented here emphasizes that structured quantitative thermography evaluation approaches, coupled with modern computer vision techniques and data augmentation strategies, lead to above industry-level classification performance on thermal images.

1.1 Breast Cancer Prevalence

Breast cancer is the most diagnosed cancer in women worldwide, with up to 12.3% of the female population affected in their lifetime (Bray et al., 2018). Patients' treatment involves significant stress and financial burden for women and their families and cannot guarantee a complete recovery (Campbell & Ramsey, 2009). Survival rates for this disease vary worldwide and are highly dependent on the healthcare system in a specific country. Advanced and metastatic breast cancer is currently incurable, but treatable, with a median survival rate of 2-3 years. Early diagnosed localized cancers have a 99% survival rate (Noone et al., 2018). Therefore, early detection remains the best way to reduce breast cancer morbidity and mortality.

Although breast cancer in young women is relatively rare, more than 250,000 people living in the US today were diagnosed with it under 40 (YCS, 2019) In young females, the disease tends to be detected at later stages and be more aggressive. Young women also have a higher mortality rate and a higher risk of metastatic recurrence as well as dense breast tissue that cannot be adequately analyzed by mammography (Anders et al., 2011). More and more evidence tells that breast cancer before age 40 has biological differences to cancer types faced by older women (Partridge et al., 2014). Despite this, young women remain underrepresented in many research studies due to the lack of reliable screening data, mainly mammograms.

1.2 Limitations of Mammography

Mammography is a breast cancer detection method that uses low energy X-Ray to detect abnormalities inside breasts. Regular mammogram screening has shown to decrease breast cancer mortality and is used in many government-supported breast cancer screening programs. Despite its prevalence, mammography is not an ideal method for breast cancer screening for several reasons. The cumulative effect of routine mammography screening, which exposes patients to low levels of radiation may increase the risk of developing radiation-induced breast cancer (Miglioretti et al., 2016). Kennedy et al. (2009) assess the risk of annual radiation exposure through mammography starting at different ages. According to their findings, annual screening starting at the age of 20 would cause more radiation-induced breast cancer than it would prevent. Testing women starting at age 30 would unlikely result in a reduction of breast cancer mortality. At age 40, the benefits of screening would only be evident if mortality was reduced by 20%. An article by Law et al. (2007) further supports this view by suggesting that screening before the age of 35 has a higher risk to benefit ratio versus screening after 40.

Mammography is also much less effective at identifying tumors in dense breast tissue, which is usually present in young women (Yaffe, 2008) and women of Asian descent. A study with more than 25,000 Asian women under 50 revealed that almost 80% had either heterogeneously dense or extremely dense breast tissue (Tice et al., 2008), which limits the interpretability of the results

(Freer, 2015). The percentage of total relevant results correctly classified by mammography (recall) for females with low breast density is 87%, but only 62.9% for women with extremely dense breasts. Due to these restrictions, mammography is recommended only to older people.

Other methods, such as Magnetic Resonance Imaging (MRI) scans, are better at identifying existing cancer but suffer from higher costs and too many false positives to be usable as a screening tool. Ultrasonography is another promising method that might become an excellent alternative for younger patients. However, it is unable to identify calcifications, which are a sign of breast cancer in a very early stage (Madjar, 2002). For these reasons, mammogram prevails so far, and self-examination remains the main and sometimes the only alternative for young females.

1.3 Thermography

1.3.1 Background

Given the limitations of mammography, researchers continuously evaluate alternative approaches for breast cancer screening. Infrared thermography is one of the most widely-studied methods. It is a safe and tissue-agnostic method that can close the gap in the early prevention of breast cancer in young women and women with dense breasts.

The idea behind using thermography for cancer detection is that tumors develop blood vessels that deliver nutrients and oxygen to support their growth. Increased blood flow causes a localized temperature increase that can be observed as a small change in skin surface temperature on a thermogram. Gamagami (1996) reported that hypervascularity and hyperthermia could be shown in around 90% of nonpalpable breast cancers and pre-cancerous conditions. Additionally, there is some evidence that thermography can identify physiological changes in tissue that precede pathological changes. Gautherie (1989) states that thermography can detect abnormalities 8-10 years before mammography can detect a cancerous mass. Therefore, there might be opportunities for early intervention to reverse breast cancer development. Due to these properties, thermography was used as a breast cancer screening tool as early as 1956 in the US. It was accepted widely by medical professionals at that time (Kennedy et al., 2009).

1.3.2 Research

The publication by Feig et al. (1977) compared different breast cancer detection methods. In this study, thermography came out as the least effective method, with a sensitivity of 39%, which is unacceptable for population-wide screening. The study has been widely criticized for its inadequate quality control of the thermography procedures and interpretation (Keyserlingk et al., 2000). For example, thermography was conducted and interpreted by untrained technicians without clearly established quantitative guidelines, while mammograms were analyzed by experienced radiologists. Despite inadequate quality control, thermography was subjected to the same statistical analysis as other methods and was dropped from any further consideration as a stand-alone screening method. Since then, both the quality and resolution of the infrared cameras and the standardized protocols have improved significantly, leading to a new spike of interest to the field of thermography.

In Yao et al. (2014), two radiologists analyzed thermograms of 2036 females (480 with breast cancer) and got 84% sensitivity and 94% specificity. This performance was comparable to two other methods they tested — mammography and ultrasonography. Another group of scientists (Parisky et al., 2003) assessed the effectiveness of thermography to evaluate mammographically suspicious lesions in 875 patients. They found thermography to have 97% sensitivity, but only 14% specificity. Comparable results were obtained on smaller sample sizes. Francis et al. (2014) got 82% sensitivity with 100% specificity using thermograms of 22 women. Tan et al. (2007) achieved a 100% sensitivity with 60% specificity on a sample of 78 patients.

Even though the results look quite promising, it is worth noting that thermography research has not been driven by sufficiently high standards. Some peer-reviewed publications have apparent experimental design flows that undermine the validity and generalizability of their findings. For example, the model presented in Silva et al. (2016) was tuned through cross-validation on the same set that was used for evaluation. As a result of this data leakage, the author achieved a suspiciously high sensitivity of 100% on a set of 80 patients, which is unlikely to be the same for real-world data. Other researchers evaluated artificially balanced sets that do not correspond to

the real occurrence rate of breast cancer, used very small datasets or reported accuracy, which is an unreliable metric for imbalanced datasets as discussed in the Data section. In this paper, we make sure to avoid such experimental design flows to ensure the validity of our findings.

1.3.3 Limitations

Despite all the advantages of thermography, it is not an ideal cancer detection method. Cancer is not the only condition that can cause a shift in thermal patterns. Some benign breast diseases, shift in hormones, and even colds can have a similar effect. Therefore, the number of false positives can be high compared to other methods. False positives can cause significant financial and emotional stress to incorrectly diagnosed patients. Therefore, thermography might be more suitable for developing countries, where there are no countrywide early detection programs, and many people are diagnosed at later stages (Garduño-Ramón et al., 2017).

Some other limitations of thermography, as stated in Kennedy et al. (2009), are its inability to identify "cold" tumors with low metabolic activity and microcalcifications. It is notable that "cold" tumors also tend to be less aggressive since they need a longer time to grow, in some cases, up to 5 years (Nakashima et al., 2019). Thus, it is less critical, while still undesirable, to miss them during the initial screening.

2 Data

This project is written in association with Eva Tech. Eva Tech is a female health technology company that offers AI-powered solutions to early breast cancer detection. The company runs multiple FDA-approved thermography clinics around Mexico that gather screening data from the general population. The team also focuses on running clinical trials at cancer centers and nonprofit organizations with a high number of confirmed breast cancer cases to ensure that models can learn from the positive samples. With hundreds of patients coming to Eva clinics and participating in clinical trials every week, the company has, to our knowledge, the biggest thermography database ever created.

In this project, we use a sample of 1861 unique observations gathered from different sources, including Eva Centers and clinical trials. The proportion of positive cases is 6%, which includes both histologically confirmed and highly suspicious of malignancy cases. We split these data into train, test, and validation tests and use k-fold cross-validation to ensure that we do not overfit to a single test set and exclude any opportunity for data leakage during feature selection and hyperparameter tuning. The test data must not be used in any way to make choices about the models. Otherwise, we are at risk of getting overly optimistic outcomes that do not generalize well.

2.1 Preprocessing

The first step towards the automatization of thermography interpretation is breast segmentation. At Eva, it refers to identifying relevant sections of thermal images that correspond to each breast, axillary region, and nipple area. By performing segmentation, we ensure that irrelevant segments of the image, such as abdomen and background, are excluded from the analysis.

We perform the segmentation using the ResNet architecture to minimize the problem of the vanishing gradient (repeated multiplication may make the gradient very small during backpropagation in deep networks.) As a result, its performance gets saturated or even starts degrading. The authors of the ResNet solve this problem by reusing activations from a previous layer until the adjacent layer learns its weights, a concept known as *skip connections* (He et al., 2015). The ResNet that we use has 34 layers and can be easily fine-tuned to work with our images that are quite different from the ImageNet dataset.

After cropping out the segmented areas and setting all the background values to 0, the images go through the pre-processing steps that include contrast enhancement and the application of the CLAHE algorithm that highlights local temperature differences. The combination of these techniques not only improves the performance of final models but also contributes to domain adaptation. We use T-distributed Stochastic Neighbor Embedding (t-SNE) to ensure that multiple centers do not introduce bias to the model. This method is a nonlinear dimensionality reduction

technique designed for embedding high-dimensional data for visualization in a low-dimensional space. It represents each high-dimensional observation using a two-dimensional point, where nearby points model similar observations.

Our test consists of creating t-SNE plots for multiple parameter values and determining whether exams cluster by some factor that is not the cancer status. A problematic t-SNE plot would reveal clusters by the confounding factor. For example, clusters formed by the center would mean that the model can discern the origin of the data based on the extracted features. This would be a problematic source of bias since some centers have a much higher proportion of cancer cases. Thus, we ensure that pre-processing steps homogenize images that come from different sources and prevent models from learning biased representations. The Anomaly Detection section presents an example of such a plot.

Final pre-processed explorations consist of four images: lateral and frontal thermograms of right and left breasts. We rescale the images to the size of 96×96 pixels. In the figure below, you can see an example exploration of a random Eva patient. Red spots correspond to hotter areas of the breast that can be indicative of physiological and pathological angiogenesis:

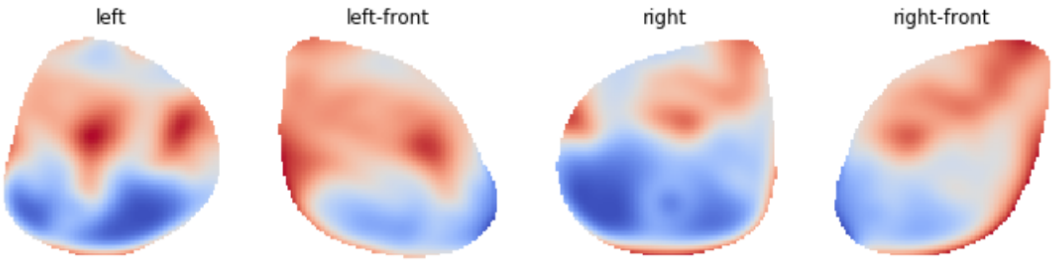

Figure 1: Four thermal images that correspond to the frontal and lateral views of left and right breasts.

2.2 BI-RADS

Breast cancers can have different histological types, clinical features, tumor markers, and stages (Li et al., 2005). Some benign conditions look like cancers, and the status of the patient is not always known until they get a biopsy. Thus, binary labels are not ideal for data labeling during the initial screening phase. To address this issue, the American College of Radiology devised a classification called Breast Imaging Reporting and Data System (BI-RADS.) BI-RADS is a quality assurance tool designed for use with mammography, but it is also applicable for thermal images. It classifies breast imagery into seven distinct classes, as shown in Table 1.

BI-RADS	**Status**
0	Incomplete
1	Negative
2	Benign
3	Probably benign
4	Suspicious
5	Highly suggestive of malignancy
6	Proven malignancy, known biopsy

Table 1: BI-RADS Categories.

The ultimate goal of an early breast cancer detection model is to identify patients with a high risk of breast cancer. Ideally, the ground truth data should correspond to the gold standard for breast cancer detection, which involves a screening mammogram followed by a biopsy upon suspicious mass identification. It is desirable that all patients follow this procedure so that we

know for sure whether they have cancer (BI-RADS 6) or no cancer (BI-RADS 1 and 2.) However, as we do not always have immediate access to test results, a considerable number of patients have BI-RADS 3, 4, or 5. Each of these results has varying degrees of uncertainty. Patients with BI-RADS 3 are often suggested to wait six months to get an additional mammogram since their risk of cancer is only about 5%. Patients with BI-RADS 4 and 5 are typically advised to go through further tests immediately since there is a higher risk of breast cancer. The goal is of the initial screening is to classify BI-RADS 4-6 as abnormal to ensure that we do not miss any cancer cases.

2.3 Metrics

A dataset is considered to be *imbalanced* if one class contains significantly more samples than the other. This problem is widespread in disease diagnostics since only a small proportion of the general population are affected by the disease. With imbalanced datasets, the ability of a classifier to distinguish members of the minority class often degrades. Models attempt to minimize the cost function, and if errors in both classes are treated equally, they develop a bias towards the majority class. This is fine in case both classes have equal importance. However, it is the minority class that is usually of primary interest. For cancer detection, the cost of false negatives is very high. It is generally better to perform some additional tests on a healthy patient than overlook a cancer case.

Despite being one of the most popular metrics, accuracy is inadequate for breast cancer detection. It is defined as the number of correctly classified cases divided by the total number of cases. Thus, a model that always predicts "healthy" has very high accuracy. Unfortunately, high accuracy scores do not translate into prediction reliability.

Two metrics that are commonly reported in breast cancer detection literature are *sensitivity* and *specificity*. While sensitivity, also known as *recall*, represents the proportion of positive cases that are correctly identified, specificity describes the proportion of correctly identified negative cases. Most screening methods want to minimize specificity, ensuring that only a small fraction of healthy patients get incorrectly classified as affected. While high specificity is essential to avoid sending too many healthy patients for additional tests, in cancer screening scenario, it is even more important not to miss any actual cancer cases. *Precision* is a useful metric that describes the model's ability to identify only relevant data points. For example, if we classify all patients as sick, the sensitivity is 100%. However, the precision of such a model would not be particularly impressive.

In general, we prefer models with high precision and recall scores. However, there is a trade-off between these two metrics: an increase in precision often results in a decrease in the recall and vice versa. Ideally, we need a single number performance summary that includes information about different metrics. The F1-score is a harmonic mean of the sensitivity and specificity that is often used for this purpose. F1-score's main flaw is that it gives equal weight to precision and recall. Classifying a sick person as healthy has a different cost from classifying a healthy person as sick, and this should be reflected in the metric. Thus, it is better to pick a different metric that can maximize metrics of interest by penalizing false negatives and false positives proportional to the real-life costs of these errors.

The Area Under a Curve (AUC) metric is an intuitive metric that illustrates the best trade-off between sensitivity and specificity. Graphically, model performance for different classification thresholds can be represented with the help of the receiver operating characteristic curve (ROC), which plots a true positive rate on the Y-axis and a false positive rate on the X-axis. The ROC-AUC metric provides the flexibility of adjusting the classification threshold for a particular context, such as an oncology center with many cancer patients or a regional hospital that performs annual screenings of the general public. Thus, we use the AUC metric paired with ROC plots and recall and precision scores for the model evaluation.

3 Human Evaluation

One crucial question to ask when evaluating a machine learning model for medical imagery interpretation is whether it exceeds human performance. In the case of thermography, doctors

and researchers usually rely on subjective reasoning since there are no internationally recognized standards/training programs. The establishment of a formal approach will not only benefit the medical community but also provide useful insights into feature engineering and model architecture decisions for automated models. This section takes the first steps towards defining such standards.

3.1 Quantitative Methods

Several studies have proposed quantitative thermography evaluation methods. González (2011) introduced a thermal score to analyze infrared images quantitatively. This score is defined as the sum of the vascularity level and the temperature difference in °C at the lesion site compared to the contralateral breast. The author obtained a significant correlation between thermal score and tumor size in González (2011).

Keyserlingk et al. (2000) designed a technique that is based on assigning low numerical values to cases with absent or moderate symmetrical vascular patterns and higher scores to samples with multiple "abnormal signs", as presented in Figure 2.

Abnormal Signs
1) Significant vascular asymmetry.*
2) Vascular anarchy consisting of unusual tortuous or serpiginous vessels that form clusters, loops, abnormal arborization, or aberrant patterns.*
3) A 1°C focal increase in temperature (ΔT) when compared to the contralateral site when associated with the area of clinical abnormality.*
4) A 2°C focal ΔT versus the contralateral site.*
5) A 3°C focal ΔT versus the rest of the ipsilateral breast when not present on the contralateral site.*
6) Global breast ΔT of 1.5°C versus the contralateral breast.*
Infrared Scale
IR1 = Absence of any vascular pattern to mild vascular symmetry
IR2 = Significant but symmetrical vascular pattern to moderate vascular asymmetry, particularly if similar to prior imaging
IR3 = One abnormal sign
IR4 = Two abnormal signs
IR5 = Three abnormal signs
*Unless stable on serial imaging or due to known noncancer causes (e.g., abscess or recent surgery).

Figure 2: Keyserlingk Infrared Grading Scale. Source: Keyserlingk et al. (2000)

The authors showed that an addition of quantitative thermography evaluation had a statistically significant positive effect on performance, compared to mammography alone.

3.2 Method Comparison

We obtained data from two independent evaluators who received training in infrared thermography. The evaluators scored a subset of 114 thermograms (50% cancer cases) based on the criteria listed in Appendix A. The evaluators performed the task separately and did not have access to the ground truth and patients' medical records.

We took the average of two evaluations to increase robustness and calculated the Gonzalez and Keyserlingk scores as defined in the corresponding papers[1]. Unlike subjective evaluations that are often used to interpret thermograms, quantitative scores can be easily calculated and are less interpreter-dependent. However, their effectiveness has not been evaluated against the subjective approach. Thus, the interpreters were also asked to give their judgment on the thermogram malignancy on the scale from 1 (*clearly normal*) to 10 (*clearly abnormal*). This score aims to identify whether proposed quantitative methods outperform an utterly subjective evaluation.

To make a comparison, we normalized the values of three scores from 0 to 1 and calculated the Area Under a Curve (AUC) score for each of them:

[1]The implementation details are given in the `human_evaluation` notebook.

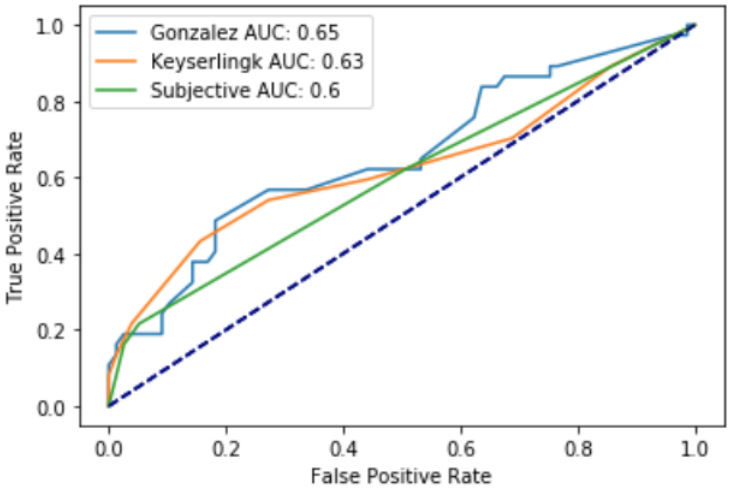

Figure 3: ROC curves for human evaluation scores. The dotted line represents a baseline with no predictive ability.

The ROC curves indicate that quantitative scores have a better predictive power compared to subjective evaluations, at least for this set of patients. Therefore, it should be possible to use such quantitative measures as a baseline for automated methods and to train new interpreters.

The score from González (2011) has a higher AUC than the Keyserlingk score. A potential reason for it is that the Keyserlingk score is based on the quantity of abnormal thermal signs, while the Gonzales score also takes into account the magnitude of these abnormalities. Thus, it provides a more precise understanding of the patient's clinical picture, especially in extreme cases of a significant temperature difference in two breasts. That is why we use the Gonzalez score as a baseline for automated models.

Since it would be unfair to compare only a small subset of balanced data to the performance of machine learning models on the entire dataset, we automated the thermal score from González (2011). The details of the implementation follow in the next section.

4 Feature Engineering

In recent years, deep learning approaches have gained popularity across multiple domains. In theory, neural networks of sufficient depth can approximate any continuous function. They learn features from high-dimensional data, which eliminates the need for feature extraction. Several researchers have applied these novel approaches to thermography images. For example, Koay et al. (2004) claim to achieve reasonable performance using deep learning networks on a sample of 19 thermograms. While it is crucial to explore new opportunities in the field, results presented in the paper are likely to be obtained due to chance, overfitting, or data leakage. Deep learning relies heavily on the sample size, which restricts its practical application in problems with a limited number of labeled samples from each class. One should not expect a model with tens of thousands of parameters to be tuned with a disproportionally small number of positive cases and still generalize well.

Once we gather thousands of positive cases, it will become feasible to use deep learning models to classify thermal images. In the meantime, we need to rely on less complex alternatives. These alternatives are models from classical machine learning, such as logistic regression, random forest, support vector machine, and ensembles of models like these.

The objective of this section is to extract meaningful, low-dimensional representations of the data. These representations can be later used by traditional machine learning models to ensure the generalizability of results.

4.1 Histogram of Oriented Gradients

The Histogram of Oriented Gradients (HOG) is a feature descriptor commonly used in computer vision. This method was a state of the art way of doing object detection before the deep learning era and is still used to extract features from small datasets.

The logic behind the HOGs is that the distribution of intensity gradients can represent shapes within an image. Abnormal shapes and dramatic shifts in intensity can be signs of pathological changes in breast tissue and, therefore, can be useful for breast cancer detection.

Raghavendra et al. (2016) extracted HOGs from a balanced set of 50 patients and demonstrated that this approach, in combination with dimensionality reduction techniques, leads to a significant improvement in several metrics compared to other feature extraction approaches. Ergin and Kilinc (2014) obtained similar findings and proposed a HOG-based computer-aided diagnosis framework to aid radiologists. In this section, we partially replicate the approach used in Raghavendra et al. (2016) and experiment with different dimensionality reduction techniques to achieve better results.

4.1.1 Implementation

The algorithm was first presented in Dalal and Triggs (2005). It contains the following steps:

1. Normalize target images using the power low compression.
2. Calculate first-order gradients to capture edges and global texture information.
3. Divide images into small regions, called cells, to capture local image content.
4. Calculate a histogram of gradient directions for each pixel in the cell. Each histogram divides the gradient angle range into a fixed number of bins.
5. Calculate the intensity across a broader region of the image, a block, and normalize all cells within this block. This normalization results in improved invariance to illumination, shadowing, and contrast, which is essential since Eva clinics have different cameras and lighting. Each cell can appear in several blocks, but since normalization is block-dependent, they are all different. Thus, a single cell appears in the final output vector multiple times.
6. Combine the HOGs from all blocks into a feature vector.

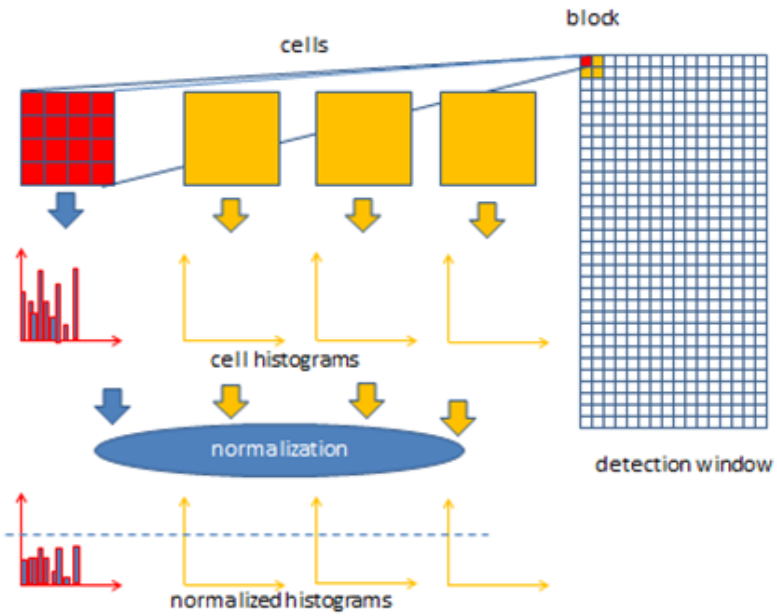

Figure 4: A schematic representation of the HOG algorithm. Source: Intel (2019)

The values of parameters should be carefully selected so that they are better suited to Eva images while producing a descriptive yet not too large feature vector. First of all, pixels per cell and cells per block should be divisors of the image shape. Since pre-processed Eva images have a dimension of 96×96 pixels, we decided to have 4×4 cells per block and 12×12 pixels per cell. We also calculated the gradients in four orientations from 0 to 360 degrees to ensure that the result is rotation invariant. The resulting feature vector has a length of 6,400. Figure 5 presents a visualization of this feature vector. Blue color indicates a homogeneous surface, while red indicates a significant change in image intensity. Similarly, rotated shapes indicate that there is a change in the gradient direction:

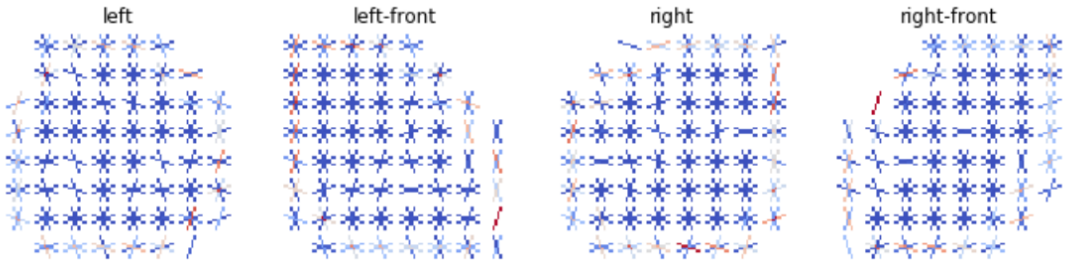

Figure 5: A visualization of the Histogram of Oriented Gradients.

4.1.2 Dimensionality Reduction Techniques

The HOG feature descriptor yields highly-dimensional vectors. Processing such data for thousands of patients is computationally intensive. Additionally, these vectors have redundant features that add no relevant information because they are correlated or only contain zeros. Thus, dimensionality reduction techniques are necessary to reduce training times and minimize data redundancy (Raghavendra et al., 2016).

Many dimensionality reduction methods have been proposed in the medical imagery literature. We compared three techniques — Kernel Principal Component Analysis (KPCA), Independent Component Analysis (ICA), and Locality Preserving Projections (LPP):

- KPCA is an extension of the most popular dimensionality reduction method — PCA. KPCA uses kernels to project data into a higher dimensional feature space, where it is linearly separable. The resulting components are uncorrelated and ordered in such a way that the first principal components explain most of the variance of the original data.
- ICA has been effectively used in the electroencephalogram (EEG) analysis and has been shown to outperform the classical PCA approach (Bugli & Lambert, 2007). ICA does not require a Gaussian distribution of the input data and can describe localized shape variations. The second property is specifically useful for malignancy identification because cancer manifests itself through changes in local breast structures.
- LPPs are linear projective maps that are obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Beltrami operator on the manifold. Raghavendra et al. (2016) found that this method was the most effective for the HOG dimensionality reduction in thermograms.

Unlike PCA, the components obtained using ICA and LPP are not ordered. Thus, we used a t-test value-based ranking technique proposed in Raghavendra et al. (2016) to select relevant components. The idea behind this method is to calculate the t-values and p-values for normal and malignant samples. The features with low p-values ($p < 0.05$) and high t-values are preferred for classification.

After implementing the methods and ordering the ICA and LPP components, we tested the performance of the low-dimensional feature vectors using the vanilla logistic regression model and

a validation set. We gradually increased the number of components to find an optimal number of components for each method. The implementation details can be found in the `ML_pipeline.ipynb` notebook of the GitHub repository. We found that KPCA consistently outperformed other methods independent of the kernel, number of components, and the train-validation split. Thus, we picked this approach to produce the final vector.

The resulting HOG feature vector consists of 30 features that contain information on essential components of the original HOG vector and can be used to train machine learning models.

4.2 Local Binary Patterns

Local Binary Patterns (LBPs) are a texture descriptor popularized by Ojala et al. (2002). The value of this feature vector in cancer detection is that it helps to identify the hottest areas of the breast that might indicate a pathological increase in the temperature. It also highlights areas of increased vascularity that are indicative of abnormality.

This method is very common in the breast cancer detection literature and has been used both with thermographic (Abdel-Nasser et al., 2019) and mammographic (Pereira et al., 2014) imagery. It has also been applied for vascular areas segmentation (Li et al., 2005) and for classification (Li et al., 2005), (Abdel-Nasser et al., 2019), (Pereira et al., 2014).

Choi et al. (2012) classified mammograms as either breast masses or healthy tissue, using the rotation-invariant LBPs approach that we explain below. The experiments were performed using 303 images. The reported AUC score was around 90%. Li et al. (2005) proposed a method for mass segmentation and detection using a rule-based algorithm called MCL (multiple concentric layers). They trained the model on 125 images and achieved the average recall of 76.8%. The main criticism of this approach is that MCL is empirically optimized and, thus, might not generalize well. Another study by Abdel-Nasser et al. (2019) reports the AUC score of 99% on a sample of 50 patients that they used for training, validation, and test. Though we might not obtain the same performance on a more massive dataset, these findings make a strong case for the use of LBPs for automatic detection. We aim to extract this feature vector from Eva data and see whether the method performs well on real-world breast cancer patient data.

4.2.1 Implementation

The idea behind the LBP operator is quite simple: it replaces pixels of the image with numbers that encode the local structure around each pixel. Specifically, we compare each central pixel with its eight neighbors. If the value of the neighboring pixel is smaller than that of the central pixel, it gets a value of 0. If the value is equal to or greater than that of the central pixel, it receives the value of 1. After that, we generate a binary number for each of the central pixels by concatenating these binary bits. The resulting decimal value replaces the original central pixel value. Figure 6 illustrates the first stage of this process:

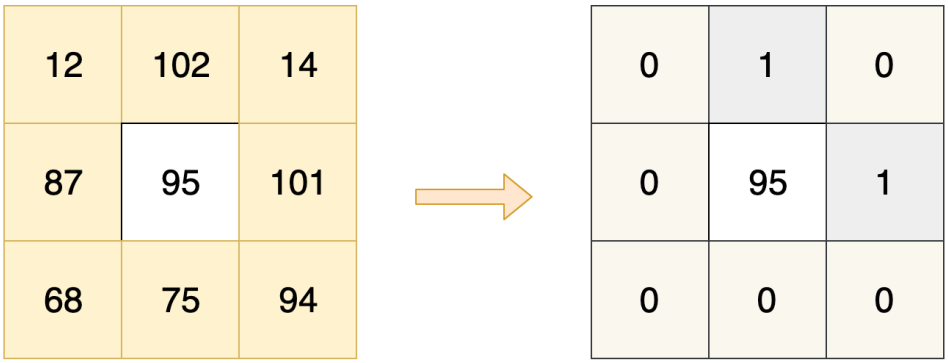

Figure 6: The calculation of the LBP of the central pixel.

While the total number of combinations equals 2^n, where n is the number of neighbors, there are not as many unique patterns to consider. The patterns that consist of all 0s (or 1s) are rotation-invariant because they remain the same no matter what pixel we start to concatenate from. However, this does not apply to other patterns. Thus, rotation invariance is required to ensure that the extracted feature vector is robust to some classical augmentations such as rotations and flips. The authors of the original paper present 36 unique rotation-invariant patterns:

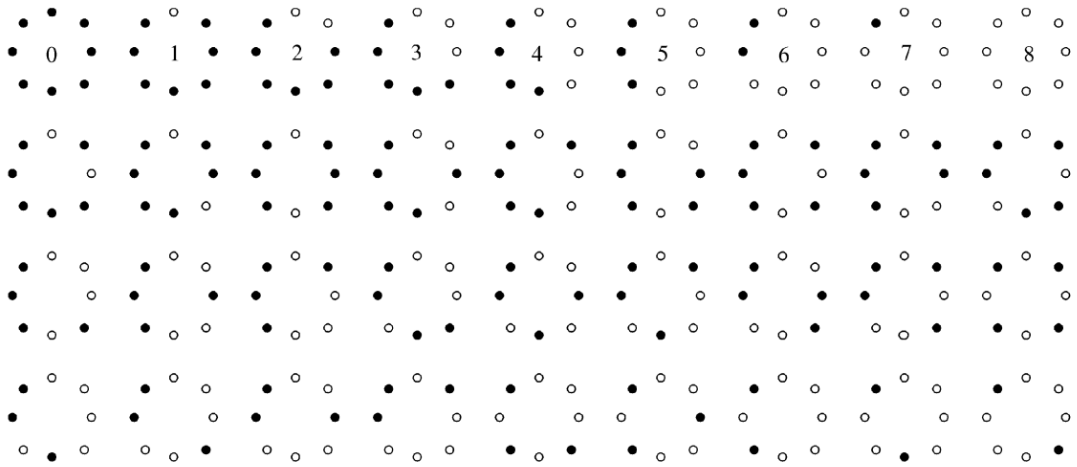

Figure 7: 36 unique rotation-invariant LBPs. Black and white circles correspond to 0s and 1s. Source: Ojala et al. (2002)

Ojala et al. (2002) have observed that some of these patterns represent fundamental properties of the texture much better than others. They called these LBPs "uniform" because they have a uniform circular structure that contains very few spatial transitions. Such uniform patterns are shown in the first row of Figure 7. They serve as templates for texture features such as bright spots (0), flat areas or dark spots (8), and edges of varying curvature (1-7) (Ojala et al., 2002).

The authors formally defined these "uniform" patterns by introducing a uniformity measure that corresponds to the number of spatial transitions between 1s and 0s. If this number is less or equal to 2, the pattern is considered uniform. That definition makes intuitive sense because multiple transitions indicate that the environment of the pixel is heterogeneous, and there is no significant change in the texture that can be used to analyze the image. Additionally, the relative proportion of such patterns is so small that their probabilities cannot be reliably estimated. Therefore, the resulting feature histogram is a histogram with 9 bins, where the first 8 bins are fundamental uniform patterns with the corresponding number of 1s, and the ninth bin is a combination of all the non-uniform patterns. The figure below shows how this feature vector highlights the vascular patterns of the breast. Red dots and lines highlight vascular areas of increased temperature. Blue lines represent areas of hypothermia, which are less relevant for breast cancer detection but might be a sign of other conditions such as breast trauma or lipomas (Piana & Sepper, 2015).

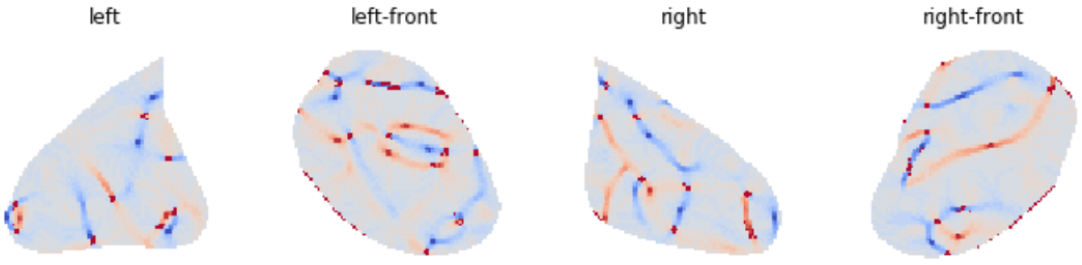

Figure 8: The visualization of LBPs.

The final vector consists of 9 features (histogram bins) that represent the difference in the uniform patterns of the left and right breasts. This approach helps to highlight vascular asymmetry, which is one of the leading indicators of breast cancer.

4.3 Automatic Thermal Score Extraction

A thermal score is the sum of vascularity and delta-T. In the Human Evaluation section, we showed that this method outperformed completely subjective evaluation as well as an alternative quantitative score. The following paragraphs describe how we can automatize this method with the help of computer vision techniques.

4.3.1 Delta-Ts

The delta-T function is introduced in the paper by González (2011). It is defined as the temperature difference between the hottest point in the asymmetrical vascular region and its corresponding location on the opposite breast.

Theoretically, the automatic extraction of this difference can be accomplished by identifying the hottest pixel on the image. However, the edge of the breast is usually hotter than the center because it has direct contact with the body. Therefore, such a function would select physiologically hotter places that are not indicative of the malignant activity.

The solution to this problem is LBPs that were presented above. The LBP value of 0 corresponds to the hottest area of the breast because it means that the pixel does not have any hotter neighbors. Since the hotter edges of the breast are quite homogeneous (they are equally affected by the proximity of the patient's body), they cannot have the LBP value of 0 and, therefore, cannot be considered as candidates for delta-T extraction.

Thus, the calculation of delta-Ts turns into a simple matter of mapping all the 0 LBP values to their corresponding temperatures and selecting the hottest spot. After that, the function identifies the corresponding spot on the other breast and determines the temperature difference. The image below illustrates the difference between the naive and LBP-based approaches:

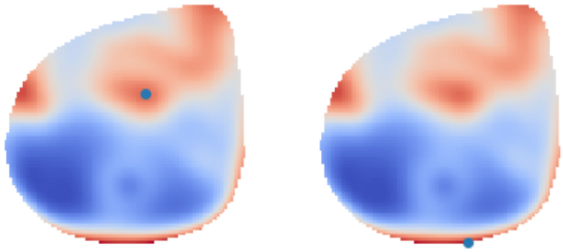

Figure 9: The difference between the delta-T feature (left) and the maximum temperature approach that is prone to edge selection (right). The hottest area is marked by the blue dot.

We visually examined the resulting hot spots on a total of 500 raw and processed images to confirm that the LBP approach eliminates the edge detection problem. The proposed approach is also robust to the change in the LBP parameters and varying image pre-processing techniques.

4.3.2 Vascularity Score

The second component of the thermal score is vascularity. It is determined using the following scale (González, 2011):

1. Absence of vascular patterns.
2. Symmetrical or moderate vascular patterns.
3. Significant vascular asymmetry.

4. Extended vascular asymmetry in at least one-third of the breast.

While humans can quickly evaluate the extent of the vascular asymmetry, it is a relatively hard task for a computer. One approach is to use segmentation algorithms to identify hot areas of the breast. Some methods, such as the Chan-Vese algorithm (Chan & Vese, 1999) and the Active Contour method (Kass et al., 1988), do not always identify the hottest areas of the breast due to their unsupervised nature. Therefore, we decided to use a semi-supervised method called Random Walker.

In the Random Walker algorithm, users need to mark seed regions: areas of the highest and lowest temperatures. For every unlabelled pixel, the algorithm initializes a random walker that can go anywhere in the image. Then, it calculates the probability that each random walker reaches one of the labeled pixels. By assigning each pixel to the class with the highest probability, it is possible to obtain a segmentation of good quality (Grady, 2006).

To account for the underlying image texture and do not produce the same segmentation for different images with the same seed, the authors added a Gaussian weighting function given by:

$$w_{ij} = \exp\left(-\beta(g_i - g_j)^2\right)$$

where g_i indicates the image intensity at pixel i. The value of β is a free parameter that the authors kept at a constant value of 900. By applying this method to Eva data, we managed to segment the areas of high vascularity:

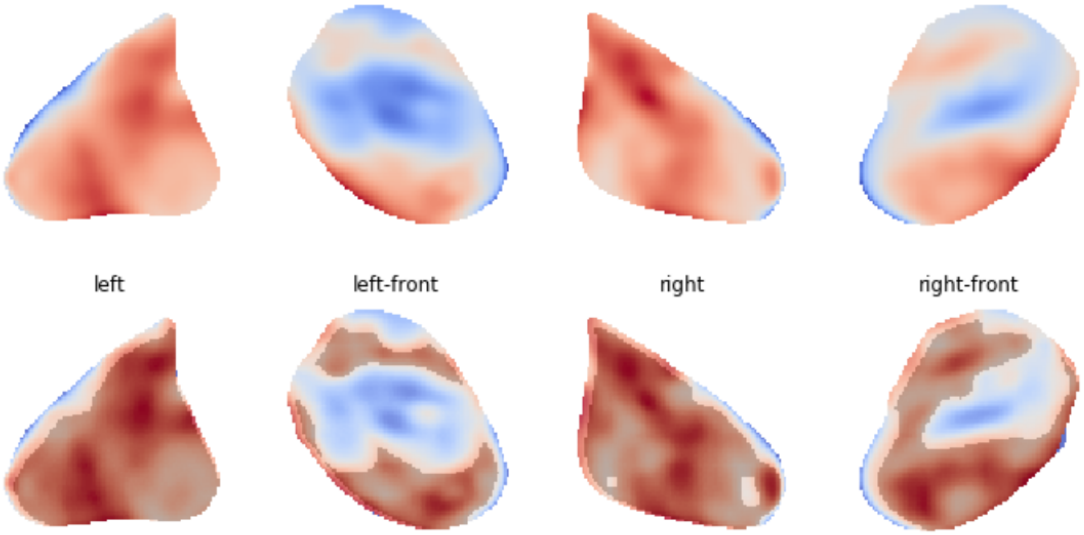

Figure 10: The outcome of the Random Walker segmentation. The darker areas in the second row indicate the areas of increased vascularity.

We designed an algorithm that calculates the vascularity score based on three main principles. Firstly, we optimized our approach by increasing the correlation of the automated score with 114 manual evaluations to ensure their similarity. Secondly, we adjusted the values to improve the correspondence to the natural occurrence of different scores in the population with the help of ground truth values. Lastly, we reached out to Javier Gonzalez, the author of the thermal score paper, to manually inspect several automatic evaluations and confirm that our understanding of the logic behind the score is correct.

Vascularity and delta-T are necessary components for the Gonzalez score automatization. The automatization of the score makes it possible to evaluate a vast number of thermographies in a short time. It is important to note that the thermograms selected for manual evaluations in the Human Evaluation section present clear healthy and cancer cases. Therefore, it might be impossible to get the same performance on more diverse and less clear cases. Additionally, a very high correlation to the manual evaluations might be indicative of overfitting. Therefore, in the Preliminary Results section, we evaluate the performance of this score on unseen data to ensure its generalizability.

4.4 Feature Selection

HOG and LBP vectors are difficult to interpret, which makes the current model a black box that outputs a prediction without giving any insights into individual feature contributions. To understand how each of the newly engineered features contributes to the model performance and identify a combination of features that result in the best performance, we used SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). This approach is based on Shapley's values — a game-theoretical concept that represents the average of the marginal contributions across all permutations. Compared to other approaches, such as the recursive variance-based feature elimination approach implemented in `scikit-learn`, SHAP has two main advantages:

- *Global interpretability* — SHAP values can show not only the strength of the correlation of each predictor to the target variable but also whether this relationship is positive or negative.
- *Local interpretability* — each observation gets its own set of SHAP values. Thus, it is possible to explain why a thermogram receives a certain score and analyze the contributions of individual predictors.

We perform the feature selection process on a separate subset than the one we used for the final model evaluation. Otherwise, we would be overfitting to that specific subset and potentially get overly optimistic results.

The plot below sorts features by the sum of SHAP value magnitudes for all observations and shows the distribution of impacts each feature has on the model prediction. More formally, the estimated SHAP value is the contribution of a specific feature value to the difference between the actual prediction and the mean prediction. For example, the plot below illustrates that high values of the HOG12 feature increase the risk of breast cancer while HOG26 decreases it.

Interestingly enough, high values of the delta-T feature do not seem to increase the risk of breast cancer, as suggested in González (2011). Instead, a higher temperature difference seems to decrease one's risk of being sick. A potential explanation for this observation is that many women have physiological differences in the temperature of their breasts that are not necessarily caused by breast cancer. Thus, taking it as a primary factor in the human evaluation method might be misleading. It is worth noting that lower values do not seem to have any effect on the model output. A further study is required to investigate the nature of this phenomenon:

Figure 11: Top-10 SHAP values and their effect on model output.

We iteratively eliminated features from the bottom of the SHAP list while optimizing for the AUC score to pick a set of features that maximizes model performance. While the SHAP values and the iterative elimination process provide insights into what a good feature set should be, it is a process that cannot account for the interaction factors between features that might play a significant effect on the model performance. We decided to stick to this method because it is much less computationally expensive compared to the exhaustive search over all possible subsets, is easy to interpret, and can be performed on a large scale.

4.5 Model Selection

In the context of machine learning, model selection can have different meanings. Firstly, we might refer to selecting the best algorithm from a set of machine learning methods. For instance, we might wonder whether a logistic regression model or a support vector machine classifier yields the best classification performance. Secondly, we might be interested in selecting the best hyperparameters for the algorithm we picked. Unlike regular parameters that are optimized during the model fit (model coefficients), hyperparameters are the parameters that we have to specify before fitting the model. For example, we need to specify the regularization strength and type in the logistic regression before the training. Finding the optimal model and corresponding hyperparameters is crucial for model performance. Thus, we do it in this section.

4.5.1 Algorithm Selection

It is essential to take the bias-variance trade-off into account to select algorithms with the best predictive power. Bias refers to simplifying model assumptions, which cause the model to underfit the data. A high-bias model cannot adequately capture the structure of the data. On the other hand, variance indicates how much the model changes as we modify the training data. A high-variance model is susceptible to small fluctuations in the training data, which causes the model to overfit. Simple models are usually high-bias, low-variance, while more complex models tend to be low-bias, high-variance. The goal is to find models that are not too simple, e.g., linear regression that cannot model nonlinear relationships, but also not too complex, e.g., deep neural networks that would overfit to the limited training data we have.

Thus, after analyzing a variety of `scikit-learn` models, we picked several candidate models that seem to have a reasonable bias-variance trade-off. We trained these models on the training set and, to avoid overfitting and data leakage, evaluated their performance on the validation set. The resulting performance is presented in the graph below:

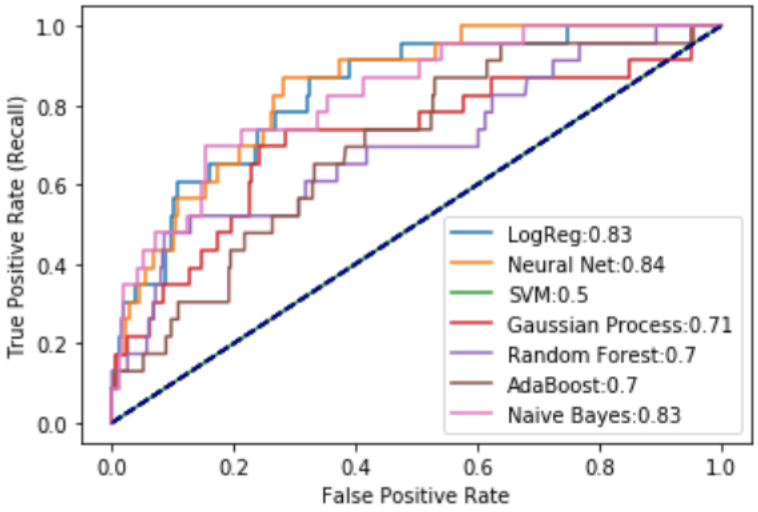

Figure 12: Candidate models performance.

The figure indicates that all models have a better-than-random performance on the validation

set. Our goal is to ensure that no matter what manipulations we perform on the data or models, the performance on the unseen data is still above the blue dashed line and ideally above the human performance.

4.5.2 Hyperparameter Tuning

To minimize the number of computational resources that are required to perform hyperparameter tuning, we picked three candidate models from the previous step: Logistic Regression (LR), Naive Bayes (NB), and Neural Network, or Multilayer Perceptron (MLP). The NB model does not have any hyperparameters, so there is no need to perform tuning for this classifier. For the other models, we implemented K-fold cross-validation (CV) using the `GridSearchCV` function from the `scikit-learn` library. This approach exhaustively considers all parameter combinations specified by the user. We decided to use this exhaustive method since we only tune two models that have a relatively small number of hyperparameters, which makes this calculation feasible and not too time-consuming. Cross-validation works as follows:

1. Split the training set into K smaller sets and make sure that the proportion of the positive cases is similar in all splits.

2. Set aside each of the K folds one time. Train as many models as there are different combinations of model hyperparameters on the remaining K-1 folds and compute the validation score on the hold-out fold. We decided to optimize for recall and precision values separately and see which approach leads to better results.

3. Compute the mean validation score for each set of hyperparameters and select the best set.

4. Train the model with the chosen hyperparameter set on the full training set and estimate the generalization error using the test set.

By following this simple procedure, we identified the best-performing algorithm in terms of sensitivity — MLP. We decided to optimize for sensitivity since our primary goal is to make sure no cancer cases go unattended. Other models had a perfect specificity score, but also much lower sensitivity values. It is worth noting that high recall (sensitivity) values always come at the expense of precision. Thus, a single "perfect" model does not exist. Instead, we need to optimize the values of interest, usually at the expense of other metrics.

Model	Sensitivity	Specificity	Precision
LogReg	0.09	1.0	0.67
NB	0.09	1.0	1.0
MLP	0.22	0.94	0.2

Table 2: The performance of tuned models on the validation set.

The manual inspection of the optimal hyperparameters indicated that the best performing model has three layers with 50, 100, and 50 nodes. These numbers seem to be reasonable since the model is not overly complex, as some modern neural network architectures. However, it can grasp the structure of the data better than a regular logistic regression model. Thus, the model satisfies the bias-variance constraint that we specified at the beginning of this section.

4.6 Preliminary Results

Finally, we want to evaluate model performance by estimating the generalization error of the selected classifier on unseen data. A good machine learning model should not only perform well on the training data (simple memorization is enough to achieve excellent results in this case) but also on unseen data. Thus, we should be sure that the model's performance does not degrade when confronted with new data. Additionally, we want to evaluate the performance of the automated Gonzalez score on unseen data and compare it to the metrics obtained from human interpreters.

The graph below illustrates some of the preliminary findings. Despite being lower than the values obtained during the hyperparameter tuning, the AUC score calculated on the test set is still decent (71%.) New features contain HOG and LBP feature vectors selected using iterative elimination of SHAP values as well as the thermal score features that seem to improve performance. "Thermal score" stands for the automated human evaluation score. It has the AUC score of around 60%, which is comparable to the value obtained from human interpreters and proves that the score generalizes well to bigger datasets and can be used as a baseline for the machine learning approaches.

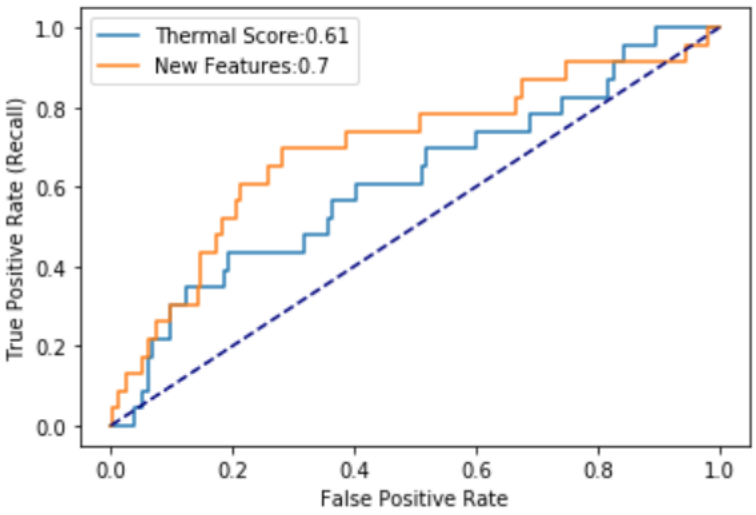

Figure 13: The ROC curve for different feature vectors.

Despite the high AUC score, a further investigation of the model performance uncovered that the sensitivity of the model is currently at 9%, which means that the model misclassifies many cancer cases as healthy. The automated score does not suffer from this problem as much with the recall value of 35%. Thus, our model performs worse than the baseline in terms of false negatives.

A potential explanation might be the lack of positive samples the model can learn from since the training set had only 6% of positive cases. This is a usual problem for imbalanced datasets. The next section presents data augmentation methods that deal with this issue. The motivation behind using data augmentation methods before feeding the dataset into the model is that classifiers are more sensitive to detecting the majority class and less sensitive to the minority class. Thus, if we do not take care of the issue, the classification output would remain biased, in many cases resulting in always predicting the majority class.

5 Data Augmentation

Data augmentation is a strategy that enables to increase the diversity of data available for training without collecting new data. This section explores three data augmentation techniques: anomaly detection, undersampling, and oversampling. Anomaly detection is an unsupervised framework that makes it possible to use unlabeled samples to identify "not-normal" cancer instances. Undersampling eliminates samples from the majority class to improve the ratio of positive samples to negative. Oversampling creates synthetic samples to increase the number of observations in the minority class.

5.1 Anomaly Detection

One of the critical development bottlenecks in thermography interpretation is data labeling. Eva centers collect hundreds of patient explorations every week. However, the data cannot be used in the machine learning pipeline until it is being labeled by a medical professional, which often

requires additional tests such as the mammogram, ultrasound, and biopsy. Since the tests usually cannot be performed on a single day, many explorations remain unlabeled and cannot be used in the supervised learning context.

Despite not being labeled, these explorations can provide relevant insights into the structure of the data using unsupervised methods. One of such approaches is anomaly detection. Anomaly detection is the process of identifying items in datasets that differ from the norm. Unlike standard classification tasks, anomaly detection is performed on unlabeled data. It takes into account only the internal structure of the dataset and relies on two assumptions:

1. Anomalies differ from the norm.

2. Anomalies are rare compared to normal instances.

Both assumptions are reasonable in the breast cancer detection scenario. First of all, thermograms of breast cancer patients are structurally different from healthy thermograms because of the physiological changes in the blood flow and surface temperature in the proximity of the tumor. Secondly, breast cancer is relatively rare in the population. As stated in the Data section, the proportion of positive cases in Eva data is around 6%.

Unsupervised anomaly detection is used in many practical applications, including the medical domain. For example, Quinn et al. (2019) use anomaly detection to classify cancerous tumors. Apart from the benefit of not requiring any labeled data, the authors claim that anomaly detection has other advantages over traditional methods. Specifically, traditional machine learning methods can only identify cancerous cases that resemble those that the algorithm had access to during training. Thus, such methods may be unsuitable for cancer classification because the possibility space of cancer is vast due to the variety of forms, stages, and cell types. Instead, the authors proposed a method that detects anomalous samples based on their deviation from the typical structure, a variation of which we explore in this section.

We use anomaly detection to extract insights from the unlabeled data and increase the robustness of our model. A detailed implementation is given in the `anomaly_detection.ipynb`. In the notebook, we are testing the hypothesis that anomaly detection techniques are useful in identifying suspicious thermograms using Python Outlier Detection (PyOD) Toolkit.

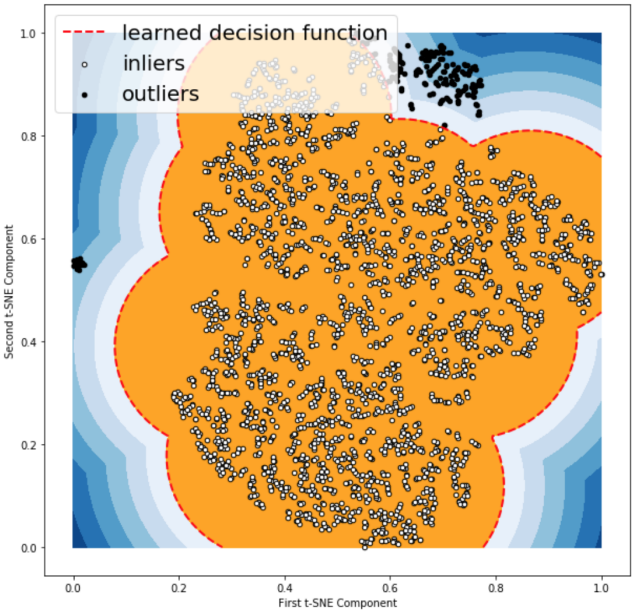

Figure 14: The anomaly detection t-SNE plot performed using the Cluster-Based Local Outlier Factor (CBLOF) method.

PyOD is a library that includes traditional and deep-learning-based anomaly detection approaches. Since anomaly detection is unsupervised, we used both training and unlabeled data to train the model and assign the probability of being an outlier to each observation in the dataset. We tried both individual methods and ensembles of methods since anomaly detection approaches often suffer from model instability due to their unsupervised nature (Kalaycı and Ercan (2018)). Thus, combining various detector outputs can improve robustness. The PyOD library has built-in methods for combining classifiers.

Figure 14 is a t-SNE plot of anomaly detection segmentation in 2D space. Note that there are no visible clusters on the plot. That means that it is impossible

to separate cancer cases from healthy cases easily. At the same time, it also means that the data is not clustered by center location, patient age, or other factors that might bias the model. Orange space illustrates the normal space of the dataset with the red line signifying the learned decision function. Black dots represent outliers.

If our hypothesis is correct, most of the outliers correspond to cancer cases. In this case, we would see good model performance on the test set after training the model on unlabeled and training data. To test this hypothesis, we ran a set of experiments and evaluated anomaly detection methods on the test set. The resulting AUC scores were in the range 64-68% for all models (both traditional and deep learning), which is significantly smaller than the results obtained with the traditional model alone. Thus, the anomaly detection approach alone is not sufficient for making predictions regarding patient status, which disproves our hypothesis. However, we can modify our hypothesis by checking if a combination of the anomaly detection technique and the traditional model leads to better results.

To test this hypothesis, we calculated the probability of being an outlier for each observation in the dataset using the best-performing anomaly detection model, fed this score into the traditional model, and evaluated model performance on the test set. The resulting model demonstrated a slight increase in the AUC and sensitivity values. The plot below puts this change in perspective:

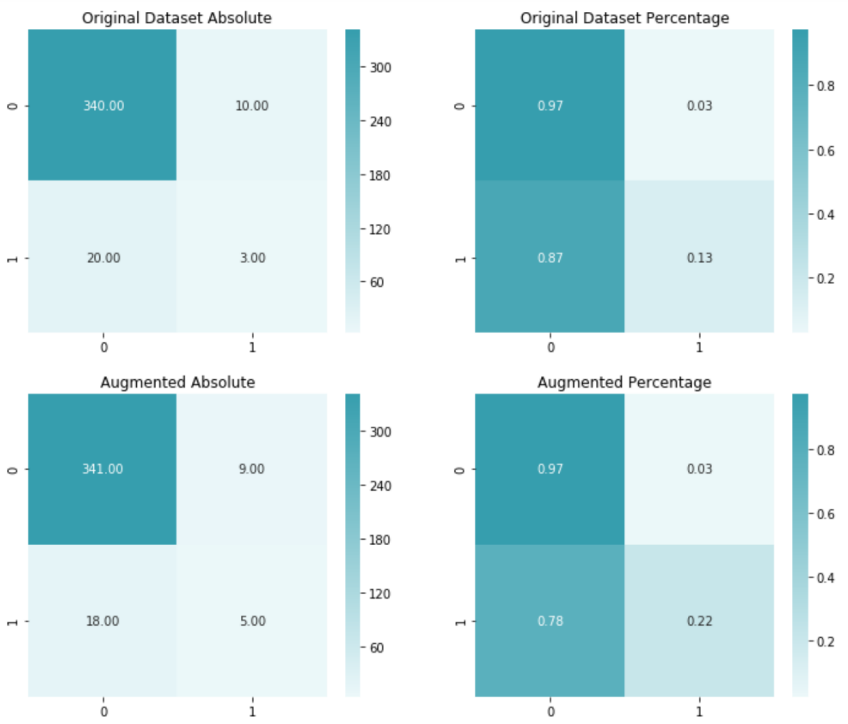

Figure 15: The confusion matrices that illustrate model performance in terms of false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) both in absolute and percentage values. 0 stands for "healthy". 1 means "cancer". The upper row shows the original dataset. The lower row shows the anomaly detection augmented version.

While we can see an increase in the percentage of true positives, the change is not very impressive when translated into absolute values. The number of correctly classified cancer cases (5) is still lower than the number of missed cancers (18) as well as the number of patients with false-positive results (9). Thus, we also try other approaches that might make it possible to get closer to human-level classification performance.

5.2 Undersampling

Undersampling refers to methods that balance the class distribution in imbalanced datasets by removing members of the majority class. The most straightforward undersampling technique removes random observations from the training set. This is referred to as random undersampling. This method removes observations without considering their importance in determining the decision boundary between the classes, which can lead to information loss. In the breast cancer detection context, it is better to remove borderline negative cases that have BI-RADS status 3 and keep healthy cases (BI-RADS 1) and proven benign conditions (BI-RADS 2.) Thus, we use a focused undersampling approach, Tomek Links, to exclude majority examples located on the border between the two classes.

Tomek Links is an enhancement of the nearest neighbor approach. The algorithm works as follows (Tomek et al., 1976):

1. Let x be an instance of cancer and y be a benign condition.
2. Let $d(x, y)$ be the distance between x and y.
3. (x, y) is a Tomek Link if for any instance z:

$$d(x, y) < d(x, z) \text{ or } d(x, y) < d(y, z)$$

   In other words, a Tomek link exists if the two samples are the nearest neighbors of each other.

4. Once a Tomek link is identified, the sample from the majority class is removed.

As a result of applying this algorithm, the boundary between classes becomes more explicit, as illustrated in the figure below:

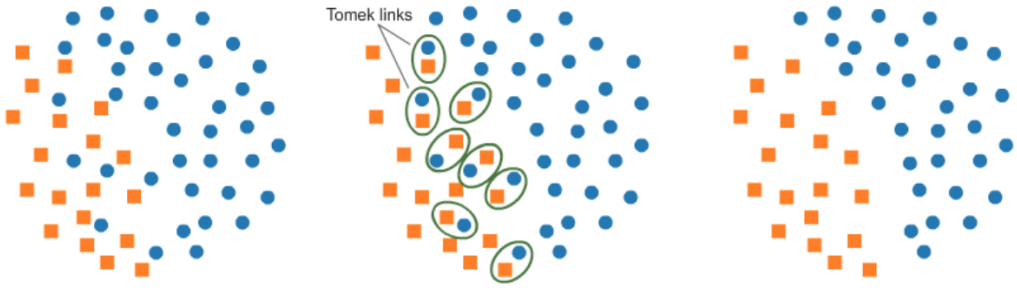

Figure 16: Illustration of Tomek Links. Different colors (blue and orange) represent different classes that we want to separate. Source: Kaggle (2018)

To analyze the performance of sampling methods, we used t-SNE plots introduced in the Data section. Given the complexity of breast cancer detection, it is unlikely that this method would map data to two distinct clusters of healthy and sick patients, as illustrated in the figure above. However, the plot can give some idea of how points are distributed in space and allow analyzing the dataset for signs of overfitting, which is likely to occur in the case of intensive artificial sample generation.

It is important to note that t-SNE does not analyze points in a high-dimensional space. Instead, it looks at the distance between points and tries to respect these relations in the low-dimensional representation. Thus, the t-SNE transform function needs to fit the data every single time, which is why the plots look slightly different for different subsets of data.

The figure below illustrates the representation of data in two dimensions before and after undersampling. The members of the positive class seem to be more clustered after performing the procedure.

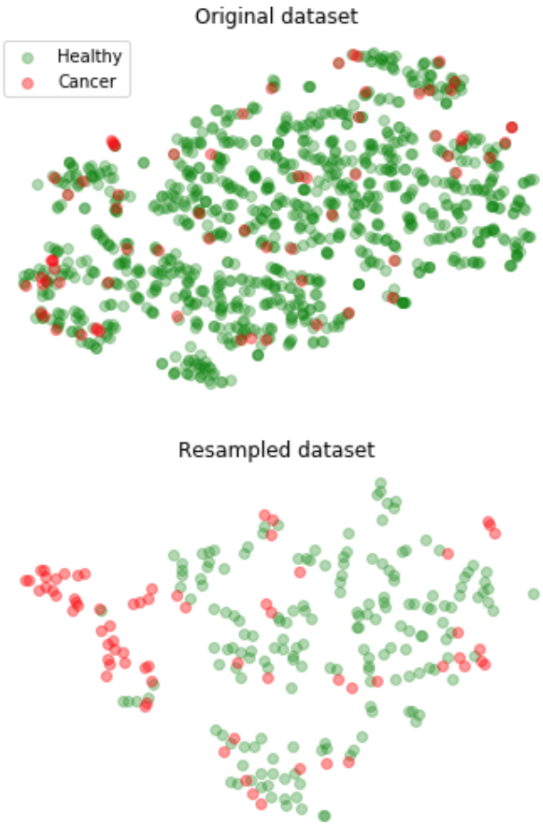

Figure 17: The t-SNE representation of the data before and after the removal of Tomek Links.

The removal of Tomek links had a significant effect on the recall of the classifier that increased from around 10% to 70%, surpassing the recall of the automated human evaluation method. We witness a significant improvement in the score despite training the model with a significantly smaller number of samples. That indicates that class balance can be more important than the sample size. It is also worth noting that an increase in the sensitivity score usually happens at the expense of the specificity score, which means that we misclassify more negative classes as positive. While it is unacceptable in many applications, we believe that it is less harmful than potentially missing a cancer case.

Model	**Sensitivity**	**Specificity**	**Precision**
Original	0.13	0.98	0.13
Undersampling	0.74	0.67	0.33

Table 3: The performance of the original and undersampled datasets.

5.3 Oversampling

Oversampling is a technique that is used to adjust the class distribution of a dataset by generating synthetic data that resembles the underlying distribution of the real data. This section describes how we applied Synthetic Minority Oversampling Technique (SMOTE) to balance out the dataset and improve model performance.

Earlier work on oversampling relied on duplicating random positive class observations. This approach has an inherent problem because positive samples are created without adding any diversity to the dataset, resulting in overfitting to limited positive cases. Thus, we decided to use a more sophisticated oversampling technique known as synthetic data generation.

Synthetic data are simulated data that have the same distribution as the original dataset.

SMOTE is an oversampling method that relies on the nearest neighbors to create its synthetic samples. It was proposed by Chawla et al. (2002) and has been widely used for a variety of problems, including cancer diagnostics (Fotouhi et al., 2019). The SMOTE samples are linear combinations of two similar observations from the minority class (x_i and x_{zi}) and are defined as:

$$s = x_i + u(x_{zi} - x_i)$$

where x_{zi} is randomly chosen among the k nearest neighbors of x_i and u is a parameter that takes on a random value from 0 to 1 to make sure that the newly generated sample is on the line between x_i and x_{zi}:

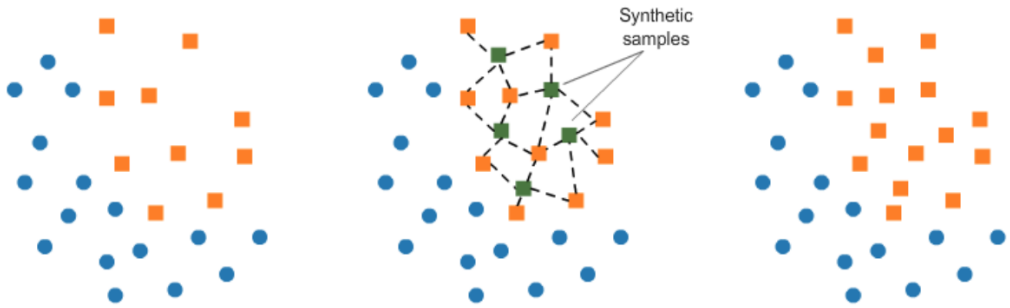

Figure 18: A schematic representation of the SMOTE approach. Different colors (blue and orange) represent different classes that we want to separate. Source: Kaggle (2018)

Oversampling can lead to misleading results if not correctly performed. One important consideration is to account for data leakage by making sure that the procedure is only performed on the training set. The test set should always be separated from the training process and only be used for the final model evaluation. Moreover, the addition of new synthetic samples can lead to the issue of overfitting. This is especially the case for traditional oversampling methods where the samples are duplicated instead of being synthesized. To evaluate the possibility of overfitting, we ran a test where we performed the oversampling procedure on the minority class of the training set. We gradually increased the float parameter that corresponds to the final proportion of positive samples until it reached the ratio 1:1 and calculated the average AUC score for over 100 repetitions for each float value. The resulting curve is presented below:

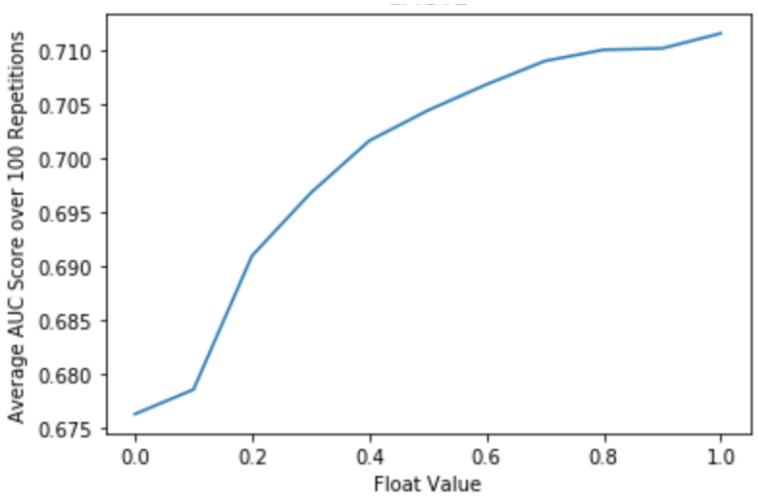

Figure 19: AUC scores averaged across 100 repetitions for oversampled set with different float values.

The graph indicates a gradual improvement in the AUC score of the model that was also translated into improved positive class recall values. Therefore, we concluded that it is safe to use this data augmentation approach in this context.

5.4 Sampling Results

The generation of synthetic data as a standalone method, as well as coupled with undersampling of the majority class, has not outperformed the undersampling method in terms of sensitivity and AUC:

Model	Sensitivity	Specificity	Precision	AUC
Undersampling	0.74	0.67	0.33	0.78
Oversampling	0.22	0.91	0.13	0.72
Both	0.22	0.98	0.38	0.73

Table 4: The classification report for SMOTE, SMOTETomek and Tomek alone.

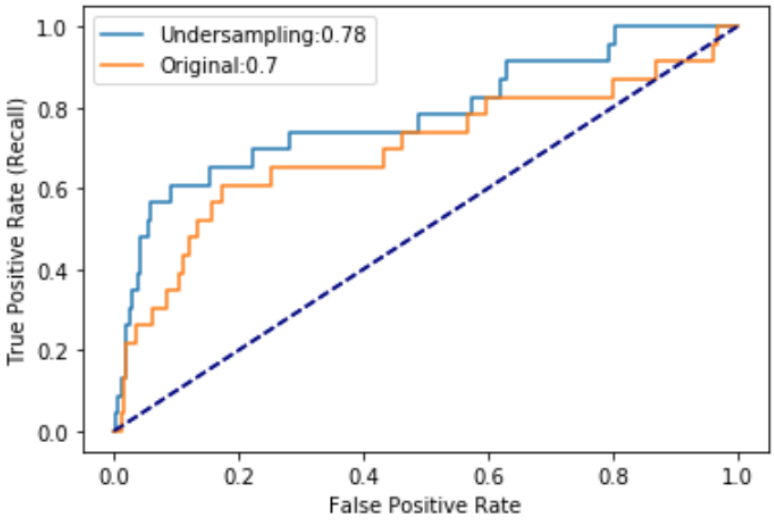

Figure 20: The AUC score of the final augmented model vs. the original model.

Thus, the undersampling of the majority class produced the best results in terms of sensitivity and AUC. While still not perfect, this model is capable of identifying cancer cases while still having a relatively high precision score when it comes to the negative class. The confusion matrix below further evaluates the quality of the classifier:

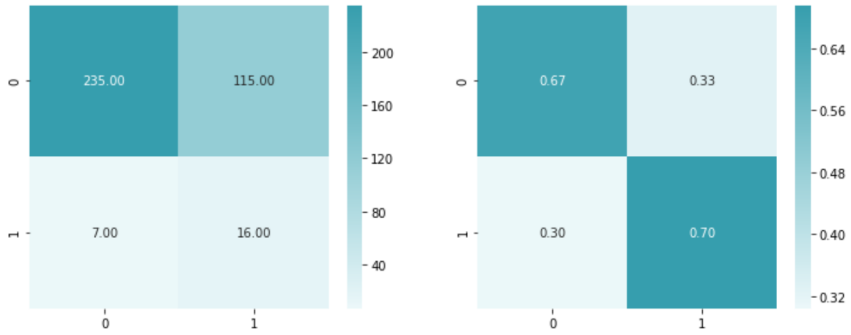

Figure 21: The confusion matrix of the classification on the held-out test set. Left: absolute values. Right: percentages.

The matrix indicates that the model can correctly classify the majority of positive and negative cases. It is possible to catch all cancer cases at the expense of a higher number of false positives by setting the sensitivity-specificity threshold to lower values.

To put these results in perspective, the reported recall (sensitivity) of the screening mammography varies a lot depending on the patient's age and breast tissue density. The average recall for females with breasts with lower tissue density is 87%, but only 63% for women with extremely dense breasts. The mammogram's precision is also affected by tissue density, with an average decrease in the final score of around 10% from almost perfect to around 90% (Carney et al., 2003).

Overall, we demonstrated that careful feature engineering coupled with methods that alleviate the imbalanced dataset issue led to the average recall values that are at least 10% higher than the screening mammography average for females with dense breast tissue. Given that mammography exposes the breasts to harmful radiation, is not perfect at the identification of cancer in women with high-density breast tissue, and can only be performed in older women and infrequently, thermograms should be considered as a useful supplementary method for young females. Additionally, the machine learning methods used in the present work prove that automatic thermography evaluation can be as effective as human-performed interpretation and lead to even more consistent and scalable performance.

6 Conclusion

Thermography for breast cancer detection is a widely studied approach that has received mixed evaluations over time. Some advantages of this method include reduced costs and improved screening accessibility, as well as the mitigation of many risks associated with other detection methods. Thermography can be performed more frequently and lead to earlier cancer detection, improving patient survival rates.

This paper advances the thermography literature by offering a novel perspective on the performance of machine learning methods for thermography evaluation in comparison to the industry-approved mammography results as well as human-performed thermography evaluation techniques.

We compared two quantitative human interpretation methods using the evaluations obtained from independent medical professionals. We concluded that the Gonzalez score outperformed both the subjective evaluation and the Keyserlingk score due to its ability to capture relevant quantitative characteristics of thermal patterns. We analyzed 114 manual thermography interpretations to automate the calculation of the thermal score and used it as a baseline model to evaluate the performance of machine learning methods.

Since the number of positive samples in the dataset is not enough to train or fine-tune deep learning models, we turned to classical feature engineering by extracting the LBP and HOG vectors and reducing their dimensionality to make computations on bigger sets feasible. We performed SHAP feature selection and hyperparameter tuning using cross-validation. Despite a very high AUC score, the model did not outperform human evaluators due to its inherent bias towards the majority class.

One of the main contributions of this paper is the exploration of data augmentation methods that mitigate the imbalanced dataset problem that is highly prevalent in cancer detection and medical imagery domains. To increase the diversity of data, we employed insights from the unlabeled observations that are present in abundance at Eva Centers as well as experimented with sampling methods. While anomaly detection and synthetic sample generation methods demonstrated modest improvements in the sensitivity score, undersampling helped to increase the recall of the model from 10% to almost 74%, which is 11% higher than the industry-accepted mammography benchmark as well as the human evaluation results obtained on the entire set.

Thus, this work shows that automatic thermography evaluation coupled with careful feature engineering and data augmentation methods offers a fresh promise for early diagnostic of breast cancer that we should not exclude from the medical decision-making process. Further work should focus on testing promising methods from the literature on bigger datasets to increase the trustworthiness of thermography research and provide more statistically significant evidence in support of thermography to the broader medical community.

References

Abdel-Nasser, M., Moreno, A., & Puig, D. (2019). Breast cancer detection in thermal infrared images using representation learning and texture analysis methods. *Electronics*, *8*(1), 100.

Anders, C. K., Fan, C., Parker, J. S., Carey, L. A., Blackwell, K. L., Klauber-DeMore, N., & Perou, C. M. (2011). Breast carcinomas arising at a young age: Unique biology or a surrogate for aggressive intrinsic subtypes? *Journal of Clinical Oncology*, *29*(1), e18.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394–424.

Bugli, C., & Lambert, P. (2007). Comparison between principal component analysis and independent component analysis in electroencephalograms modelling. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *49*(2), 312–327.

Campbell, J. D., & Ramsey, S. D. (2009). The costs of treating breast cancer in the us. *Pharmacoeconomics*, *27*(3), 199–209.

Carney, P. A., Miglioretti, D. L., Yankaskas, B. C., Kerlikowske, K., Rosenberg, R., Rutter, C. M., Geller, B. M., Abraham, L. A., Taplin, S. H., Dignan, M., Et al. (2003). Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of internal medicine*, *138*(3), 168–175.

Chan, T., & Vese, L. (1999). An active contour model without edges, In *International conference on scale-space theories in computer vision*. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Choi, J. Y., Kim, D. H., & Ro, Y. M. (2012). Combining multiresolution local binary pattern texture analysis and variable selection strategy applied to computer-aided detection of breast masses on mammograms, In *Proceedings of 2012 ieee-embs international conference on biomedical and health informatics*. IEEE.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection.

Ergin, S., & Kilinc, O. (2014). A new feature extraction framework based on wavelets for breast cancer diagnosis. *Computers in biology and medicine*, *51*, 171–182.

Feig, S. A., Shaber, G. S., Schwartz, G. F., Patchefsky, A., Libshitz, H. I., Edeiken, J., Nerlinger, R., Curley, R. F., & Wallace, J. D. (1977). Thermography, mammography, and clinical examination in breast cancer screening: Review of 16,000 studies. *Radiology*, *122*(1), 123–127.

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, *90*, 103089.

Francis, S. V., Sasikala, M., & Saranya, S. (2014). Detection of breast abnormality from thermograms using curvelet transform based feature extraction. *Journal of medical systems*, *38*(4), 23.

Freer, P. E. (2015). Mammographic breast density: Impact on breast cancer risk and implications for screening. *Radiographics*, *35*(2), 302–315.

Gamagami, P. (1996). Indirect signs of breast cancer: Angiogenesis study. *Atlas of mammography. Cambridge, Mass: Blackwell Science*, 321–6.

Garduño-Ramón, M., Vega-Mancilla, S., Morales-Henández, L., & Osornio-Rios, R. (2017). Supportive noninvasive tool for the diagnosis of breast cancer using a thermographic camera as sensor. *Sensors*, *17*(3), 497.

Gautherie, M. (1989). Atlas of breast thermography with specific guidelines for examination and interpretation. *Milan, Italy: PAPUSA*, *256*.

González, F. J. (2011). Non-invasive estimation of the metabolic heat production of breast tumors using digital infrared imaging. *Quantitative InfraRed Thermography Journal*, *8*(2), 139–148.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11), 1768–1783.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. corr abs/1512.03385 (2015).

Intel. (2019). Image color conversion, volume 2, image processing, intel integrated performance primitives for intel architecture developer reference.

Kaggle. (2018). Resampling strategies for imbalanced datasets.

Kalaycı, İ., & Ercan, T. (2018). Anomaly detection in wireless sensor networks data by using histogram based outlier score method, In *2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit)*. IEEE.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, *1*(4), 321–331.

Kennedy, D. A., Lee, T., & Seely, D. (2009). A comparative review of thermography as a breast cancer screening technique. *Integrative cancer therapies*, *8*(1), 9–16.

Keyserlingk, J. R., Ahlgren, P., Yu, E., Belliveau, N., & Yassa, M. (2000). Functional infrared imaging of the breast. *IEEE Engineering in Medicine and Biology Magazine*, *19*(3), 30–41.

Koay, J., Herry, C., & Frize, M. (2004). Analysis of breast thermography with an artificial neural network, In *The 26th annual international conference of the ieee engineering in medicine and biology society*. IEEE.

Law, J., Faulkner, K., & Young, K. (2007). Risk factors for induction of breast cancer by x-rays and their implications for breast screening. *The British journal of radiology*, *80*(952), 261–266.

Li, C., Uribe, D., & Daling, J. (2005). Clinical characteristics of different histologic types of breast cancer. *British journal of cancer*, *93*(9), 1046.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions, In *Advances in neural information processing systems*.

Madjar, H. (2002). Advantages and limitations of breast ultrasound. *Gynakologisch-geburtshilfliche Rundschau*, *42*(4), 185–190.

Miglioretti, D. L., Lange, J., Van Den Broek, J. J., Lee, C. I., Van Ravesteyn, N. T., Ritley, D., Kerlikowske, K., Fenton, J. J., Melnikow, J., De Koning, H. J., Et al. (2016). Radiation-induced breast cancer incidence and mortality from digital mammography screening: A modeling study. *Annals of internal medicine*, *164*(4), 205–214.

Nakashima, K., Uematsu, T., Takahashi, K., Nishimura, S., Tadokoro, Y., Hayashi, T., & Sugino, T. (2019). Does breast cancer growth rate really depend on tumor subtype? measurement of tumor doubling time using serial ultrasonography between diagnosis and surgery. *Breast Cancer*, *26*(2), 206–214.

Noone, A., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Et al. (2018). Seer cancer statistics review, 1975-2015. *Bethesda, MD: National Cancer Institute*.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), 971–987.

Parisky, Y., Sardi, A., Hamm, R., Hughes, K., Esserman, L., Rust, S., & Callahan, K. (2003). Efficacy of computerized infrared imaging analysis to evaluate mammographically suspicious lesions. *American Journal of Roentgenology*, *180*(1), 263–269.

Partridge, A. H., Goldhirsch, A., Gelber, S., & Gelber, R. D. (2014). Breast cancer in younger women, In *Diseases of the breast: Fifth edition*. Wolters Kluwer Health Adis (ESP).

Pereira, E. T., Eleutério, S. P., & Carvalho, J. M. (2014). Local binary patterns applied to breast cancer classification in mammographies. *Revista de Informática Teórica e Aplicada*, *21*(2), 32–46.

Piana, A., & Sepper, A. (2015). Contemporary evaluation of thermal breast screening. *Pan American Journal of Medical Thermology*, *1*(2), 93–100.

Quinn, T. P., Nguyen, T., Lee, S. C., & Venkatesh, S. (2019). Cancer as a tissue anomaly: Classifying tumor transcriptomes based only on healthy data. *Frontiers in genetics*, *10*, 599.

Raghavendra, U., Rajendra Acharya, U., Ng, E., Tan, J.-H., & Gudigar, A. (2016). An integrated index for breast cancer identification using histogram of oriented gradient and kernel locality preserving projection features extracted from thermograms. *Quantitative InfraRed Thermography Journal*, *13*(2), 195–209.

Silva, L. F., Santos, A. A. S., Bravo, R. S., Silva, A. C., Muchaluat-Saade, D. C., & Conci, A. (2016). Hybrid analysis for indicating patients with breast cancer using temperature time series. *Computer methods and programs in biomedicine*, *130*, 142–153.

Tan, T. Z., Quek, C., Ng, G. S., & Ng, E. (2007). A novel cognitive interpretation of breast cancer thermography with complementary learning fuzzy neural memory structure. *Expert Systems with Applications*, *33*(3), 652–666.

Tice, J. A., Cummings, S. R., Smith-Bindman, R., Ichikawa, L., Barlow, W. E., & Kerlikowske, K. (2008). Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Annals of internal medicine*, *148*(5), 337–347.

Tomek, I. Et al. (1976). Two modifications of cnn.

Yaffe, M. J. (2008). Mammographic density. measurement of mammographic density. *Breast Cancer Research*, *10*(3), 209.

Yao, X., Wei, W., Li, J., Wang, L., Xu, Z., Wan, Y., Li, K., & Sun, S. (2014). A comparison of mammography, ultrasonography, and far-infrared thermography with pathological results in screening and early diagnosis of breast cancer. *Asian Biomedicine*, *8*(1), 11–19.

YCS. (2019). Breast cancer in young women: Statistics and disparities.

A Human Evaluation Metrics

1. **Regions of interest:** The number of separate vascularity regions that researchers considered necessary for the analysis. The most common value is 1, with 0 in cases where no asymmetries are present, and values higher than 1 if there are multiple areas of potential abnormality.

2. **Asymmetry size:** Size of the asymmetric vascularity region on a scale from 1 to 10. If there is more than one asymmetrical vascular region, only the larger one is considered.

3. **Asymmetry shape:** How abnormal the asymmetry looks to the interpreter on a scale from 1 to 10. Only the most abnormal looking vascularity region is considered.

4. **Nipple asymmetry:** A binary value to indicate whether the thermal patterns around the nipples are asymmetrical.

5. **Breast shape asymmetry:** A binary value to indicate whether the breasts significantly differ in shape.

6. **Delta-T:** Temperature difference between the hottest point in the asymmetrical vascularity region, and its corresponding location on the opposite breast.

7. **Vascularity:** A vascularity score between 1 and 4. This value combines both asymmetry size and shape. This metric was added to replicate the thermal score method.

8. **Subjective score:** The interpreter's level of suspicion on a scale from 0 to 10, where 0 represents certainty that the patient is normal, and 10 certainty that the case is abnormal. The interpreter is asked to give this score based on their gut feeling without using any metrics.

B Author Contributions & Repository Access

Both the report and the code repository were created without any coordination within Eva Tech and can be seen as my individual work. Even though some parts of my code are used in the company's machine learning pipeline and have evolved over time, the code presented in the repository does not include any of the changes made by other contributors.

I generated all results, figures, and metrics using a subset of real data, but not necessarily the same subset that is used to evaluate performance at Eva Tech. The model that I designed is not related to the model used by the company. I also do not claim that the model performance presented here has any relation to the performance obtained by the company.

Due to the sensitivity of the content of the GitHub repository, it cannot be made publically available. Project contributors can be added upon request. If you want to get access to the code, have any questions, suggestions, or have identified an error, please do not hesitate to email me at `rita@minerva.kgi.edu`.

C Applications of the Minerva Curriculum

C.1 Project-Specific Learning Outcomes

#cs156modelmetrics

In the Metrics section, I provide an overview of popular machine learning metrics and their statistical and contextual meaning. I explain the accuracy paradox and justify that this metric is unsuitable for the breast cancer detection context. Specifically, I argue that my dataset is highly imbalanced due to the low incidence of breast cancer in the population and explain the high cost of false negatives that could cost people their lives. I disregard the F-1 score for this exact reason since it gives the same weight to precision and recall, while the primary goal of initial screening is minimizing the number of false negatives.

I argue that the area under the ROC curve should be the primary metric to evaluate models since it is suitable for imbalanced data and illustrates the diagnostic ability of a binary classifier at different thresholds. The second property is especially useful since it becomes possible to use

our model in contexts with different base rates of cancer. A model with a higher area under the ROC curve would have increased sensitivity with a lower trade-off in specificity. At the same time, I still report the sensitivity of my models because high AUC values in the context of imbalanced classifiers might trick people into thinking that the model performance is better than it is due to almost perfect specificity values. Thus, I also put my results into perspective by including classification reports and confusion matrices both in absolute values and percentage points.

Apart from using machine learning metrics to evaluate models, I also used SHAP values to compute the relative importance of features and reduce their number. This approach gave me not only insights into the feature contribution to the model prediction, but also the direction of their effect. I used this property to observe the unexpected effect of the delta-T on the model output, as described in the Feature Importance section.

#cs156classification

Early breast cancer detection is a classification problem that is characterized by an unclear cut between positive and negative class, imbalanced datasets, high cost of false negatives, and an undefined baseline. In the BI-RADS section, I explain that breast cancer is not a single condition but rather a set of diseases that can have multiple symptoms, thermal manifestation, and histological types. Therefore, I built my binary classification model based on the breast cancer radiologic quality assurance tool that divides images into distinct categories. Since the primary goal of my model is early detection, I classify all thermograms that have a significant shift in thermal patterns (BI-RADS 3 and higher) as suspicious to ensure that no cancer cases go unattended. While it can result in an increased number of false positives, some other conditions, such as benign diseases, might need medical assistance. Thus, some additional exams might be beneficial for patients. Additionally, the task becomes easier for the model, since it only needs to identify whether the thermogram is "normal" or "suspicious."

In the Human Evaluation section, I present my analysis of manual thermography evaluation performance, which is, to my knowledge, the first attempt to set a baseline for machine learning approaches. The automation of the quantitative evaluation methods gave me further insights into how humans classify thermograms, which I could use for my model. For example, I found both delta-T values and vascularity scores highly predictive of breast cancer (both in top-10 features) despite the unintuitive direction of the effect illustrated by the SHAP model. A further study is needed to evaluate the effect of these specific features.

The field of disease diagnostics is usually characterized by small and imbalanced datasets. Thus, I followed strict rules to ensure that the model does not develop a bias towards the positive class by implementing class balancing techniques. I also selected a model that has an optimal bias-variance trade-off to ensure it does not overfit to the limited positive samples. To ensure a low number of false negatives, I tuned hyperparameters to optimize for recall. Lastly, I was aware of the opportunities for data leakage and took actions to eliminate this possibility with the help of cross-validation and held-out test data. To ensure the generalizability of my model to different contexts, I used the ROC curves that provide an ability to choose a specific threshold. These manipulations enabled me to get a drastic improvement in the AUC score and sensitivity values compared to the baseline score and industry-level mammography, benchmark as discussed in the Sampling Results section.

#cs110codereadability

My code is well-structured, well-documented, and relates to other files and the report where necessary. The README file has clear instructions on the purpose of each file and the preferred execution order. Given the fact that I use a lot of proprietary libraries, I made sure to give context to each of the unclear function calls using inline commenting and markdown cells. Since not all required dependencies can be made available for the review, I made sure to package my efforts into Jupiter notebooks that show outputs such as data visualizations and result tables, follow the writeup chronologically, and have clear navigation. My code adheres to the PEP-8 standard, uses clear variable names and extensive docstrings with the characterization of inputs and outputs and their types. To minimize the repetition of code, I gathered all frequently-used functions in the `utils.py` file so that all of them can be imported a single time and used in different notebooks. Wherever possible, I used popular high-reputed libraries such as `NumPy` and `scikit-learn` to ensure

smooth and highly optimized implementations. I also display my fluency in Python by following the best Python idiomatic practices, such as using list comprehensions, highly efficient structures such as dictionaries, and vectorization instead of for loops.

#RKDataAugmentation

While feature engineering, model selection, and hyperparameter tuning are vital for obtaining good results, I would not be able to achieve industry-level performance without dealing with the imbalanced dataset problem. While some breast cancer detection papers balance their datasets artificially by manually picking clear positive and negative cases, a company like Eva needs to be able to classify incoming observations that follow the real-life distribution. Thus, these observations are highly skewed to negative cases. Machine learning models are at risk of developing a bias towards the majority class in case they only have access to a limited number of positive samples. In the Data Augmentation section, I tried several approaches proposed in the literature.

First, I augmented my data using insights extracted from multiple unlabelled samples gathered at Eva Centers. While this approach, in combination with the traditional machine learning model, led to a statistically significant improvement in the model sensitivity, it is not as effective in practice because the number of missed cancers remained high. The AUC score of the anomaly detection model alone was around 65%, which is, although small, gives us a reason to assume that with a rising amount of data, the model can be retrained and be able to learn the structure of the data much better to the extent that we could use it as a stand-alone method for the automatic sample labeling.

The second approach was based on artificial sample generation using the SMOTE approach. Similar to the anomaly detection method, the improvement was consistent but not very impressive. Since oversampling has a significant chance of overfitting, I made sure to use a more profound approach than simple duplication and also ran a series of tests to ensure there is no possibility for overfitting even with large float values. I only added synthetic samples to the training set since the augmentation of the validation and test sets would result in a more balanced distribution and defeat the purpose of using this model to classify real-life data.

Lastly, I undersampled the positive class by eliminating the Tomek links and creating a more explicit boundary between two classes. This approach led to especially good results, which is also demonstrated by the cluster formed by red points on the t-SNE plot in the respective section of the report. Thus, a combination of the anomaly detection approach and undersampling led to the AUC score of 78%, with a sensitivity value of 74%.

Finally, I would like to briefly mention that I also experimented with the Generative Adversarial Network (GAN) to generate synthetic patient explorations. While this approach shows a great promise in the medical imagery domain, it barely worked in my case since each of my observations contains four images, and in most cases, only one of the images in positive observations had cancer. Since Eva does not have information on the tumor location for most patients, the resulting dataset was way too small to generate reliable samples. Thus, I plan to get back to this approach later once more information on the tumor location becomes available.

#cs156overfiting

Overfitting and data leakage are two main problems that I have identified in several peer-reviewed articles on breast cancer detection. Given the imbalanced nature of the dataset and the limited number of samples, it is easy to fall into the trap of getting the most out of the data without holding out a subset for testing. While relevant in some contexts, leave one out cross-validation (LOOCV), and other similar validation methods should not be performed in conjunction with hyperparameter tuning and model selection. This approach leads to inevitable overfitting and reduced generalizability of results. Since my main objective is to provide evidence regarding the thermography effectiveness, I was very aware of the potential for overfitting and always followed the best methodologies to avoid it.

First of all, in the feature engineering section, I fit my dimensionality reduction techniques on the train data exclusively to ensure the model does not have access to the structure of the test and validation sets. At the early stages of the project, this mistake led to suspiciously good results. Thanks to the careful review of my work by my manager at Eva Tech, we managed to identify this issue before it went into production.

Secondly, both hyperparameter tuning and feature selection should not be performed using the test set due to the increased probability of overfitting. Thus, I used a validation set for intermediate result evaluation and only checked the model performance on the final set once I made all modeling decisions. While I experienced a small decrease in the AUC score compared to the validation set (80% and 78%), I now know that my model exhibits very consistent performance on different subsets of unseen Eva data, which indicates high generalizability.

Lastly, I was aware of the issue of overfitting in the context of artificial sample generation. Thus, I ran a series of experiments on different train-validation splits and averaged my results over 100 runs for each float value. The AUC score remained consistently high and increasing with a higher float value, showing that overfitting was of little concern for my specific dataset.

C.2 Capstone Universal Learning Outcomes

#cp193.024-qualitydeliverables

The final version of my Capstone contains two main deliverables: a report and a code repository. Compared to the previous submission, I have significantly improved the flow of the paper by incorporating feedback from three reviewers. Specifically, I improved the readability of the report by splitting long paragraphs into shorter ones, added clarifications to some figures, and extended the interpretation of results. I also added two new sections on Hyperparameter Tuning and Model Selection that had a significant positive impact on the final score. I added additional justification to the feature engineering section and listed some successful previous applications of several methods.

I updated the repository by adding clarification where necessary and making connections to the relevant sections of the report. I also optimized some of my functions to ensure shorter running times.

#cp193.024-planningarchitecture

After submitting the final draft, I have thoughtfully analyzed the components that were missing from my submission or required some extra work. I assigned different priorities to each of the tasks and put them on a timeline that included 2 hours of work every day for two and a half weeks. This approach was beneficial because I did not spend time deciding what specific task I want to complete on a given day. Since I usually work well under pressure, I assigned more significant and less exciting tasks closer to the deadline to minimize procrastination. I started with missing components to ensure that I do not get too absorbed by reviewing almost ready sections. Additionally, it gave me some freedom to experiment with my code before finalizing all components.

Since I only worked for two hours per day, I was not overwhelmed by the amount of work and felt very productive. I managed to complete the project 5 days before the original deadline and had enough time for final revisions.

#cp193.024-feedback

I saw these three weeks after the submission of the final draft as the best and most fruitful time for receiving feedback since the project was almost finished. One of my reviewers, who is a professor at my previous university, gave me an important piece of feedback regarding the use of the validation set. In the previous submission, I used a test set for all evaluations and, thus, have overfitted to this limited sample. I tried running my model on an unseen set of data and got a worse performance. Thus, I created a validation set and used it to make modeling decisions, such as model selection and hyperparameter tuning. I evaluated the performance of the model on the held-out test set and reported this final result. The performance was stable when I tested it on other subsets of Eva data, indicating that I solved the issue.

I also have significantly improved my HC and LO justifications based on the feedback I received from Professor Ribeiro. This time, I tried to emphasize not only what I did but also why it led to better results and why it made my project better. I also paid closer attention to the HC rubrics to make sure my justifications cover all components required for a score of 4.

#cp193.024-accountability

This time, I went over what was in my original plan and not only finalized my deliverables but also added new crucial parts on Model Selection and Hyperparameter Tuning. This addition

required a careful revision of my results and rerunning the pipeline to ensure consistency and reproducibility.

I set this goal while realizing that it might require an additional time commitment. So, I carefully planned these changes and started working on them right after the previous submission. I know that I am usually more productive under pressure, so I set up an accountability group with two other Minervans and talked with them every three days to report on my progress. As a result, all three of us managed to finish our projects before the final deadline and have enough revision time.

#cp193.024-research

I completed most of the research in my junior year. Details are given in the **#sourcequality** and **#evidencebased** HC justifications. In this final submission, I added additional research-based justifications for the use of feature engineering methods. For example, I researched the use of HOGs in thermography and breast cancer detection and found out that the ICA dimensionality reduction method that I initially used did not outperform other methods proposed in the literature. I reran my analysis based on this finding and identified that the KPCA method led to better outcomes. Additionally, I researched the best hyperparameter tuning practices that prevent overfitting that I summarized in the Model Selection section of the writeup.

#cp193.024-connect

Reaching out to medical and machine learning experts was crucial for my project. In the initial stages of the Capstone, I was in touch with the members of the FutureTalks community who provided guidance and support during my time in Berlin. I shared my ideas during community-wide presentations and gathered feedback and critique from AI and ML experts as well as medical professionals that were necessary to finalize my idea.

This project would not be possible without the support of my colleagues from Eva Tech, who commented on my work, helped me navigate the code repositories at the initial stages of the project, and always emphasized the practical side of my work.

My advisor, Professor Ribeiro, provided useful feedback regarding the structure, content, and academic rigor of my project and shared her knowledge in machine learning, specifically on feature engineering and imbalanced dataset issue. She also provided emotional support throughout this eventful year and motivated me to achieve extraordinary despite difficulties and uncertainties.

My peers, who reviewed my Capstone at different stages, not only helped me to look at my project with different eyes but also see how my project compares to theirs and showed me some ways of improving the project that I have not considered before. For example, looking at Jennifer's pretty formatting made me think that I could also improve my **#presentation** even though I did not consider it before.

Finally, I am very thankful for the feedback from my final set of reviewers who took their time to look through my work before this final submission. I got feedback from a Data Science professor, Machine Learning Engineer, and a Minerva student. All of them looked at my project through different lenses and highlighted some parts that needed additional work.

#cp193.024-metrics

To evaluate my progress, I have always constructed a detailed timeline. The tools have changed over time, starting from a Google spreadsheet, moving to a Notion notebook, and finishing with the Todoist app, where I kept track of my goals before the final submission. I set my objectives and key results (OKR) with the help of the Capstone assignments to make sure I am on track with Minerva's expectations.

My leading performance indicators were the amount of time spent every week and the number of tasks I crossed out of my TODO list. I presented my progress in our Capstone group and received feedback from Professor Ribeiro and my peers every two weeks. Additionally, I reported my progress to my Eva manager every week by making a presentation for the demo day.

During my last semester, I started to use a different metric — completeness. For example, if I believed that the section was at least 80% complete, I did not spend too much time on it and moved to other sections that required more attention. This approach helped me to cope with my perfectionism that made me unproductive in cases I needed to move on instead of revising existing

sections multiple times.

C.3 Foregrounded Habits of Mind and Foundational Concepts

#evidencebased

I provide structured evidence for three components of the project. First of all, in my Background section, I argue that thermography is a plausible, safe, and effective method of breast cancer detection using evidence from the literature. I provide a biological explanation of the effectiveness of thermography by explaining the mechanism behind the tumor formation and increased angiogenesis that can be captured by the infrared cameras. I also cover some of the past research on automated and human-performed thermography evaluations that indicate that thermography can be as effective as mammography, especially for young females and females with dense breast tissue.

Secondly, I chose the Gonzalez score as the baseline for model evaluation based on the evidence presented in the Human Evaluation section. The analysis of the assessments obtained from two independent interpreters indicates that the Gonzales score outperforms a fully subjective evaluation and the Keyserlingk score both in terms of their correlation with the BI-RADS status and the ease of implementation (it is impossible to implement an entirely subjective score since it cannot be quantified.)

Lastly, I justify the application of every single machine learning method that I use in the paper. For example, in the implementation of the LBP features, I cite an article that has successfully implemented this approach for breast cancer detection (Raghavendra et al., 2016) and explain how the resulting local binary patterns relate to the areas of increased vascularity and potential tumor location. I also reduced the dimensionality of the resulting vector by following the guidelines presented in the original paper.

#sourcequality

I demonstrate my use of sources in the Background section, where I conduct a thorough review of breast cancer and thermography literature. I focused my attention on the most up-to-date research as well as the most influential papers in the field. I close read and critically evaluated more than thirty peer-reviewed articles. In the Thermography Research section, I argue that past research has been driven by insufficiently high standards. Specifically, I criticize poor quality control in the influential study by Feig et al. (1977) and express my concerns regarding the experimental design of Silva et al. (2016), which is prone to data leakage. At the same time, I make it clear that there is promising evidence that thermography evaluation, both human-performed and automated, can achieve an industry-accepted level of performance and should not be disregarded by the medical community.

In the Feature Engineering section, I have not only identified and analyzed influential papers in the fields of thermography and breast cancer detection but have also employed methodologies therein suggested to extract conclusions from my dataset and see which approaches work better in my specific context. For example, Raghavendra et al. (2016) identified that the LPP was the best performing dimensionality reduction method. However, I identified a significant increase in the metrics of interest with the KPCA method, which was proposed in another paper on HOGs (Ergin & Kilinc, 2014).

In the machine learning section of my project, I used open-source libraries with a large number of contributors and reviewers, namely `scikit-learn`, `pandas`, and `NumPy`. These libraries are some of the most reputable sources that have extensive documentation. I referred to this documentation as my primary source of knowledge to follow established standards and best coding practices. This approach is the most evident in my `ML_pipeline.ipynb`, where I store and operate with temperature matrices and newly generated features using `NumPy` arrays to ensure smaller memory consumption and better runtime behavior.

#organization

One obvious application of this HC is the way I structured this report. I organized the sections using the standard academic paper breakdown: Introduction, Methods, Results, and Conclusion. This should let the reader focus on the content more than the structure. I also added a glossary

and a table of contents to make the navigation easier.

I did my best to craft the report in a way that integrates theoretical justification with practical implementation in a clear and easy-to-understand way. I kept in mind my target audience and tried to give just enough background information to make my reasoning clear without getting into unnecessary details. I also used the Zoom method in all of my sections by giving a short description of the section before diving into its specificities. I analyze the performance of each approach inside the section as well as in the Conclusion section to iterate on my findings. I also put my results into perspective by comparing them with the mammography and human evaluation baselines.

I believe that this enhanced organization structure tells the reader a story of my Capstone by emphasizing not only my final results but also the process I went through and what I learned on the way. For example, I present preliminary results before the Data Augmentation section to come up with potential areas of improvement and ideate solutions.

Besides this application, I would like to highlight the organization of my code repository. It contains a README file that explains the purpose of each notebook and makes it easier to navigate. I kept all the setup and utility functions in separate files to not distract readers' attention from the data analysis and results. The notebooks can be easily associated with relevant parts of the writeup in case further justification is required. Each Jupyter notebook has a table of contents that highlights essential components. I also added clarifying comments in the markdown cells or inline. Thus, although it is impossible to run the code due to the sensitivity of the data, the reader can still see my results in the order I obtained them.

#professionalism

This project is a result of almost two years of brainstorming, research, replication, critical evaluation, and iteration. I believe that the quality of my literature review and the depth of my analysis indicate that I took this work very seriously and produced deliverables of high professional quality.

I strictly followed the guidelines from the Capstone handbook, such as the recommended word count, the number of HCs and LOs that should be tagged in the appendix, and paper formatting. My code is well-documented and adheres to the PEP-8 standard, includes docstrings and clarifying markdown cells.

To ensure the highest quality of presentation, I produced the report in LaTex and rendered it using overleaf.com. I checked all sections for grammar and plagiarism using Grammarly. I carefully cited all my references in the APA format. Figures and tables have clear captions and sources in case I did not produce them myself. I also included a glossary with machine learning and cancerology terms.

Since my Capstone contains some sensitive information and is subject to intellectual property limitations, I added a copyright note. I also made sure no protected/identifiable information is shared without the company's consent.

#responsibility

The application of this HC is mostly evident in the planning and execution phases of the project. My responsibilities included but were not limited to brainstorming project ideas, researching project plausibility and relevance, writing readable code and experimenting with various machine learning methods, finalizing my research in the writeup, and actively seeking feedback from Minerva peers and advisers, as well as external experts.

While working on the paper, I developed a strong internal locus of control by realizing that the outcome and quality of my project depend solely on my hard work and carefully planned execution. Thus, to ensure the success of the project, I assigned levels of commitment (low, medium, high) to specific responsibilities based on the Minerva expectations, as stated in the Capstone Handbook and expressed by my Capstone Advisor. I assigned the highest levels of commitment to milestones, such as assignment deadlines. For example, I focused most of my time and effort on the literature review before the respective deadline in the Junior year. This assignment not only enabled me to learn more about the topic I am passionate about but also minimize the time spent on writing the Background section in the Senior year thanks to the high quality and depth of my work. Some other responsibilities received a lower level of commitment since they were less vital for the project's success (not a part of the MVP) but still relevant. For example, I produced multiple visualizations

that, although not a Capstone requirement, serve a vital role in increasing the clarity and integrity of the project that is restricted by intellectual property rights and where the code in the notebooks cannot be executed because of several proprietary dependencies.

To hold myself accountable, I not only adhered to Minerva deadlines and met with my advising group regularly, but I also had a running timeline with all the tasks and subtasks. I reviewed it at least once per month by crossing the tasks off or reassigning them to a new date. During the planning phase, I talked a lot to the members of the FutureTalks community, who guided me through the brainstorming process. They helped me to focus on some of the relevant problems that humanity faces and develop a growth mindset by realizing that I can tackle some of the toughest issues if I am genuinely passionate about them. Additionally, I would like to highlight my peer-network of students who have similar projects and who motivated me to progress and keep myself accountable by giving feedback, namely Anna and Jennifer.

C.4 Transfer-Eligible Habits of Mind and Foundational Concepts

#rightproblem

In the Background section of my report, I explain that there exists a gap in the early detection of breast cancer in young females that is mainly attributed to the lack of reliable diagnostics methods. The gold standard for breast cancer detection, mammography, is inappropriate for detection in young females due to the increased risk of radiation-induced breast cancer and the reduced efficiency of the method. As a result, the mortality rate among young females is high because tumors are often identified at later stages and also tend to be more aggressive.

Since 250,000 women under 40 are diagnosed with breast cancer every year in the US alone, the identification of a reliable early diagnostics method in this group would significantly benefit society. I argue that thermography is an alternative approach that has the potential to solve this problem. Unlike mammography, it is a safe and tissue-agnostic method. Additionally, it might be able to see pathological changes before the actual tumor forms, opening opportunities for early intervention and continuous diagnostics.

I also point out some potential downsides of this solution, such as a high number of false positives, but also explain that this downside does not outweigh the benefit of possible early detection of a deadly disease that can save people's lives. Before tackling the problem, I also identified some of the constraints such as lack of well-defined quantitative standards and thermography's bad reputation that can be overcome as more high-quality research in the field of breast thermography is being published. Some of the obstacles include the physical limitations of the method (e.g., the fact that it cannot identify "cold" tumors.) However, as stated in the Thermography Limitations section, cold tumors are rare and are usually less aggressive.

The rest of my report comes up with a way to close this gap using manual and automated thermography interpretation methods and proves the thermography has immense potential in the field of early diagnostics.

#cognitivepersuasion

One of the goals of my Capstone is to persuade the reader that thermography is an excellent tool for early diagnostics of breast cancer in young females. To do it, I target personal, emotional, and logical modes of persuasion that are also known as ethos, pathos, and logos.

Ethos is the ethical appeal that I used to convince the reader of my credibility. I developed this mode of persuasion by choosing an appropriate language, making myself sound unbiased by analyzing and citing influential papers that present different views on thermography, and by using correct grammar.

Pathos is an emotional appeal. I used it to evoke an emotional response in my readers. For example, I cite a credible paper that states that every eighth female is likely to be affected by cancer in her lifetime. This is a scary number that makes the readers immediately think of their personal risk or the risk of their loved ones. I also explain that breast cancers in young women are usually diagnosed later and are more aggressive. These statements add additional urgency to the problem and make readers empathetic to millions of young females who fight this disease at an early age.

Logos is the appeal to logic, where I use reason to convince my audience. For instance, I explain

the biological justification behind thermography. Mainly, how tumor formation affects angiogenesis, and why thermography is so efficient at identifying cancer in young people. I also cite papers that achieved industry-accepted performance using small datasets and prove that similar results can be achieved on bigger samples.

The purpose of the logical appeal is to cause cognitive dissonance in my readers, where there is a disconnect between what people think (we can save lives of millions of young females with the help of thermography) and do (there are no screening programs for women under the age of 40 in most countries.) Thus, my goal is to make my readers, persuaded by the report's credibility and feeling emotional about the topic, want to resolve the cognitive dissonance in their heads by, for example, learning more about screening programs in their countries or supporting early detection initiatives.

#dataviz

My implementation of this HC consists of four components: creation, analysis, interpretation, and presentation. I created all visualizations using `matplotlib` and `seaborn` libraries in Python or took them from reputable resources and carefully cited. Since I worked with high-dimensional data, the data visualization process took a fair bit of pre-processing and creativity. For example, I could not plot anomaly detection results directly since the algorithms operate in high-dimensional space. Thus, I used t-SNE plots that are particularly well-suited for the visualization of high-dimensional datasets. This approach not only gave me insights into the distribution of cancer cases but also allowed me to check for the domain shift problem. As explained in the Data section, the presence of BI-RADS-independent clusters would indicate that the model can discern the cancer status based on the bias introduced by centers, which was not the case for my data.

Secondly, I used visualizations to not only convey information to the readers but also to inform my analysis. For example, the plot in the Model Selection section illustrates the ROC curves for different models that enabled me to pick three candidate classifiers for further analysis without spending time on less suitable models.

Interpretation is another crucial aspect of this HC since one of the primary goals of data visualization is to provide insights that are hard to get from text or tabular data. For example, the SHAP plot from the Feature Selection section made me question some of the main assumptions behind the Gonzalez score by showing that high values of delta-T do not seem to have a positive effect on the probability of having breast cancer.

Lastly, I wanted my visualizations to convey information to the reader in a concise and easy-to-understand manner. I carefully thought out the purpose and placement of each data visualization and made sure that all of them a) convey information without ambiguity b) have a clear caption, legend, and axis labels c) are appropriately placed and referenced in the text.

#plausibility

One of the primary goals of my Capstone is to evaluate the plausibility of using thermography as a method for early breast cancer detection. Throughout the project, I perform various plausibility tests with increasing levels of complexity to test the strength of the hypothesis and the plausibility of three premises:

- The formation of tumors requires substantial heat generation that can be captured by thermograms.
- Humans can evaluate the change in the surface temperature using quantitative methods.
- Machine learning models can learn the differences between malignant and healthy thermograms to make reliable predictions regarding cancer status.

First, I analyzed evidence from the literature in support of using thermography both for manual evaluation and for machine learning classification. To eliminate confirmation bias, I also looked at papers that discredited this approach. In the Thermography Background section, I critiqued an influential paper by Feig et al. (1977) that reduced the acceptance of thermography in the 70s. I pointed out the non-rigorous processes and untrained technicians that confounded the results. Newer papers did better on these aspects and obtained promising results both for manual and

computer-aided methods. There is also strong biological evidence in support of thermography presented in the Background section, which proves the plausibility of this approach.

The second important plausibility test was showing that it is possible to obtain human-level performance by automatizing quantitative scores from the literature. The Preliminary Results section illustrates that the automatic score extraction algorithm that I implemented generalizes well to large datasets and has the AUC and sensitivity scores comparable to that obtained from human interpreters.

Lastly, I proved that it is possible to obtain decent classification performance using machine learning methods on an imbalanced dataset using feature extraction and data augmentation methods, as described in the respective sections. In the Results, I show that the performance of my classifier is comparable to the state-of-the-art benchmarks for thermography and mammography. Thus, the hypothesis that thermography can be used for breast cancer detection using both manual and automated methods passed all plausibility tests explored in this paper.

#studyreplication

I used the concept of replicability on multiple levels. First of all, I critically reviewed thermography and breast cancer detection literature to identify the most promising methods and approaches. I implemented these methods into the pipeline while following the best practices from the past research and avoiding mistakes such as the data leakage iSilva et al. (2016). For example, I replicated the Raghavendra et al. (2016) paper that used HOGs to classify thermograms. I not only implemented the algorithm for the feature vector extraction but also replicated their t-test value-based method comparison approach described in the Dimensionality Reduction Techniques section. However, I did not take their approach for granted. Instead, I experimented with dimensionality reduction methods presented in a broader domain of medical imagery literature and found out that the KPCA approach led to better results than the method proposed in the original paper.

On a different level, I ensured that my team members could quickly reproduce my work if needed. It is often the case that the result of one's work cannot be shared and recreated easily due to several obstacles such as different package versions, varying pre-processing steps, or cherry-picked data. To avoid such problems, I followed the company's guidelines on the environment setup and used a subset of data that was randomly selected from all patients to ensure the generalizability of results. Additionally, I followed the best machine learning practices. For example, I performed hyperparameter tuning using k-fold cross-validation and used a held-out test set only for final evaluation. While these rules seem trivial, there exist peer-reviewed publications in which a model is optimized through cross-validation and then evaluated on the same data. The final results that I report are generalizable to other subsets of Eva data and can be easily obtained by running the notebooks without changing any components.

#algorithms

The paper and the code repository both demonstrate my understanding of how to use algorithmic thinking to solve problems at multiple scales. Apart from carefully replicating algorithmic methods from the literature, some of which are secured by the property rights or pre-implemented in publicly available packages, my efforts focused on establishing the right sequence of steps that are run at the right time in a reliable and replicable manner. For example, I broke down my code into notebooks that contain essential components of the pipeline, such as feature engineering, model selection, or hyperparameter tuning. I provided a README file to make sure that reviewers look at these components in the correct order. I documented all the code inside the notebooks and util files and added detailed docstrings to all functions to demonstrate the depth of my understanding and make it easier to read. I also provide a step-by-step explanation of the logic behind some of the most complicated algorithms in the corresponding sections of the writeup using visualizations (e.g., HOG and LBP features) and concrete examples (SHAP values.)

#biasmitigation

Bias mitigation is essential when working with methods that have a controversial reputation in the scientific community. It is easy to favor information that confirms one's previously existing beliefs. Thus, to mitigate this confirmation bias, I made sure to review papers both in favor and

against breast thermography. While studies that are around 50 years old express their concerns about thermography and do not recommend it as a stand-alone approach for breast cancer screening, newer studies with modern measuring equipment and transparent methodology obtain decent classification performance. Thus, I weighed evidence from both sides and made an informed decision that thermography has excellent potential in breast cancer detection instead of believing the first article that would confirm my intuition.

Another important application of this HC is in the data preparation and model selection. First, I wanted to make sure that no irrelevant information such as data source or patient age can be used by the model to classify observations. For example, original images were not scaled to 96×96 pixels. Since images that came from the cancer center had a higher resolution, the model learned to predict the images from this center as positive independent of the thermogram. While it led to higher model performance, I realized that this performance was due to the differences in the original images. Thus, I made sure all images are pre-processed similarly, independent of the source. I used t-SNE plots to confirm that there is no clustering by the confounding variables. While the information on Eva centers cannot be shared, I also used t-SNE plots in other sections of the report, and there was no clustering in these plots.

Another important aspect that I took into account is the bias-variance trade-off of the models. Overly simplistic models cannot capture the structure of the data. Thus, they produce highly biased predictions. I used a multilayer perceptron with three hidden layers to make sure that my model is sophisticated enough that it can model the data and produce unbiased predictions while still having reasonable variance to not overfit to the data severely. The Hyperparameter Tuning section describes the methodology that I followed in further detail.

Lastly, in the Data Augmentation section, I explain that models that are trained on imbalanced datasets tend to develop a bias towards the majority class and, in some cases, always predict 0. I tried different data augmentation techniques to mitigate this behavior and managed to achieve above industry-level performance, as discussed in the Results and Conclusion sections.

#selfawareness

In this HC justification, I would like to analyze how my engagement with the project developed over time and how my strengths and weaknesses played out in the planning and execution phases. I consider myself a person who chooses their interests carefully, but once I develop a passion for something, it usually stays with me for a long time. Female health and cancer prevention specifically have long been some of my primary interests. Thus, it was easy for me to dive into the topic and spend hours analyzing it from different sides, making the preliminary research stage of the project very exciting. However, I always found it hard to summarize my findings on paper in an engaging manner. Being aware of this, I made sure that I start the report early and have most of it done by the end of the first semester and ensure I have enough time to iterate on my writing and incorporate feedback. Since it is always easy to find excuses and postpone the writing, I practiced the Pomodoro method at the beginning of some Capstone sessions until I felt fully observed by the process. Once I realized that "getting into the right mood" takes me some time, I started to allocate specific days to working on the project instead of trying to work a couple of hours every day, which did not feel productive.

While I am usually good at making plans and following them, I tend to underestimate the time required to complete a specific task. For example, I spent much time iterating on the feature engineering part and finalizing my literature review. However, I could not afford this luxury with the data augmentation methods since I did not expect that the method replication and the finalization of results would take me that much time. One important take away was that it is easy to fall victim to the planner's fallacy and be overly optimistic about the time required to complete the task. At the same time, I realized that I become much more productive under pressure. I worked the most before the assignment deadlines and Capstone class sessions, which helped me to focus and move on if I hang on certain parts unreasonably long. I also realized that it is impossible to achieve perfection, and after some iterations, the progress curve flattens a lot. Because of that, I always pushed myself to the next steps, even if I was not 100% satisfied with my results. Coming back to a specific section after some time and looking at it with fresh eyes ended up being the most effective approach. At the final project stages, as the situation in the world escalated because of the pandemic, I made sure to allocate at least three additional days solely for revision and final

HC justifications, which was a smart decision given that I had to travel to the US unexpectedly.

Something that I found especially useful is receiving feedback from my peers. At some point, I realized that I became immune to my writing and could not spot some apparent mistakes. I generally find it hard to accept critique and get frustrated if I hear negative comments. However, I also realized that I would not be able to progress further without some external insights. Thus, I asked some of my friends to critique my work, and, to my relief, they found the right balance between compliments and critique that motivated me to work even harder. At times when I became frustrated or felt like what I was doing was not enough, I told myself that the work I am doing is not just for my thesis paper but also for the good of the medical community since some parts of my Capstone will be published in the Quantitative InfraRed Thermography journal. Thus, the more negative feedback I receive, the more space for improvement I have. Other days I would just remind myself that this is the beginning of my academic journey, and while it is great to strive for perfection, getting frustrated because it is impossible to achieve is not productive.

#utility

Considering costs and benefits is crucial in the context of breast cancer detection. I explicitly applied this HC in the literature review, where I evaluate the risks of screening tools such as mammography, and prove that the cost of performing this procedure outweigh the benefits for young females due to the increased exposure to radiation, potential deformation of the breast, and decreased evaluation performance. I also explain why it is worth adding thermography to government-run screening programs of young females, even if it does not guarantee the highest specificity.

Costs and benefits also influence my choice of metrics. While the cost of incorrectly diagnosing a patient without cancer is relatively high due to the psychological and financial stress they might experience, it still does not outweigh the benefit of not missing an early stage cancer and potentially save a person's life. Additionally, patients with suspicious mammograms might have some benign conditions that require medical attention even if they do not have cancer. Thus, I prioritize sensitivity over specificity in my analysis. I also explain that the precision-recall trade-off differs in different contexts. Thus, I propose to use the ROC curves to be able to adjust the threshold accordingly and maximize the benefit.

Another application of this HC that might not be as explicit in my report is the consideration I put into how to enable the most effective interaction with the company to maximize the benefit for both sides. From my side, I wanted to make sure to bring value to the company while keeping up with Minerva's academic requirements. From the company side, it was vital that I not only focus on the academic part of the project but also meet my weekly objectives, complete some routine tasks, and start on projects that might be outside of the scope of my Capstone. For example, my manager was always willing to clarify my confusions and spend time discussing my Capstone, but he wanted to make sure it was somehow related to my weekly objectives so that the company could benefit from my efforts. Thus, I proactively sought for opportunities to work on projects that have at least an indirect connection to my Capstone. If that was not possible, I saw other projects as opportunities for growth and tried to transfer learning from one project to another. For example, I initially performed anomaly detection as a part of an automatic labeling project that did not end up in production. However, while working on it, I found a paper that explained how anomaly detection could be used to extract information from unlabeled samples and increase model performance, which I ended up adding to my final report. Despite working on unrelated projects seems to be a cost, such projects have enriched my analysis with new methodologies that I have not considered during the planning phase. One additional benefit for the company is that I could use some sections of my report as a basis for scientific papers that have now been accepted to a scientific journal. Thus, I will also provide utility to the broader scientific community by presenting my work at the conference in Portugal this July.

#interpretivelens

To ensure the highest quality of final deliverables, I approached my report and code repository from many angles and with different purposes. For example, after the submission of the final draft, I read through the report and asked myself three questions:

1. Does it make sense to someone who reads it for the first time?

2. Are there some "low hanging fruits" that will make this report much better?

3. Does it look professional?

Answering the first question enabled me to improve the readability of my writeup and include some essential details, such as figure interpretations. In this respect, getting feedback from people with different experiences and backgrounds was particularly useful. For example, Professor Ribeiro, who has experience writing scientific papers, spotted some places where she believed additional justification was needed and which I overlooked. For example, the explanation of the SHAP values and the importance of checking for data leakage. Anna, who was not familiar with the topic of breast cancer detection, pointed out that my report would be difficult to read for someone without any background in cancerology or machine learning. Thus, I followed her advice to include a glossary that explains scientific jargon. I also focused on improving some points that all of my reviewers, independent of their background, pointed out. For example, I split long paragraphs into shorter ones and added subheadings.

The second question made it possible for me to significantly improve the model performance compared to the version I presented in the final draft. In my case, the "the low-hanging fruit" was the Model Selection and Hyperparameter Tuning parts. Even though I did not plan to focus on model selection, giving priority to feature engineering and data augmentation, I realized that a vanilla logistic regression model that I used for the initial model evaluation was too simple to produce reliable results. Changing it to the optimized multilayer perceptron improved the recall of the model by 10%. This improvement was straightforward to achieve, given how optimized and user-friendly the `scikit-learn` library is.

Lastly, I took a completely different lens to check my report and code for typos and formatting errors, such as missed blank spaces or inappropriate figure sizes. I ran it through Grammarly to highlight problematic parts and read it without focusing too much on the meaning of the sentence but rather on its structure. Following the advice of my peers, I changed all first-person singular pronouns to the plural and eliminated abbreviations. I also checked my citations to ensure they follow the APA formatting.

Overall, by using these three approaches, I managed to improve the clarity, readability, and trustworthiness of my final deliverable.