Machine Learning Methods for Lung Tumour Classification

Carl Viggo Nilsson Gravenhorst-Lövenstierne
carlviggo@icloud.com

February 05, 2024

Abstract

Lung cancer stands as one of the deadliest forms of cancer. In 2020 alone, the disease claimed the lives of more than 1.8 million individuals globally [1]. Advancements in medical imaging technology, such as low-dose computed tomography, have resulted in a global accumulation of high-quality image data. However, the traditional approach of identifying and diagnosing cancer still relies on manual evaluation of such data. This workflow is time-consuming and prone to errors.

To assist clinicians, computer models specifically designed for tumour classification have recently been developed. This study aims to enhance and evaluate machine learning algorithms, specifically random forests and convolutional neural networks, in diagnosing lung cancer through the analysis of computed tomography scan images.

This study reveals that based on the analysis of the Kaggle Data Science Bowl 2017 data set [2], conventional machine learning methods outperform fine-tuned deep learning models in distinguishing between malignant and benign tumours, especially in reducing false positives and negatives.

In addition, we can conclude that the random forest model with feature extraction from the PyRadiomics library, combined with the synthetic minority over-sampling technique (SMOTE), proved more efficient on limited datasets. Compared to this, the deep learning model with feature extraction from a pre-trained AlexNet model, fine-tuned on data partly generated by a generative adversarial network (GAN), was less effective. Nonetheless, if the dataset had been larger and denser in terms of information, then the deep learning approach might have yielded improved performance.

Acknowledgements

I want to thank Dr. Mehdi Astaraki at Stockholm University, Department of Medical Radiation Physics, for his guidance throughout this project. I am also grateful for the many inspiring discussions with Daria Chamoun and for the proofreading provided by Aleksander Purik. Furthermore, I would like to express my gratitude to the Research Academy for Young Scientists (RAYS), Stiftelsen Hierta-Retzius Stipendiefond, and Stiftelsen Oscar och Maria Ekmans Donationsfond for their support in making this project possible.

Contents

1 Introduction

In 2020, cancer caused more than 10 million deaths worldwide. As our lifestyles change and the global population ages and grows, this number is expected to almost double by 2040. Among the various types of cancer, lung carcinoma, i.e. cancer in the respiratory region, is the most deadly. If the disease is detected early, the chances of survival increase significantly. [1] [3]

Improvements in medical imaging technology have made it possible to capture vast quantities of high-resolution medical images from patients. This has led to an accumulation of cancer image data worldwide. Conventional methods of assessing abnormalities in medical images include uncertain, labour-intensive methods such as manual image observation. In the image data, the presence and type of cancer is not always easily detectable, which increases the risk of misdiagnosis. This uncertainty could prohibit patients from being diagnosed early in the disease course, and thereby impact their survival chances negatively.

When cancer is detected, the malignancy of any observed tumour is classified. The next step is to identify the tumour characteristics. For instance, there exists a complex micro-environment within cancerous regions which represents different levels of aggressiveness [4]. In some cases, such characteristics can be studied by conducting an invasive surgery known as a biopsy. When a tumour is incorrectly classified and its malignancy goes undetected, it poses additional risks to the patient.

Computer-based models have recently been developed to assist clinicians with their daily work. This approach relies upon novel imaging processing techniques and machine learning models. In this project, conventional machine learning algorithms such as random forests deep learning models such as and convolutional neural networks have been applied to classify benign and malignant tumours in the lung region. Because of class imbalance and data scarcity, innovative techniques for synthetic data generation such as generative adversarial networks and synthetic minority oversampling techniques have been evaluated.

Background

A list of prerequisites and their corresponding explanations can be found in this section.

1.1 Quantitative Imaging Biomarkers

Thanks to recent improvements in medical imaging technology, for example, magnetic resonance imaging or computed tomography scans, the information content within medical images has undergone substantial growth in recent years [5]. This enhanced level of information can be utilized as a quantifiable indicator of potential disease.

Medical images provide 2D or 3D quantitative imaging data from the inside of human organs. Such information can be further processed for diagnosis, prognosis and prediction purposes in a process known as QIB, quantitative image biomarking. A subfield of QIB is radiomics, in which particular image features, radiomic features, are extracted, processed and then analyzed. [6]

Features must be chosen with respect to the set of data and the purpose of the study; in this study, lung tumour classification. Examples of features could be intensity, which corresponds to the tissue density, image entropy — a metric of how diverse the pixel values are, and image energy — a measure of variance in pixel intensity. Deep learning and machine learning models are trained to select and extract only the most relevant features with respect to the characteristics of the dataset.

1.2 Random Forests

Random forests are supervised machine learning models used for regression and classification tasks, combining the insights of multiple decision trees. Each decision tree in a random forest uses a series of conditional statements to split the data, enhancing the overall accuracy and robustness of the model.

In decision trees, the splitting condition is chosen based on information gain; the difference between entropy in the parent and child node, see equation 1. The condition that maximizes this gain is selected for the split

$$\text{Information Gain} = \Delta S = \sum_{i=1}^{c} p_i \log_2 p_i - \sum_{j=1}^{c} p_j \log_2 p_j \tag{1}$$

In which p is the probability of randomly choosing a data point in class i and c is the number of classes. The splitting continues until the entropy of all leaf nodes has reached 0, or until the parameters of the model are satisfied. To determine the information gain, all possible splitting combinations are tested, resulting in decision trees being relatively slow and inappropriate for large sets of data. Another weakness inherent to decision trees is that unless a maximum tree depth is defined, the tree will continue to split the data until all leaf nodes are in a state of zero entropy, potentially resulting in overfitting. [7]

Random forests are generally less prone to overfitting than individual decision trees. This is because they initially bootstrap, i.e., randomly select, features from parts of the dataset. These features are then used as training data by n decision trees within the forest. The model's final output is determined by the majority vote across these trees. Incorporating randomness through bootstrapping and combining votes enhances the generalizability of the model, making it more robust. [8]

1.3 Artificial Neural Networks

Artificial neural networks are data structures inspired by the biological nervous system used for function approximation purposes. The most primitive unit of a neural network is the perceptron, see figure 1.

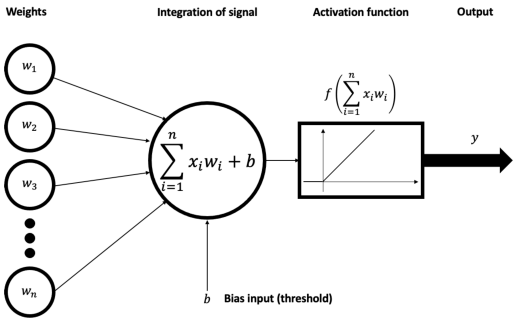

Figure 1: Schematic representation of a perceptron, including input data: x_n, weights: w_n, biases: b, weighted sum and activation function. (CC BY-SA 4.0)

In the initialisation of the model, each node connection is assigned a randomized weight and bias. The input, in the form of an RGB image matrix, is then multiplied with the weights and summarized as a weighted sum, where b is the bias:

$$\sum_{i=1}^{n}(x_i w_i + b) \tag{2}$$

The model must be introduced to a dimension of non-linearity to be able to adapt itself to the input data [9]. This is achieved by the activation function which takes the weighted sum as input. Its output either serves as input to a new perceptron or as the final classification of the model. If the activation function had been linear, then the model would only be able to learn linear relationships between the input and outputs.

The most primitive neural network is composed of an input layer, at least one hidden layer with n neurons and an output layer. In a densely connected neural network, each node is interconnected by weights. The parts of a neural network which must be manually configured, e.g. the number of nodes per layer or the activation function, are called hyperparameters. The hyperparameters make up the model architecture and can influence the performance of the model significantly. Choosing optimal hyperparameters is often the main challenge in constructing neural networks.

1.4 Loss Functions and Optimizers

In neural network training, the loss function quantifies the difference between predicted and actual outcomes, guiding the network to minimize errors. Cross-entropy loss assesses the performance of the network by comparing the predicted probability distribution with the true distribution. Optimizers, like Stochastic Gradient Descent or AdamW, iteratively compute the gradient of the loss function and adjust the weights of the network based on this gradient to reduce loss. This process, guiding the model towards higher accuracy, is part of what is known as backpropagation. The model training ideally continues until the loss function converges to a local minimum. [10]

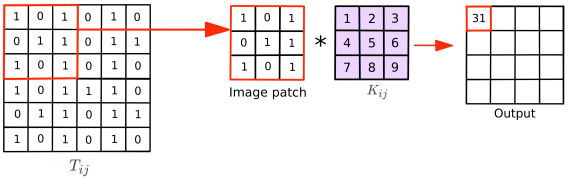

Figure 2: Schematic representation of a convolutional layer. The sum of the cross products between the image patch and the kernel is calculated and projected onto the feature map.

1.5 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are used in image analysis and automatically learn to identify useful features for categorization tasks, such as detecting edges or complex patterns in images. A typical convolutional neural network architecture is composed of convolutional layers, see figure 2, alternated by pooling layers. For classification purposes, this structure is finally interconnected with a densely connected neural network which generates the classification.

The convolutional layer transforms an input tensor, for example, an RGB image, into feature maps that contain extracted features from the data. In the domain of greyscaled medical images, this tensor is a two-dimensional matrix, see figure 2.

The process transforms an input tensor into feature maps through a kernel that moves across the tensor, as described in 3. At each value, the dot product between the kernel and the tensor value is computed, and the result is projected onto a feature map. Initially, the values of the kernel are randomized but are gradually optimized during training via backpropagation, by using the loss calculated from the output layers. As training progresses, the model extracts increasingly complex features from the input data.

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(T_{ij} \cdot K_{ij}) \tag{3}$$

Where K_{ij} is the kernel and T_{ij} is the input tensor with the dimensions $m \times n$. The position of the kernel is represented by $i \times j$. An example where the variables i, j, m, n equal 3 is visualized in figure 2.

Multiple feature maps can be extracted from a single input tensor, and are either used as input to another convolutional layer or to a densely connected neural network, where feature maps are used as predictors for classification. To minimize computing time, feature maps can be passed through pooling layers, where their dimensionality is reduced. Thus, convolutional neural networks are inherently reducing the dimensionality of the data, and work as a feature extraction method, where only more relevant features are projected onto the final feature maps.

Figure 3 visualizes examples of feature maps generated after the transformation of an image tensor by two convolutional layers, g_1 and g_2. This particular kernel extracts information about variance in light intensity.

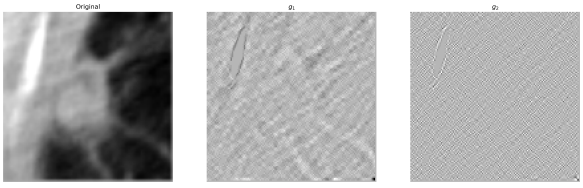

Figure 3: Application of two convolutional layers, g_1 and g_2, onto the original image.

1.6 Transfer Learning

Machine learning models tend to become overly complex, in particular when possessing too many parameters relative to the information in the training data. Just like a polynomial can be modified to fit n data points, a machine learning model can overadjust itself and thereby lose its ability to accurately classify new data. This phenomenon is known as overfitting.

Pre-trained models trained on extensive datasets such as ImageNet [11] can be imported and integrated, i.e., fine-tuned, into preexisting neural networks to improve its generalization capabilities and thereby cope with the issue of overfitting. See figure 4 for example datapoints of the ImageNet dataset.

Figure 4: 400 sample images from the ImageNet dataset [11].
(Creative Commons Licence)

1.7 Dimensionality Reduction

Applying dimensionality reduction is another strategy to prevent overfitting in high-dimensional spaces. This approach simplifies datasets by lowering the number of variables, which boosts model performance through feature selection, which identifies the most relevant features, and feature extraction, which transforms data into a lower-dimensional space while preserving vital information. Beyond dropout and pooling, which are inherently integrated into the deep learning models, this project also implements L2-regularization and sequential feature selection (SFS) for dimensionality reduction.

L2-regularization shrinks the coefficients of less important features to zero and thereby removes them from the model. This technique can be incorporated into densely connected- and convolutional layers. SFS is a feature selection method that adds or removes features iteratively based on a criterion, for example, the performance of a classifier. It aims to find the optimal set of features that improves model performance. [12]

1.8 Generative Adversarial Networks (GANs) and Synthetic Minority Oversampling Techniques (SMOTE)

Generative Adversarial Networks (GANs) are unsupervised machine learning frameworks employed to generate data with the characteristics of a given training dataset. GANs operate through an interplay between two key components: a generator and a discriminator. The loss experienced by the discriminator is used in the training process of the generator. [13]

The generator and discriminator components are typically composed of convolutional neural networks (CNNs) and densely connected networks. The discriminator's role is to assess and classify the images generated by the generator. Initially, the generator starts with random input, which is optimized to produce images that are increasingly similar to those found in the original training dataset. [13]

The SMOTE algorithm creates synthetic data points by selecting a random example from the minority class, finding its k nearest neighbours, choosing one neighbour randomly, and creating a synthetic example at a randomly selected point between the two. [14]

1.9 Model Evaluation: Metrics

When validating a machine learning model, various metrics are used to quantify its performance. While a single metric can be misrepresentative, combining multiple can give a more accurate representation of the model performance. One method for model optimization is hyperparameter tuning, where hyperparameters are adjusted while the model performance is tracked. Another method for model evaluation is K-fold cross-validation. This algorithm trains the model several times, using different validation data for each training process, and thus reduces the risk of biased datasets.

Creating a confusion matrix, containing the four possible outcomes of binary classification: true positive, false positive, true negative, and false negative, is a prerequisite for calculating accuracy, sensitivity and specificity. To clarify, a false positive diagnosis occurs when a patient's tumour is predicted as malignant when it is actually benign.

Accuracy quantifies the ratio of correctly identified samples to the overall sample count. Sensitivity, known as the true positive rate, measures the proportion of actual positives that are correctly identified. Specificity, known as the true negative rate, measures the proportion of actual negatives that are correctly identified.

2 Method

The project was divided into two pipelines: P1 and P2, where P1 involved the creation of a random forest model, data preprocessing, radiomic feature extraction, synthetic minority oversampling techniques, dimensionality reduction, hyperparameter tuning, and 10-fold cross-validation analysis. The second pipeline, P2, involved image synthetisation by applying a generative adversarial network, fine-tuning two pre-trained models with the AlexNet and InceptinetV3 architectures, and then evaluating the models using 10-fold cross-validation analysis. Finally, the the models developed in P1 and P2 were rigorously compared against one another, resulting in the selection of an optimal candidate. This pipeline is illustrated as a schematic flowchart in figure 5.

2.1 Data Preprocessing

The Kaggle Data Science Bowl 2017 dataset [2], containing 1297 low-dose computer tomography images of pulmonary nodules was used as training, testing and evaluation data. Figure 6 displays 32 samples from the dataset, where the labels '0' and '1' denote the benign and malignant classes, respectively.

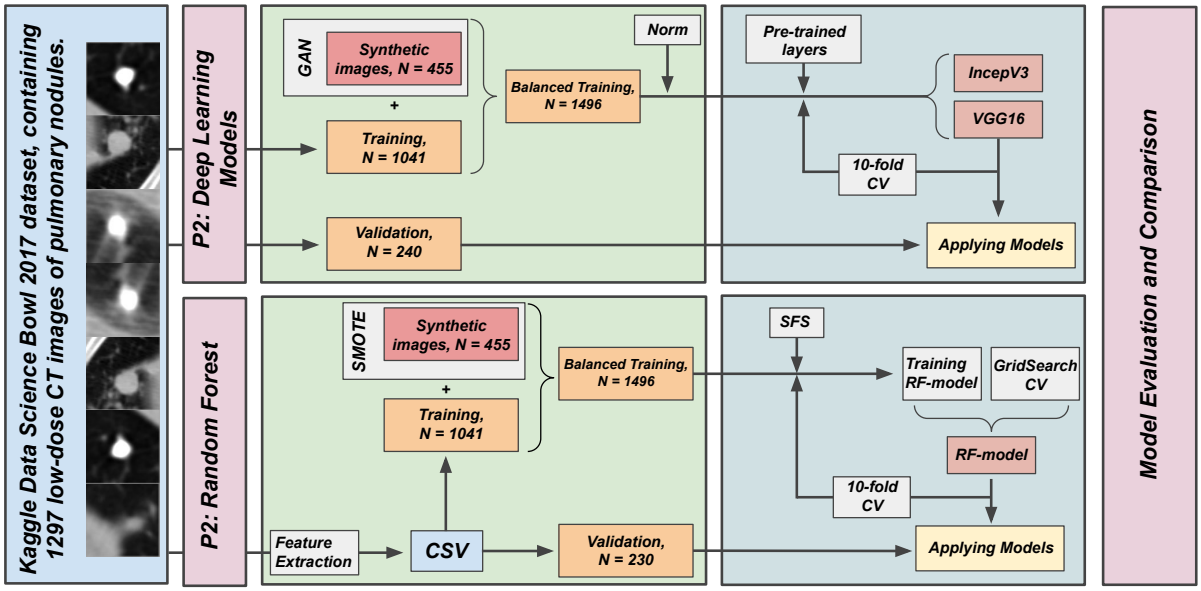

Figure 5: Flowchart describing the project pipeline; two models are developed using the same data, and are then compared against one another. CV - cross validation. N - number of samples.

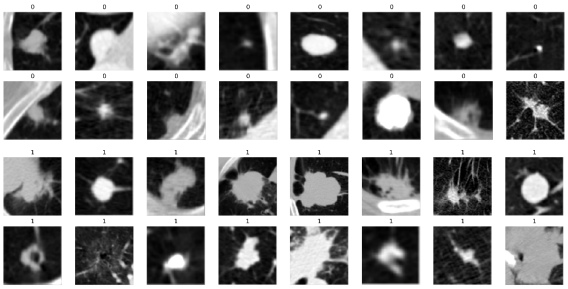

Figure 6: 32 samples from the Kaggle Data Science Bowl 2017 dataset [2]. An image is labelled as malignant if it has a title of 1, whereas a title of 0 indicates that the image is benign.

The images depicted in figure 6 are derived from low-dose computer tomography (CT) scans of the entire chest region of the patient. This is exemplified by the 2D axial images presented in figure 7.

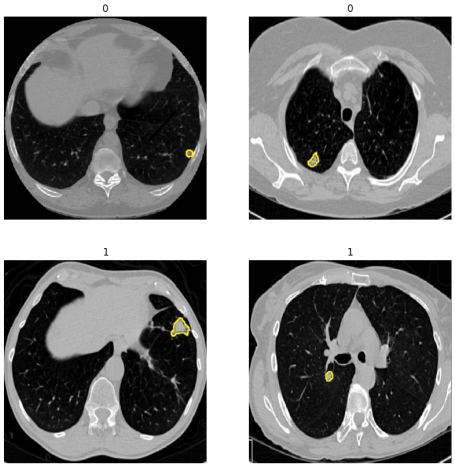

Figure 7: 2D axial CT-scans of the entire chest region, with segmented pulmonary nodules. An image is labelled as malignant if it has a title of 1, whereas a title of 0 indicates that the image is benign.

2.2 P1: Radiomic Feature Extraction

A feature extraction method from the PyRadiomics python library was implemented to extract 1070 radiomic features from the dataset. Examples of radiomic features could be energy and entropy, which measure voxel intensity distribution, as well as statistical descriptors like mean intensity, maximum intensity, and geometric properties such as volume and surface area.

2.3 P1: Synthetic Minority Oversampling Techniques

A synthetic minority oversampling technique (SMOTE) was applied to counteract the imbalance between the two classes of the dataset; benign (67.5%) and malignant (32.5%). This imbalance had a degrading effect on the performance of the model on the minority class. As a result, the relationship between the classes was balanced.

2.4 P1: Dimensionality Reduction and Feature Selection

To reduce the risk of overfitting, three different dimensionality reduction techniques were applied to the dataset sequentially: L1 regularization (lasso), principal component analysis (testing 2-31 components) and sequential feature selection (extracting 10 features during 3-fold cross-validation). The ideal method with respect to validation accuracy turned out to be the sequential feature selection, and the extracted features were used in subsequent phases of the study.

2.5 P1: Hyperparameter Tuning

The random forest machine learning model was initiated and trained on the extracted features. During training, the maximum tree depth, $0 - 200$, and the number of estimators, $0 - 120$, were tuned combination-wise via the GridSearchCv algorithm to optimize the model performance. Figure 9 captures the relationship between the different hyperparameter combinations and model performance.

2.6 P1: 10-fold Cross Validation Analysis

Ten-fold cross-validation was applied to the dataset to confirm the absence of statistical inequalities when splitting the data into training and validation sets. For each fold, the mean validation-accuracy of the model was extracted. See the the final model metrics in table 1.

2.7 P2: Generative Adversarial Network (GAN)

A generative adversarial network architecture was trained over 200 epochs, a batch size of 6, and a learning rate of 1×10^{-5}. Binary cross-entropy was used to calculate loss and Adam was used as an optimizer for both the discriminator and the generator. The training set comprised 293 low-dose computer tomography images of malignant tumours from the Kaggle Data Science Bowl 2017 data set [2]. Figure 8 illustrates the progression of the generator's performance, showing its increased refinement as the number of epochs increases. To achieve a balanced training set with a 1:1 ratio between the two classes, 455 synthetic images of malignant tumours were generated, equalizing the number of benign and malignant cases. The validation set consisted of 240 images, evenly distributed with 120 non-generated images per class.

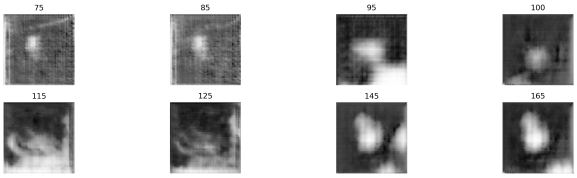

Figure 8: Feature maps produced by the generator after x epochs of training.

2.8 P2: Pretrained AlexNet InceptionV3 Fine-tuning

A convolutional neural network with the AlexNet architecture, pre-trained on the ImageNet dataset [11], containing $14,000,000$ images, was used as a feature extractor. The AlexNet was integrated with a three-layer densely connected neural network, incorporating dropout and batch-normalization layers to enhance model robustness and generalization. Leaky ReLU was selected as the activation function to address the issue of vanishing gradients. To further deal with overfitting, L2 regularization was employed across these layers. Additionally, a learning rate schedule starting at 0.01 was implemented to fine-tune the training process, adjusting the learning rate dynamically to optimize convergence and model performance. The model was thereafter fine-tuned on the training set with the added synthesised images from the generative adversarial network. The procedure was similarly applied to a pre-trained InceptionV3 architecture, replacing the AlexNet. However, this setup significantly underperformed and was quickly eliminated as a candidate. As a result, further investigation was exclusively focused on the AlexNet model.

2.9 P2: 10-fold Cross Validation Analysis

The same cross-validation procedure was applied to the AlexNet model as for the random forest model, extracting binary cross-entropy loss and validation accuracy. See table 1 for the final model metrics.

2.10 Model Evaluation and Comparison

After 10-fold cross-validation, models' performances were visualized with violin plots and Gaussian curves; validation accuracies were compared using pairwise T-tests at a 95% significance level. Sensitivity and specificity were calculated of the top-performing models.

3 Results

In this section, a comprehensive evaluation of the performance metrics from both the random forest and the AlexNet model is presented. By contrasting the models, we aim to highlight their respective strengths and weaknesses and to provide insights into their suitability for machine learning-guided lung tumour diagnosis. For clarification purposes; when we refer to 'top-performing models,' we mean those models identified as the best based on the results from 10-fold cross-validation. The InceptionV3 model was not included in the analysis because it performed significantly worse than the AlexNet model, as previously mentioned.

3.1 Random Forest Metrics

The validation accuracy of the top-performing random forest model was systematically assessed through the GridSearch-CV algorithm and compiled in figure 9. The validation accuracy initially increases rapidly and reaches a relative plateau of maxima in the intervals of 25-80 estimators and maximum depths of 110-50.

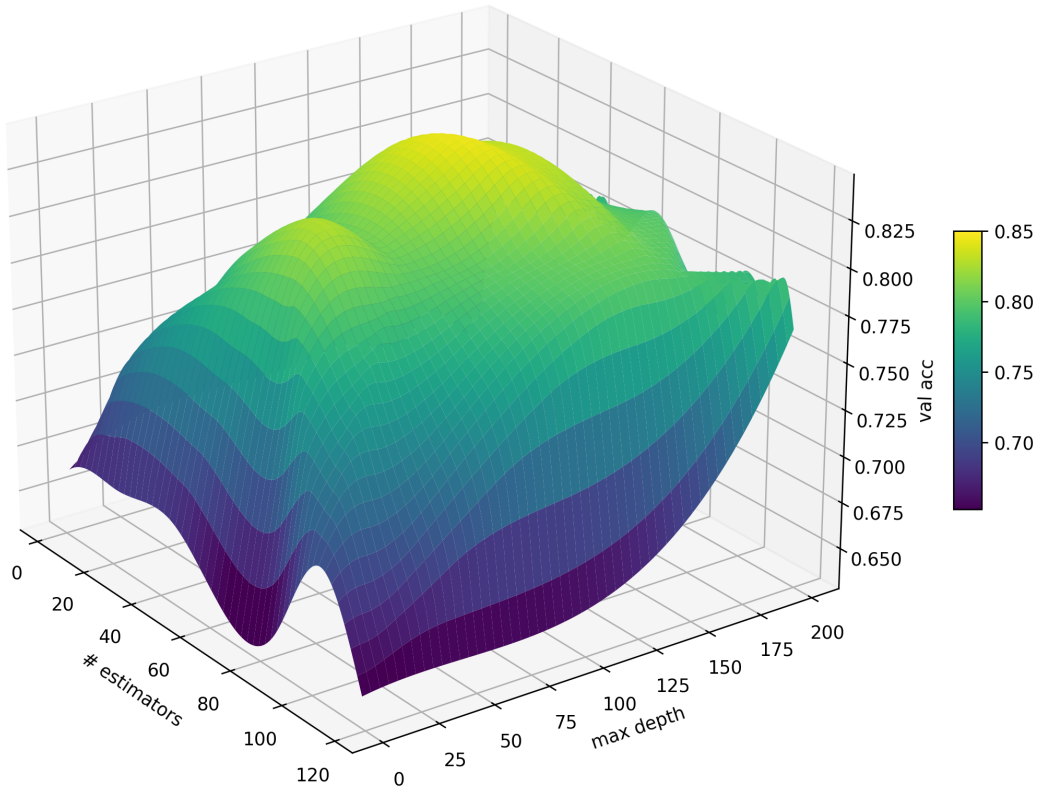

Figure 9: Visualization of mesh grid showing the relationship between estimators, maximum depth, and accuracy in model performance.

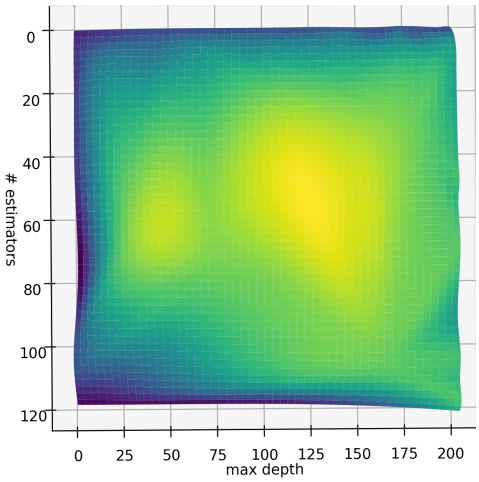

Figure 10: Complementary top-view mesh visualization of figure 9, illustrating the variance of validation accuracies across the horizontal plane.

3.2 Convolutional Neural Network Metrics

Figure 11 presents the compiled accuracy and loss of the top-performing AlexNet model from n epochs of fine-tuning the pre-trained convolutional neural network. The left graph indicates that the validation accuracy begins to converge near the 20th epoch, eventually stabilizing at an approximate mean accuracy of 0.80 with a standard deviation around 0.020. In contrast, the training accuracy moderately continues to increase, suggesting that the model is slightly overfitting on the training data. This phenomenon is confirmed by the divergence observed in the corresponding loss curves, where the training and validation loss trajectories slightly deviate from one another.

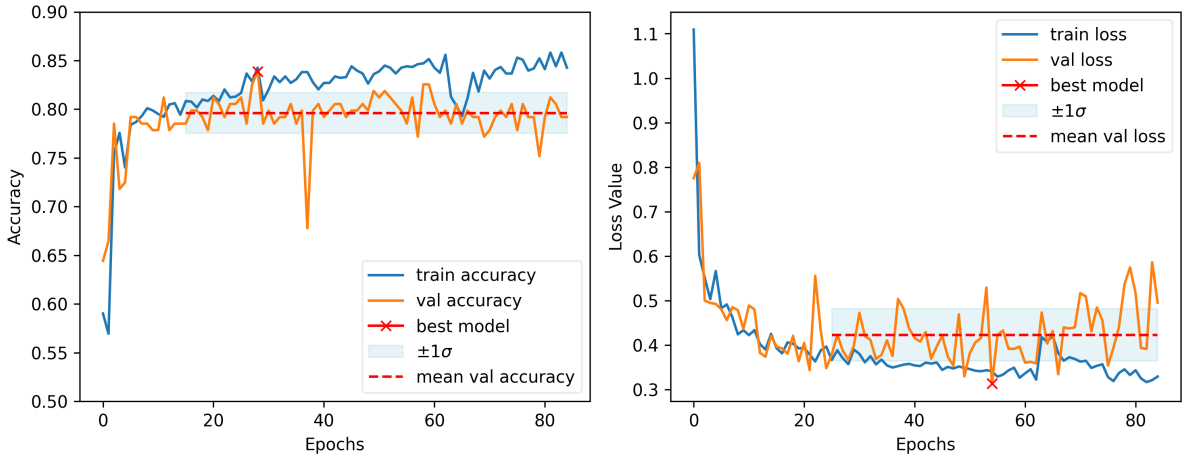

Figure 11: Fine-tuning learning curves of the top-performing AlexNet model. The orange graphs represent validation scores, while the blue graphs represent training scores. As the validation accuracy converges, the mean loss and accuracy values with standard deviation boundaries are plotted.

Table 1 displays the performance metrics for the AlexNet and random forest models, evaluated using 10-fold cross-validation. This includes the average validation accuracy and validation loss for the models, as well as specificity and sensitivity values. Note that the specificity and sensitivity metrics are computed from the top-performing model after the cross-validation folds, while the reported accuracies and losses represent the mean across all ten folds.

Table 1: Model summary.

Metric	**AlexNet**	**Random Forest**
Validation Accuracy, 10-fold	0.748 ± 0.094	0.800 ± 0.026
Validation Loss, 10-fold	0.59 ± 0.17	–
Specificity, best-model	0.78	0.97
Sensitivity, best-model	0.79	0.93

Figure 12 visually displays the distribution of validation accuracies for the random forest and the AlexNet model after 10-fold cross-validation using violin plots. The larger variance observed in the performance of the AlexNet mode can partly be attributed to an outlier model in the 6:th fold that significantly underperformed compared to the others.

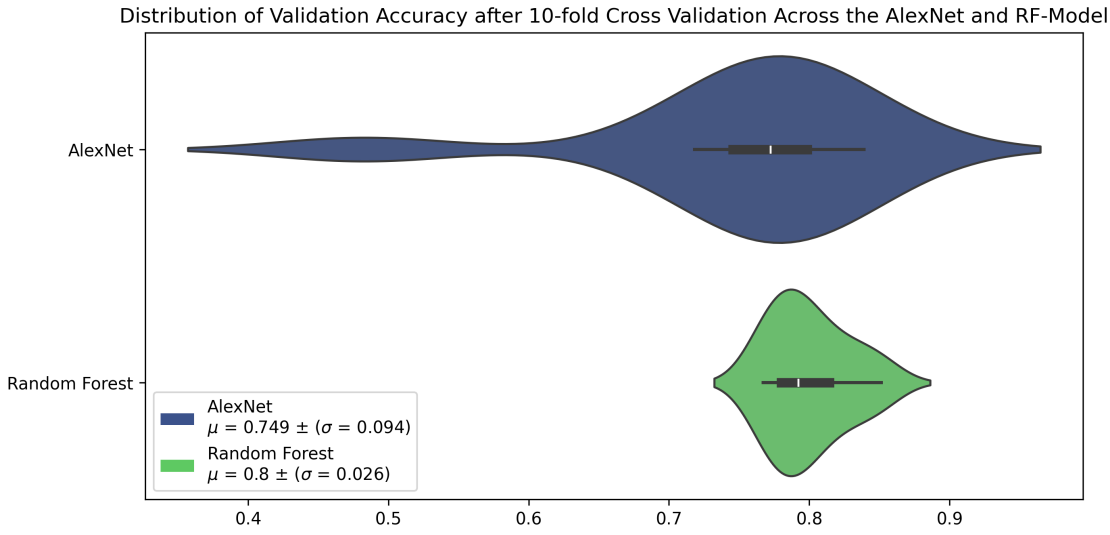

Figure 12: Distribution of validation accuracies for the random forest and AlexNet models after 10-fold cross-validation using violin plots, where the violin body represents the distribution of model validation accuracy, while the central error bar marks the first and second quartile distance.

Table 2 summarises the results of pairwise T-tests comparing the validation accuracy of the random forest and AlexNet model after 10-fold cross-validation. The data is visualised in figure 13 as two Gaussian curves, offering a visual representation of the t-test analysis. The results of the t-tests, using a significance level α of 0.05, reveal that the random forest model does not significantly outperform the AlexNet model in terms of validation accuracy. This highlights the importance of further analysis, including a closer examination of true positives, true negatives, false positives, and false negatives to gain deeper insights.

Table 2: Pairwise T-tests

Architecture-Pair	**T-statistic**	**p-value**	**Significant difference**
AlexNet vs. RF	-1.55	$p > .05$, $(p = 0.138)$	$\times$

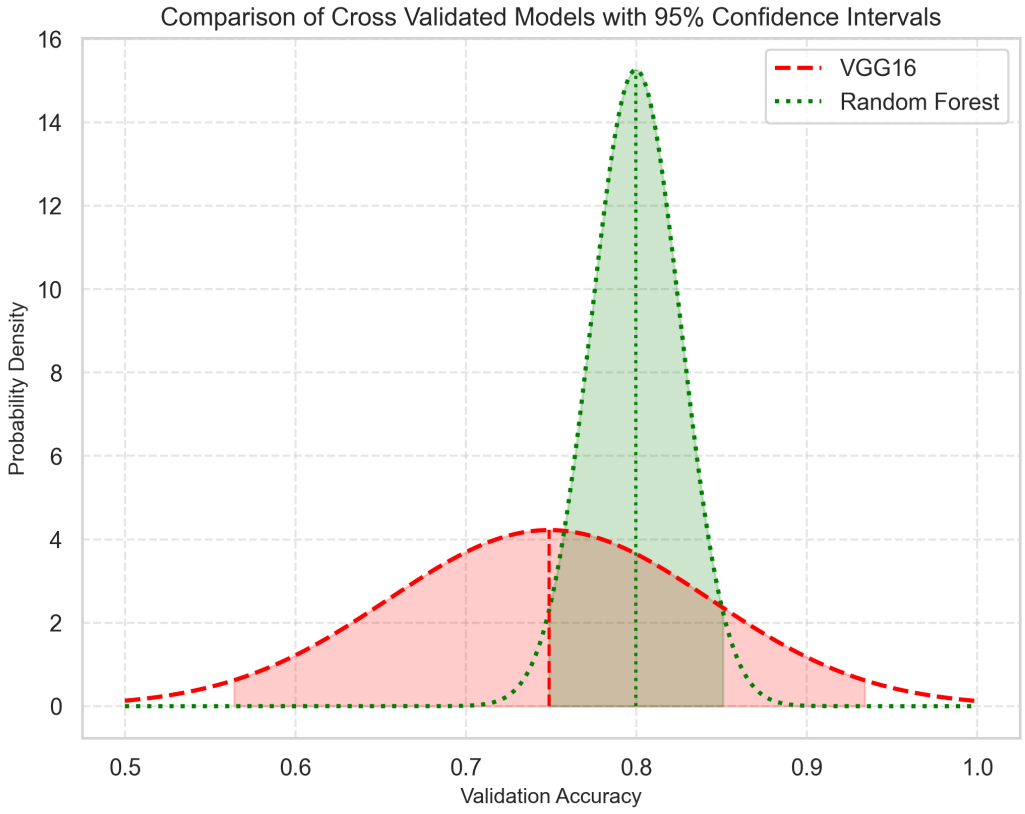

Figure 13: The distribution of validation accuracy for the two 10-fold cross-validated models is depicted as two Gaussian curves, where the area under the curves represents the 95% confidence intervals.

The confusion matrices in figure 14 reveal the details behind the reported validation accuracies. They display the counts of true negatives, false negatives, false positives, and true positives, which are used for calculating the sensitivity and specificity of the models. Sensitivity and specificity are presented in table 1. A notable distinction between the models is the random forest model's lower counts of false negatives and false positives. Such metrics are important to evaluate before deploying the model in clinical settings.

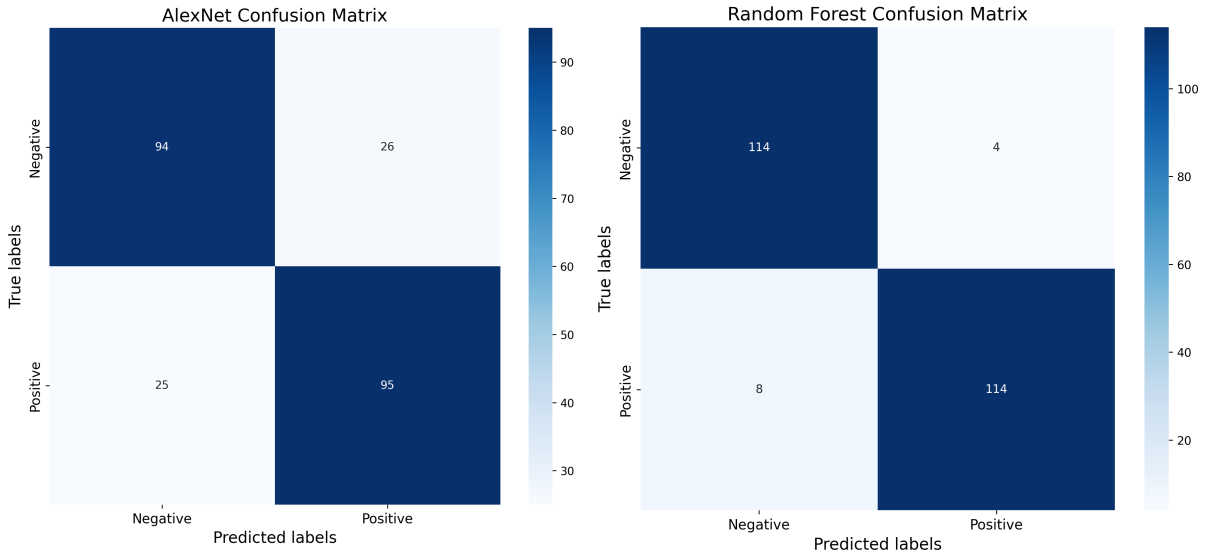

Figure 14: Confusion matrices illustrating the accuracy of the top-performing AlexNet (left) and random forest (right) model. Within each matrix, the numbers correspond to counts of true negatives, false negatives, false positives, and true positives.

4 Discussion

4.1 Model Training

Both the random forest and the deep-learning model converged during training, indicating that further training would not yield substantial performance improvements. However, the slight divergence between validation and training accuracies observed in the deep learning models could be an indication of overfitting to the training dataset, suggesting that the relationship between the amount of data in the dataset and the amount of parameters in the model, architecture is insufficient. To address this, optimizing hyperparameters such as increasing model regularization through additional dropout or L2 regularization could be considered potential measures. As the utilization of a generative adversarial network (GAN) and a synthetic minority oversampling technique (SMOTE) balanced the distribution of samples across the two classes, any performance exceeding 50% should be interpreted as an enhancement in model performance, excluding the effect of random chance. This underscores the efficacy of GAN:s and SMOTE:s in enhancing data distribution thereby balancing the learning process. In summary, we conclude that the models have converged, and the enhancements in model performances can be attributed to effective training rather than random chance.

4.2 Model Comparison

The overlapping confidence intervals and similar validation accuracies depicted in figure 13 might suggest that the performances of the random forest and deep-learning models are indistinguishable from one another. Furthermore, the pairwise t-test on the cross-validated metrics reveals no statistically significant difference in validation accuracies between the models. This raises the question if the observed differences in performance could be attributed to random chance, and highlights the need for further investigation into the details behind the validation accuracies, such as true positives, true negatives, false positives, and false negatives. In clinical settings, avoiding false negatives, i.e., diagnosing a tumour as benign when it is actually malignant, is critical and determines the ability to implement the models clinically. If a tumour is falsely diagnosed as benign, it could have the opportunity to further grow and metastasize throughout the body, potentially leading to fatal outcomes and increased diagnostic costs. [15]

The confusion matrices for the top-performing random forest and AlexNet models in figure 14 show a higher frequency of false positives and false negatives for the AlexNet model, with a nearly $1:1$ ratio, accounting for 21% of the total 240 validation samples. In contrast, the random forest model shows a much lower frequency of these errors, with a $1:2$ ratio between false positives and false negatives, meaning that false negatives are twice as likely as false positives. This results in only 5% of the 240 samples being misdiagnosed. Consequently, the random forest model demonstrates higher specificity and sensitivity. Moreover, a positive diagnosis from the random forest model is almost certainly accurate, while a negative, i.e., malignant, classification should be subject to further investigation.

4.3 Random Forest Superiority

The superior performance of the random forest model could potentially be derived from the interplay between the amount and quality of training data and the complexity of the model. While the random forest model underwent testing across a spectrum of complexity levels through the GridSearch algorithm, starting from 0 estimators and depths, the AlexNet model initially possessed a fixed number of parameters. These parameters were subsequently reduced through the application of a dropout and L2 regularization. However, this approach primarily focused on reducing complexity, and it did not systematically explore lower levels of complexity. This suggests a potential misalignment between the available dataset size and the complexity of the AlexNet model, which could have contributed to the observed differences in model performance.

In summary, the feature extraction method employed from the PyRadiomics Python library, which extracted a total of 1070 features, in combination with the synthetic minority over-sampling technique (SMOTE), demonstrated greater performance when applied to limited datasets compared to the feature extraction of the pre-trained AlexNet fine-tuned on both original and synthetically generated data produced by the generative adversarial network.

4.4 Applicability in Healthcare

A study focused on the Bahcesehir Mammographic Screening program in Türkiye evaluated the breast cancer detection rate among radiologists [16]. The study reported an initial average accuracy of 67.3% for cancer detection when evaluated solely by humans. However, when supervised by the Lunit deep learning model, this accuracy improved to 83.6%. While the random forest models developed in this study achieved an average accuracy of 80.0%, surpassing the average accuracy of the radiologists, the top-performing models still remains insufficient for clinical implementation without any human control, in particular for more mortal types of cancer than breast cancer. The much lower five-year survival rate for lung cancer compared to breast cancer, as supported by lung cancer statistics [17] and breast cancer statistics [18], underscores the urgency of implementing more extensive diagnostic procedures with higher accuracies and lower risks of false negatives. Furthermore, there are legal considerations, including the determination of responsibility, that would need to be resolved before the full integration of machine learning methods into the realm of cancer diagnosis.

4.5 Further Studies

In the preprocessing phase, the original three-dimensional NIFTI files were transformed into the 2D JPG format, which is less data-intensive. NIFTI images possess unique radiomic features, such as dimensionality, bit depth, and tissue density. For instance, lung tissue density is denoted by pixel intensity, typically ranging from -400 to 1000. However, when converted to the JPEG format, the intensity range must be adjusted to the 0 to 255 range due to JPEG's 8-bit compression. This results in information loss that could potentially degrade model performance.

Furthermore, only the axial view (the head-to-foot axis) was converted from the three-dimensional NIFTI format to JPEG. Including the sagittal (parallel to the sides of the patient) and coronal (perpendicular to the front of the patient) views would have provided a more comprehensive representation of the human respiratory region.

To circumvent these issues, the pydicom library could be employed to directly convert DICOM files into readable numpy arrays. While this approach would likely improve model performance, it would also significantly increase training time.

Moreover, the original resolution of the images was reduced to 244 x 244 x 1 to fit the VGG16 model's requirements. Given that pulmonary nodules occupy a small number of pixels, this reduction in resolution may result in the loss of critical information about their internal heterogeneity and shape, further impacting the model's effectiveness.

Additionally, all convolutional layers of the pre-trained models were utilized, which are highly optimized for the ImageNet dataset. This dataset does not necessarily reflect the characteristics of the pulmonary nodules. Reducing the number of imported layers and fine-tuning more layers could potentially enhance the generalizability and performance of the deep-learning models.

5 Conclusion

This study reveals that based on the analysis of the Kaggle Data Science Bowl 2017 data set [2], conventional machine learning methods outperform fine-tuned deep learning models in distinguishing between malignant and benign tumours, especially in reducing false positives and negatives.

In addition, we can conclude that the random forest model with feature extraction from the PyRadiomics library, combined with the synthetic minority over-sampling technique (SMOTE), proved more efficient on limited datasets. Compared to this, the deep learning model with feature extraction from a pre-trained AlexNet model fine-tuned on data partly generated by a generative adversarial network (GAN), was less effective. Nonetheless, if the dataset had been larger and denser in terms of information, then the deep learning approach might have yielded improved performance.

6 Code Access

The code, models, figures and relevant material used in this project are openly accessible on the following GitHub repository:

`https://github.com/CarlViggo/ML_lung_tumour_classification`

References

[1] Kleber, H., D., and Gold, M., S., , "Use of psychotropic drugs in treatment of methadone maintained narcotic addicts," *Ann N Y Acad Sci*, vol. 311, pp. 81–98, 1978.

[2] "Data science bowl 2017." https://www.kaggle.com/c/data-science-bowl-2017. Accessed: 07 08, 2023.

[3] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., Jemal, A, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *A Cancer Journal for Clinicians*, vol. 68, pp. 394–424, 2018.

[4] Xu, J., Liao, K., Yang, X. et al, "Using single-cell sequencing technology to detect circulating tumor cells in solid tumors," *Molecular Cancer*, 2021. https://doi.org/10.1186/s12943-021-01392-.

[5] Otero, Hansel J. and Rybicki, Frank J. and Greenberg, Dan and Neumann, Peter J., "Twenty Years of Cost-effectiveness Analysis in Medical Imaging: Are We Improving? ," *Radiology*, vol. 249, no. 3, pp. 917–925, 2008. https://doi.org/10.1148/radiol.2493080237.

[6] Park, J. E. and Kim, H. S. , "Radiomics as a Quantitative Imaging Biomarker: Practical Considerations and the Current Standpoint in Neuro-oncologic Studies," *Nucl Med Mol Imaging*, vol. 52, pp. 99–108, Apr 2018.

[7] Wikipedia contributors, "Decision tree — Wikipedia, the free encyclopedia." https://en.wikipedia.org/w/index.php?title=Decision_tree&oldid=1164832164, 2023. [Online; accessed 11-July-2023].

[8] David O'Connor and Evelyn M.R. Lake and Dustin Scheinost and R. Todd Constable, "Resample aggregating improves the generalizability of connectome predictive modeling," *NeuroImage*, vol. 236, no. 11, p. 118044, 2021. https://www.sciencedirect.com/science/article/pii/S1053811921003219.

[9] N. Kulathunga *et al.*, "Effects of the nonlinearity in activation functions on the performance of deep learning models," 2020. Funded by the National Science Foundation (NSF), grant no: CNS-1831980.

[10] M. G. M. Abdolrasol, S. M. S. Hussain, T. S. Ustun, M. R. Sarker, M. A. Hannan, R. Mohamed, J. A. Ali, S. Mekhilef, and A. Milad, "Artificial neural networks based optimization techniques: A review," *Electronics*, vol. 10, no. 21, 2021.

[11] "Imagenet." https://www.image-net.org/index.php. Accessed: 07 08, 2023.

[12] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality reduction: A comparative review," Tech. Rep. TiCC TR 2009–005, TiCC, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, October 2009. Available online at http://www.uvt.nl/ticc.

[13] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How generative adversarial networks and their variants work: An overview," *arXiv preprint arXiv:1711.05914*, 2017.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[15] N. Iragorri and E. Spackman, "Assessing the value of screening tools: reviewing the challenges and opportunities of cost-effectiveness analysis," *Public Health Reviews*, vol. 39, no. 1, p. 17, 2018.

[16] Kizildag Yirgin, I. and Koyluoglu, Y. O. and Seker, M. E. and Ozkan Gurdal, S. and Ozaydin, A. N. and Ozcinar, B. and lu, N. and Ozmen, V. and Aribal, E. , "Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis," *Technol Cancer Res Treat*, vol. 21, p. 15330338221075172, 2022.

[17] Di Girolamo, C. and Walters, S. and Benitez Majano, S. and Rachet, B. and Coleman, M. P. and Njagi, E. N. and Morris, M., "Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013," *BMC Cancer*, vol. 18, p. 492, May 2018.

[18] DeSantis, C. E. and Bray, F. and Ferlay, J. and Lortet-Tieulent, J. and Anderson, B. O. and Jemal, A. , "International Variation in Female Breast Cancer Incidence and Mortality Rates," *Cancer Epidemiol Biomarkers Prev*, vol. 24, pp. 1495–1506, Oct 2015.

A Complementary Data

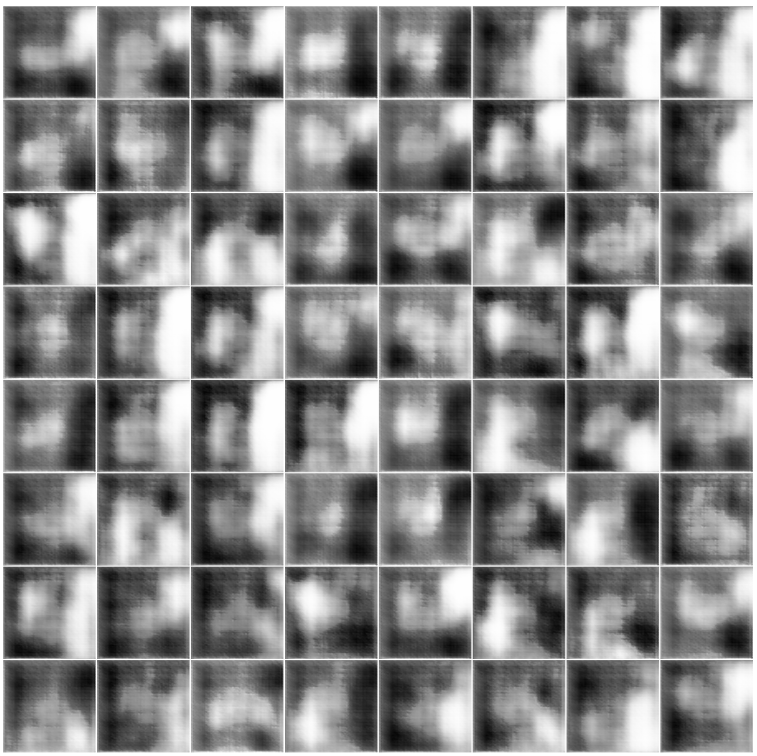

Figure 15: 64 sample images generated by the generative adversarial network used in this study.