mathematics

MDPI

Article

Semi-Supervised Approach for EGFR Mutation Prediction on CT Images

Cláudia Pinheiro [1,2], Francisco Silva [1,3], Tania Pereira [1,*] and Hélder P. Oliveira [1,3]

[1] INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[2] FEUP—Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
[3] FCUP—Faculty of Science, University of Porto, 4169-007 Porto, Portugal
* Correspondence: tania.pereira@inesctec.pt

Abstract: The use of deep learning methods in medical imaging has been able to deliver promising results; however, the success of such models highly relies on large, properly annotated datasets. The annotation of medical images is a laborious, expensive, and time-consuming process. This difficulty is increased for the mutations status label since these require additional exams (usually biopsies) to be obtained. On the other hand, raw images, without annotations, are extensively collected as part of the clinical routine. This work investigated methods that could mitigate the labelled data scarcity problem by using both labelled and unlabelled data to improve the efficiency of predictive models. A semi-supervised learning (SSL) approach was developed to predict epidermal growth factor receptor (*EGFR*) mutation status in lung cancer in a less invasive manner using 3D CT scans.The proposed approach consists of combining a variational autoencoder (VAE) and exploiting the power of adversarial training, intending that the features extracted from unlabelled data to discriminate images can help in the classification task. To incorporate labelled and unlabelled images, adversarial training was used, extending a traditional variational autoencoder. With the developed method, a mean AUC of 0.701 was achieved with the best-performing model, with only 14% of the training data being labelled. This SSL approach improved the discrimination ability by nearly 7 percentage points over a fully supervised model developed with the same amount of labelled data, confirming the advantage of using such methods when few annotated examples are available.

Keywords: semi-supervised learning; adversarial training; generative adversarial networks; medical image analysis; genotype prediction

MSC: 68T01; 68T07; 68T20

check for updates

Citation: Pinheiro, C.; Silva, F.; Pereira, T.; Oliveira, H.P. Semi-Supervised Approach for EGFR Mutation Prediction on CT Images. *Mathematics* **2022**, *10*, 4225. https://doi.org/10.3390/math10224225

Academic Editor: Jakub Nalepa

Received: 6 October 2022
Accepted: 9 November 2022
Published: 12 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

According to the 2019 report of the World Health Organization [1], trachea, bronchus and lung cancer deaths are ranked as the 6th leading cause of death worldwide. Although it was not the most common cancer in terms of new cases in 2020, lung cancer was, by far, the most lethal cancer in the same year [2]. When diagnosed at an early stage, this ailment can present a favourable prospect, with 5-year survival rates around 60% for localised cancer (limited to the lungs) [3]; however, early-stage detection remains challenging, with more than half of all cases being diagnosed when cancer has already spread to other organs, with a corresponding 5-year survival rate of only 6% [3]. For this reason, it is crucial to work on individualised treatments according to lung cancer type and stage, leaving behind the traditional approaches relying almost exclusively on chemotherapy and radiotherapy treatments for patients with advanced disease. Targeted therapies have emerged as a strategy to enhance the outcome of lung cancer, improving patient survival. Some gene mutations linked to lung cancer have already been identified, with epidermal growth factor receptor (*EGFR*) and Kirsten rat sarcoma viral oncogene homolog (*KRAS*) being the most common ones. Due to their prevalence, they are important biomarkers to be identified, although only *EGFR* has approved targeted therapies. Consequently, assessing *EGFR* mutation has become a determinant step when deciding on the possible treatments for each individual, enabling more effective patient management in precision medicine. *EGFR* mutations are usually detected using DNA extracted from tumour tissue samples obtained during biopsy or resection; however, this method is an invasive procedure with clinical implications. In this context, the need to find alternative less-invasive methods to determine gene mutation status arises. Computer-aided diagnosis (CAD) systems can play an essential role in this assessment. These systems allow clinicians to have more information (often not accessible to the human eye) to support decision-making. Therefore, the analysis of medical images such as computed tomography (CT) scans may be the key to overcoming the aforementioned problem. Medical images have already proven to be able to provide valuable information on the understanding of biological characteristics of cancer and on tumour genomic profiling [4–6]. Moreover, previous works have highlighted and revealed the connection between *EGFR* mutation status (mutant or wild-type/non-mutant) and CT scan imaging phenotypes [7–9] using supervised approaches. By establishing this link, some light is shed on a less invasive way of identifying mutations driving cancer; however, studies so far were limited to the small size of the available datasets with the *EGFR* mutation information.

Some studies have developed machine learning (ML) models to predict *EGFR* mutation status using features extracted from different ROIs. Pinheiro et al. [9] used different combinations of input features, obtaining the highest value of the averaged area under the curve (AUC) of 0.7458 ± 0.0877 with hybrid semantic features (features describing not only the nodule but also other lung structures than the nodule). Having shown the importance of a holistic lung analysis, Morgado et al. [10] presented an approach extending the latter by assessing *EGFR* mutation status using radiomic features extracted from the entire volumetric region of the lung containing the tumour instead of focusing on the nodule region only. The best-performing model recorded an average AUC of 0.737 ± 0.018. Deep learning (DL) models have shown to be able to capture relevant information and patterns directly from images, avoiding all the feature engineering processes. An end-to-end pipeline based on DL was presented by Wang et al. [11] using only the tumour-region CT images, which were previously manually identified. The developed model comprised two subnetworks. The first one shares the same structure with the first 20 layers in DenseNet, with weights acquired from the ImageNet dataset [12] in a transfer learning manner. The second subnetwork was trained with a dataset consisting of nearly 15,000 CT images to identify the *EGFR* mutation status. With this DL model, an AUC of 0.85 was achieved. Using a 3D perspective of the nodules, Zhao et al. [13] developed a 3D DenseNet framework to analyse cubic patches containing the tumour region in an end-to-end approach, attaining an AUC of 0.758. Although these studies cannot be directly compared, as they used fairly different methodologies, a tendency is evident: a holistic assessment can provide more discriminative information relating to the alterations induced by this mutation, and DL methods seem to be able to capture these patterns. These aspects are fundamental when assessing mutation status through CT images since it is not yet fully known which structures/tissues can exhibit alterations induced by genetic mutations; therefore, the entire image might contain many mineable data.

Some of the presented works considered ROIs containing only the nodule [11,13], although studies so far have demonstrated that a holistic analysis is able to provide better results. Other studies were usually limited to the small size of the available datasets with the *EGFR* mutation information [9,10]. More robust and reliable models could be developed if more data were used. Despite the availability of larger datasets with CT scans, they lack the intended labels. This is a recurrent problem in medical imaging analysis due to the evident difficulty in collecting such labels, since the process is expensive, time-consuming, and oftentimes requires additional invasive exams, as is the case when collecting labels regarding mutations that drive cancer. For this reason, the usage of semi-supervised learning (SSL) techniques, which make use of a combination of labelled and unlabelled data, going further than traditional supervised approaches, comes as a solution to overcome the problem of scarcity of annotated data and might enhance the predictive abilities of the models.

In recent years, several SSL methods have been proposed, and some works have already applied these approaches in medical imaging in order to deal with the small proportion of labelled data in the training datasets. Martins and Silva [14] used a teacher-student-based pipeline in the classification of chest X-ray images, intending to evaluate the improvement in the performance of a DL model when additional unlabelled data is used. The registered performance gain was higher when smaller datasets were used, with enhancements going as high as around 7 percentage points with only 2% of labelled data when compared to the fully supervised counterpart. Similar comparative studies were performed by Sun et al. [15] and Al-Azzam and Shatnawi [16] regarding breast cancer diagnosis in digital mammography using a graph-based approach and a self-training technique, respectively. Exploiting the power of adversarial training, Das et al. [17] and Xie et al. [18] proposed semi-supervised adversarial classification models for different tasks: breast cancer grading through histopathological images and classifying lung nodules as benign and malignant on chest CT scans, respectively. The results revealed excellent performances, with AUCs above 0.90 in both studies.

The current work intends to use a semi-supervised learning approach to predict *EGFR* mutation status using CT images. This study represents the first implementation of SSL dedicated to such a complex biomarker prediction. The mutation status prediction is not possible to identify via the human naked eye, which implies no visible features in the image related to the genotype. However, deep learning models can capture more abstract features that can be used for mutation status prediction. Additionally, the extremely small datasets with this kind of label information have limited the prediction capacity of the learning models due to the variability of the cases that not are covered by the current labeled dataset, which is expected to be an overfitting issue for the supervised learning models. SSL algorithms attempt to create more robust predictive models by taking advantage of a broader set of data and using information that unlabelled data are able to provide. Exploiting the power of adversarial training, the approach used consists of combining a variational autoencoder (VAE) and adversarial training, intending that the features extracted from unlabelled data to discriminate images can help in the classification task. This method is expected to significantly reduce the necessary labelled data required to train such a classification model. The development of this methodology contributes to: supporting medical decision-making in the use of targeted therapies by providing a method for lung cancer characterisation; the development of a DL classification model with a small labelled dataset; and the comparison of important aspects when developing such classification models, including losses applied and tackling imbalanced datasets and different proportions of an unlabelled set.

2. Material and Methods

2.1. Datasets

Two datasets with CT images were used to develop the proposed work: one including clinical data with the *EGFR* mutation status label, and the other without this label. A detailed description of each dataset is provided hereafter.

2.1.1. NSCLC-Radiogenomics Dataset

The NSCLC-Radiogenomics dataset [19] is a publicly available collection developed from a cohort of 211 NSCLC patients, comprising clinical and imaging data. The records were acquired between 2008 and 2012 and are related to patients from the Stanford University School of Medicine and the Palo Alto Veterans Affairs Healthcare System. This is a unique dataset containing imaging data paired with genomic data, including mutation status information for *EGFR* (172 patients, 43 mutant and 129 wild-type), *KRAS*

(171 patients,38 mutant and 133 wild-type), and *ALK* (157 patients, 2 translocated and 155 wild-type). In addition to CT and PET/CT scans, this dataset provides semantic annotation of the tumours in a controlled vocabulary and binary tumour masks. The latter result from a manual delineation made by a radiation oncologist. From the NSCLC-Radiogenomics dataset, just 117 patients were considered, as only these suited the following inclusion criteria: having an *EGFR* mutation test result, having an available CT scan, and binary tumour masks. The CT scans contained in this database were acquired using different CT scanners and imaging protocols, resulting in a slice thickness variation from 0.625 to 3 mm (with a median value of 1.5 mm) and an X-ray tube current from 124 to 699 mA (with a mean value of 220 mA) at 80–140 kVp (mean value: 120 kVp) [19]. From this dataset, just 117 patients were considered in the present work, and regarding the distribution of the *EGFR* mutation status for this subset, the wild-type is predominant, with mutants representing $\approx$ 20% and the wild-type representing $\approx$ 80% of the cases.

2.1.2. National Lung Screening Trial (NLST) Dataset

The National Lung Screening Trial (NLST) [20] was a randomised trial of lung cancer screening tests with 53,454 registered participants between 2002 and 2004. All the subjects were individuals considered at high risk: smokers or former smokers, with ages between 55 and 74 and at least a 30 pack-year smoking history. The study aimed to evaluate the clinical effectiveness of lung screening with chest CT. Screenings took place from 2002 to 2007 at 33 medical institutions in the United States. From the cohort, 26,722 participants were randomly assigned to screening with low-dose CT, and 26,732 were assigned to screening with chest radiography. Participants were offered three exams (T0, T1, and T2) performed annually, with the first (T0) being done soon after enrolment. All abnormalities found in the exams were recorded, and, for a CT scan to be considered positive (suspicious for lung cancer), the radiologist had to observe a non-calcified nodule or mass of at least 4 mm diameter or other suspicious findings for lung cancer. The confirmation of lung cancer was made by the NLST through medical records abstraction, and participants diagnosed with this disease did not undergo any posterior screening test in this trial. In these cases, information was documented in an additional dataset containing data about each confirmed lung cancer case, including tumour size and location. The latter encompasses the following: carina, left hilum, lingula, left lower lobe, left main stem bronchus, left upper lobe, mediastinum, right hilum, right lower lobe, right middle lobe, right main stem bronchus, right upper lobe, other and unknown. All screening examinations were performed in line with a standard protocol, which specified acceptable machine characteristics and acquisition variables, resulting in a variation of the slice thickness from 1.0 to 2.5 mm and of the tube current-time product from 40 to 80 mAs with 120 to 140 kVp of voltage [20,21]. With the data collected in this trial, one of the most extensive chest CT datasets publicly available was built. The NLST database also includes clinical data, which, along with the images, are only available for researchers through the Cancer Data Access System https://cdas.cancer.gov/plco/(accessed on 5 February 2022). Out of the 26,722 patients assigned to screening with CT, only 1089 had a confirmed cancer diagnosis, and, from those, just 622 had paired image data. This last subset was the initial collection of data considered in this work, and, due to not carrying information regarding the *EGFR* mutation status, was the unlabelled set.

2.2. Data Pre-Processing

Considering the different acquisition protocols present in both datasets, the following pre-processing techniques were employed to reduce their effect on the learning process (Figure 1). First, the distance between adjacent pixels was set at 1 mm, with further resizing to a 256 × 256-pixel resolution. Then, the pixel intensities were converted to Hounsfield units (HU) by applying a linear transformation using *min-max* normalisation. Values under -1000 HU, which corresponds to air density, were assigned to 0, and values above 400 HU, which relates to the density of hard tissues, were assigned to 1.

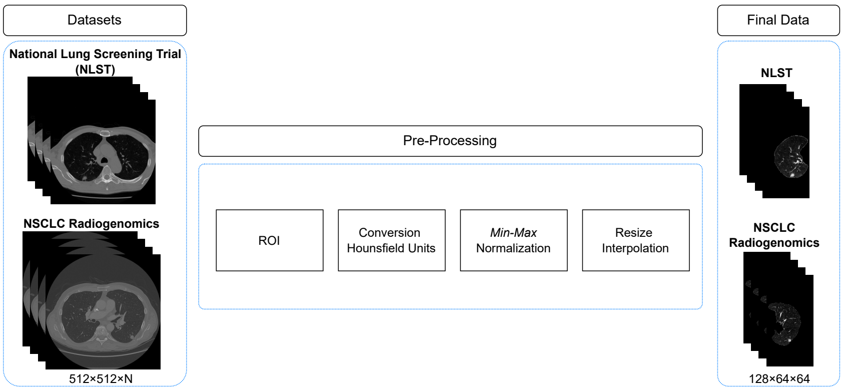

Figure 1. Overview of the pipeline developed for preprocessing the images from the two datasets.

Previous works identified the importance of not restricting the analysis to the nodule structure [22]. Additionally, the use of more "established" regions of interest for EGFR prediction in this SSL approach makes it easier to compare with the literature. The performed study used a holistic approach based on the entire lung containing the nodule in confirmed lung cancer cases. The binary masks for the lungs were obtained using a lung segmentation algorithm [22]. For the NSCLC-Radiogenomics dataset, binary masks for the nodule were available. The NLST dataset only provided the size of the tumours, the corresponding locations, and the CT scan slice number containing the largest nodule diameter. Based on the available information, each scan was cropped to the lung containing the nodule. Since the carina is located at the base of the trachea (the area where the trachea splits into the left and right bronchus), and the mediastinum is also located in the region that separates the lungs, individuals that only presented tumours in these locations were excluded. Moreover, one case was found in which both lungs exhibited primary tumours. In this situation, the two lungs were considered distinct samples (as if they belonged to different patients). This resulted in a total of 574 volumes.

In this work, it is intended to take as input 3D volumes providing information about the lung as a whole. For this reason, data uniformisation was an essential step, as CT scans from the considered datasets had a varying number of slices: from 245 to 635 in the NSCLC-Radiogenomics and between 46 and 545 in the NLST dataset. Therefore, to obtain volumes with the same number of slices, a standard depth of 64 was selected. Considering this value, the only CT scan with an inferior slice number was excluded from the NLST data collection. Additionally, the axial image size was also a challenge due to resource limitations. Therefore, each cropped slice was resized to half its size by interpolation, with (128×64) being the final axial image size. To achieve the desired standard depth, two different strategies were tested:

- Rescaling the volume, using interpolation, by a scaling factor given by

$$factor = \frac{desired\ depth}{scan\ depth} \tag{1}$$

  where the *desired depth* was set to 64 and the *scan depth* represents the number of slices of the CT scan;
- Selecting 64 uniformly spaced slices, simulating a wider space between slices. In this selection, it was imposed that the slice containing the maximum nodule mask area was included.

After the pre-processing steps, the final number of considered images from each dataset, according to the aforementioned inclusion criteria, are summarised in Table 1.

Table 1. Number of CT scans considered from each dataset.

Dataset	*EGFR* Mutation Status	# CT scans Considered
NSCLC-Radiogenomics	Available	117
NLST	Unavailable	574

2.3. Learning Models

In this study, learning models were developed for a more robust classification model for *EGFR* mutation status assessment using CT scan images as input in a combination of labelled and unlabelled data. To achieve this, the power of adversarial training was explored using a combination of an SSL generative adversarial network (GAN) and a VAE.

Autoencoders, an efficient feature extraction method, are neural networks used to learn lower-dimensional codifications (latent space) to, afterwards, generate input reconstructions. Their architecture comprises two networks: an encoder and a decoder. The former transforms the input data into an encoded representation, and the latter reconstructs, as closely as possible, the original input from the low-dimensional latent space. Thus, the decoder acts similarly to a GAN generator, projecting a low-dimensional vector to an image. A shortcoming of this kind of representation learning algorithm is that it does not allow the generation of new samples as it uses a deterministic approach. VAEs [23], generative models with a similar structure to autoencoders and a solid probabilistic foundation, replace the encoded representation by a stochastic sampling operation, learning, instead, the parameters of a probability distribution using a Bayesian approach. As the posterior distribution $p(z|x)$ (where z are the latent variables and x is the input) is an intractable probability distribution, using this variational inference, the encoder learns $q(z|x)$, a simpler and tractable distribution [24]. Typically, $q(z|x)$ is a Gaussian.

Proposed Method

In this study, the proposed method encompasses two main structures connected by a shared network: a VAE and a semi-supervised GAN, where the decoder of the VAE acts as the GAN generator, as illustrated in Figure 2.

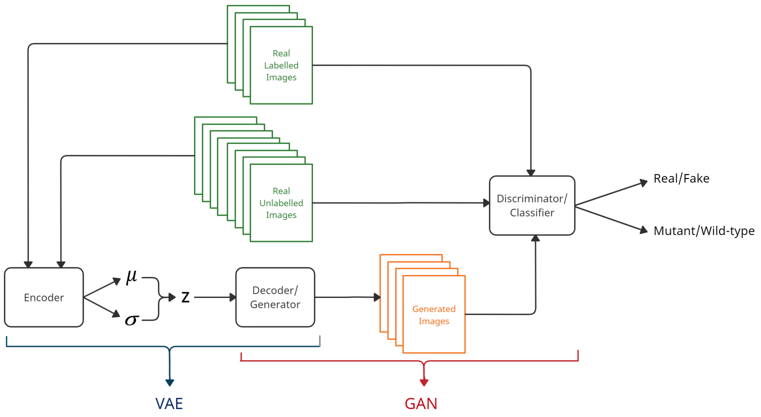

Figure 2. Proposed method, which consists of two main structures linked by a shared network: a VAE and a semi-supervised GAN, with the VAE decoder acting as the GAN generator.

The discriminator has two different outputs, both for binary classification tasks: one for the likelihood of an image being a real CT (belonging to the training data) or a generated one, and the other to classify labelled data as *EGFR* mutant or wild-type. In fact, this can be seen as having a discriminator and a classifier with a common backbone and two different output layers. The encoder of the VAE receives as input CT scan images from the training set (both labelled and unlabelled) and maps them to a distribution, providing as outputs two vectors: one representing the mean (μ) and the other representing the log-variance (σ) of $q(z|x)$. Latent vectors z, sampled from these distributions, are passed to the decoder/generator that maps the code to an image. In the generation of each of these samples z, a reparameterisation must be performed to enable backpropagation [23]. Thus, the variable z can be obtained by $z = \mu + \sigma \odot \epsilon$, where $\odot$ represents the element-wise product, and ϵ is an auxiliary noise variable (the stochastic component), $\epsilon \sim \mathcal{N}(0, 1)$. The reconstructed and original images are then provided as input to the discriminator/classifier, whose role is to undertake the two classification tasks mentioned above. To stabilise the training of the generator, avoiding a faster convergence of the discriminator early in training, the VAE part was initially fixed to be trained alone, giving the decoder/generator a better starting position.

The proposed VAE and GAN architectures were largely based on the deep convolutional generative adversarial network (DCGAN) [25]. In the case of the VAE base architecture, which was kept unchanged during the experiments and is represented in Figure 3, the encoder is similar to the DCGAN discriminator, and the decoder is similar to the DCGAN generator. Considering the GAN architecture, only a scheme of the discriminator base architecture is represented in Figure 4.

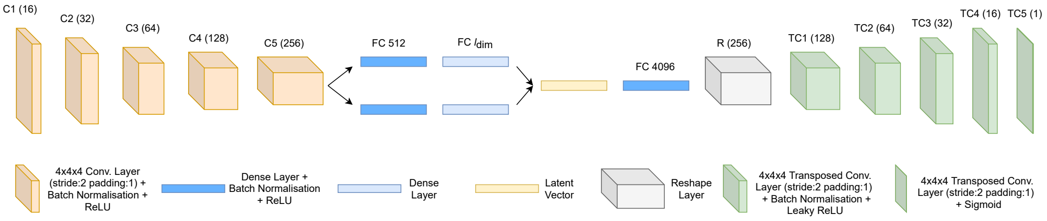
C1 (16)
C2 (32)
C3 (64)
C4 (128)
C5 (256)
FC 512
FC l_{dim}
FC 4096
R (256)
TC1 (128)
TC2 (64)
TC3 (32)
TC4 (16)
TC5 (1)
4x4x4 Conv. Layer (stride:2 padding:1) + Batch Normalisation + ReLU
Dense Layer + Batch Normalisation + ReLU
Dense Layer
Latent Vector
Reshape Layer
4x4x4 Transposed Conv. Layer (stride:2 padding:1) + Batch Normalisation + Leaky ReLU
4x4x4 Transposed Conv. Layer (stride:2 padding:1) + Sigmoid

Figure 3. Proposed VAE architecture. As can be seen, the encoder is composed of five convolution blocks (C1 to C5) and two bottleneck dense layers. Each convolution block comprises convolutional layers with an increasing number of filters, with each one followed by a rectified linear unit (ReLU) activation function and batch normalisation [26]. To reduce the size of each feature map by half, $4 \times 4 \times 4$ kernels with a stride of 2 and a padding of 1 were used. To reconstruct the original input, a latent vector with the size of the latent dimension (l_{dim}) is passed to a dense layer to, afterwards, be reshaped into 256 activation maps. This is used as input to four transposed convolution blocks (TC1 to TC4), each one composed of a $4 \times 4 \times 4$ transposed convolutional layer, with a decreasing number of filters with a stride of 2 and a padding of 1, followed by a Leaky ReLU activation (with a negative slope of 0.2) and batch normalisation. The last transposed convolutional layer (TC5) is followed by a sigmoid activation function to ensure that all output pixel values belong to the original range of values $[0, 1]$.

Additionally, a slight variation of the base architecture was tested, as depicted in Figure 4, by adding into each classification head another dense layer with a smaller number of neurons than those used in the previous layer.

Before this final base architecture was achieved, a different design for the discriminator was tested using an approach introduced by Salimans et al. [27]. The main difference between the two implementations concerned the output layer: instead of having two output layers, a single one was used with two nodes (the same number of classes in the initial supervised classification problem), and, therefore, a Softmax activation function. The unsupervised classification task used the outputs before the activation function, and a normalised sum of the exponential outputs was calculated, returning the probability of the input being fake [27].

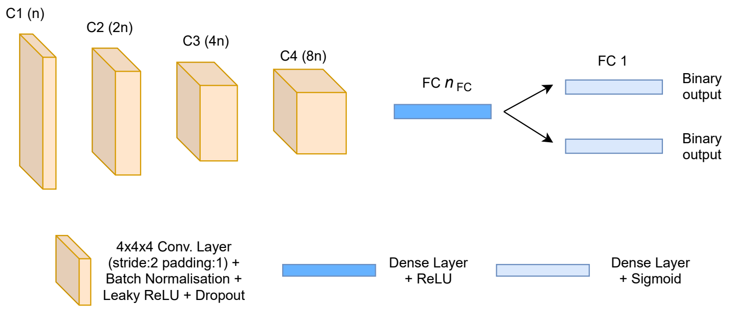

Figure 4. Proposed discriminator architecture. As illustrated, this network embodies four blocks of convolutional layers with 4 × 4 × 4, a stride of 2 and a padding of 1 for down-sampling, as well as one dense layer in the backbone followed by two classification heads. Each convolutional layer is followed by batch normalisation and uses Leaky ReLU (with a negative slope of 0.2) as the activation function. A dropout layer [28] was added after each of these convolutional blocks to decrease the number of trainable parameters, reducing overfitting. This regularisation strategy, which consists of randomly dropping out neurons during training with a certain defined probability, has the additional benefit of promoting a more robust feature extraction. Similarly to the decoder network, the number of filters increases as the network becomes deeper. Lastly, a dense layer with a variable number of neurons was included prior to the classification heads.

2.4. Training

For all the experiments performed, the labelled dataset considered was randomly split into two different sets: one for training (80%) and the other for testing (20%). With this, different training and testing groups are achieved within each random split. Given the small dimension of the dataset, only a train and test split (i.e., no validation sets created) was performed, allowing more data to be added to the training set. The divisions are performed independently, restarting all the model parameters at each one, which ensures there is no data leakage. The unlabelled set utilised was added to the training set, and the model was trained until the classifier (the discriminator) converged. To achieve a model that was as robust as possible without drawing conclusions based on a possibly biased test set and to explore data variance, 10 different random train-test splits were performed. During this training and testing process, different evaluation metrics were computed and averaged over the 10 random splits: AUC, precision, sensitivity and specificity.

2.5. Experiment Design

Different experiments were conducted in order to test some possible solutions for the model described above. In addition to hyperparameter tuning, different discriminator network architectures were tested, as well as some variations to the loss functions.

2.5.1. Loss Functions

During the training process, different loss functions were tested for the optimisation of all the networks involved, trying to find the best way to combine the VAE with the GAN with the best possible performance.

Discriminator

Starting with the discriminator, the loss functions considered for the optimisation of this network comprised an adversarial loss and a supervised classification loss used separately or combined using the average (both options were tested). In the adversarial part, as proposed in the original GAN paper [29], the goal of the discriminator is to maximise the function presented below:

$$\mathcal{L}_{adversarial} = \mathbb{E}_{x\sim p_{data}(x)}[log\, D(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (2)$$

where $\mathbb{E}_{x \sim p_{data}(x)}$ and $\mathbb{E}_{z \sim p_z(z)}$ are the expected values over the real data inputs and over the fake images $G(z)$, respectively; $D(x)$ is the discriminator probability estimation that a real image x is real; $G(z)$ is the generator output for a given input z, and $D(G(z))$ is the discriminator probability estimation that an image produced by the generator is real.

Additionally, another version of Equation (2) was tested by adding another term. The goal was to enforce the discriminator to distinguish not only images generated from latent vectors sampled from the distributions outputted by the encoder but also from random noise vectors sampled from a Gaussian distribution. Hence, the alternative adversarial loss to be maximized is given by:

$$\mathcal{L}_{adversarial} = \mathbb{E}_x[log\, D(x)] + \mathbb{E}_z[log(1 - D(G(z)))] + \mathbb{E}_{z_p}\left[log(1 - D(G(z_p)))\right] \quad (3)$$

with $z_p \sim \mathcal{N}(0, 1)$. The supervised classification loss considered was the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(Y_i \cdot log\, \hat{Y}_i + (1 - Y_i) \cdot log(1 - \hat{Y}_i)\right) \quad (4)$$

where Y_i is the ground truth label, $\hat{Y}_i$ is the predicted probability for the i^{th} image, and N is the mini-batch size. Moreover, we also tested if the addition of a manifold regularisation term would improve the overall performance. A manifold regularisation, $\Omega_{manifold}$, should enforce the discriminator to yield similar features for nearby points in the latent space.

$$\Omega_{manifold} = \left\| D(G(z)) - D(G(z + \delta \cdot 1 \times 10^{-5}) \right\|, \delta \sim \mathcal{N}(0, 1) \quad (5)$$

Decoder/Generator

For the optimisation of the decoder/generator, different combinations of loss functions (C_1, C_2 and C_3) were tested and are now described (in the equations that follow, the parameter λ represent loss weights, and the reduction or aggregation method selected is the mean):

(C_1) with a loss function composed of a reconstruction term $\mathcal{L}_{reconstruction}$ (Equation (6))—in this case, the mean squared error (MSE) between the decoder reconstruction $\hat{x}$ and the original input x, and a generator term, $\mathcal{L}_{adversarial}$. The latter is determined by maximising the log-probability of the discriminator by considering generated images as belonging to the training data or, analogously, minimizing the log-probability of the discriminator by correctly classifying fake images. This was done by applying the loss introduced by Goodfellow et al. [29] given by Equation (7) or, alternatively, the non-saturating version (Equation (8)). In such a case, the decoder/generator loss is provided by Equation (9);

$$\mathcal{L}_{reconstruction} = \frac{1}{N}\sum_{i=1}^{N}(x - \hat{x})^2 \quad (6)$$

$$\mathcal{L}_{adversarial} = \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))] \quad (7)$$

$$\mathcal{L}_{adversarial} = \mathbb{E}_{z \sim p_z(z)}[-log(D(G(z)))] \quad (8)$$

$$\mathcal{L}_{generator} = \lambda \cdot \mathcal{L}_{reconstruction} + \mathcal{L}_{adversarial} \quad (9)$$

(C_2) maintaining the same reconstruction loss mentioned above (6) and substituting the adversarial loss (Equation (8)) with the feature matching loss, introduced by Salimans et al. [27], where the generator is encouraged to synthesise data that minimises the statistical difference between the features of the real and fake data on an intermediate layer of the discriminator. Therefore, this loss is defined as follows:

$$\mathcal{L}_{feature\ matching} = \left\| \mathbb{E}_{x \sim p_{data}} f(x) - \mathbb{E}_{z \sim p_z(z)} f(G(z)) \right\|^2 \tag{10}$$

where $f(x)$ represents activations on an intermediate layer of the discriminator. Consequently, in this case,

$$\mathcal{L}_{generator} = \lambda \cdot \mathcal{L}_{reconstruction} + \mathcal{L}_{feature\ matching} \tag{11}$$

(C_3) inspired by Larsen et al. ([30]), using the feature matching loss (Equation 10) instead of a reconstruction loss combined with an adversarial loss achieved by:

$$\mathcal{L}_{adversarial} = \mathbb{E}[log(1 - D(G(z)))] + \mathbb{E}[log(1 - D(G(z_p)))] \tag{12}$$

where z_p is a sample from the prior $\mathcal{N}(0, 1)$. For this situation,

$$\mathcal{L}_{generator} = \lambda \cdot \mathcal{L}_{feature\ matching} + \mathcal{L}_{adversarial} \tag{13}$$

Encoder

Similarly to the decoder/generator optimisation, distinct combinations (C_1) and (C_2) were tested for this network (in the equations that follow, the parameter γ represent loss weights, and the reduction or aggregation method selected is the mean):

(C_1) with the traditional VAE loss, which incorporates both a reconstruction loss (6) and a latent loss $\mathcal{L}_{prior}$ (Equation (14)), the Kullback–Leibler (KL) divergence loss or relative entropy. The latter is a statistical measure and quantifies the distance between two probability distributions [31], in this case, the distribution of the encoder output and a Gaussian of mean 0 and variance 1. Therefore, for this situation, the encoder loss can be obtained by Equation (15), a loss similar to the one used in β-VAE [32] but, in this case, varying $\frac{1}{\beta}$ instead;

$$\begin{aligned} \mathcal{L}_{prior} &= \mathbb{D}_{KL}(q(zx)\ \mathcal{N}(0, 1)) \\ &= -\frac{1}{2} \sum_{i=1}^{N} (1 + log\,(\sigma^2) - \sigma^2 - \mu^2) \end{aligned} \tag{14}$$

$$\mathcal{L}_{encoder} = \gamma \cdot \mathcal{L}_{reconstruction} + \mathcal{L}_{prior} \tag{15}$$

(C_2) replacing the previously mentioned reconstruction term by the feature-matching loss (Equation (10)) while maintaining the KL divergence loss. Thus, in this case,

$$\mathcal{L}_{encoder} = \gamma \cdot \mathcal{L}_{feature\ matching} + \mathcal{L}_{prior} \tag{16}$$

2.5.2. Hyperparameters

The described networks required careful hyperparameter optimisation for fine-tuning. Table 2 presents the list of values considered for the hyperparameter manual search applied.

As the training sets used included different proportions of labelled and unlabelled data, it was decided to keep the same ratio in the mini-batch size. That is, if the used training set had, for instance, a proportion of 15% of annotated data and the remaining 85% of data without a label, the mini-batch comprised data with a similar division.

Table 2. List of values used for hyperparameter optimisation in the SSL model.

Hyperparameter	Values
Mini-batch size	8, 16, 32
Dropout discriminator	0.3, 0.4, 0.5
Learning rate discriminator	1×10^{-5}, 2×10^{-5}, 5×10^{-5}, 1×10^{-4}, 3×10^{-4}, 1×10^{-3}
Learning rate VAE	1×10^{-5}, 1×10^{-4}, 2×10^{-4}, 1×10^{-3}
Optimisers	Adam, AdamW, SGD [(a)]
Weight decay	1×10^{-7}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}
Momentum	0.1, 0.5, 0.9
l_{dim}	128, 256, 512
n [(b)]	16, 32, 64
n_{FC} [(c)]	512, 1024
γ	1, 5
λ	5, 10 ,15

[(a)] Stochastic gradient descent. [(b)] Number of filters in the first hidden layer of the discriminator network. [(c)] Number of neurons in the dense layer of the discriminator network.

2.5.3. Imbalanced Data

As usually occurs when dealing with medical diagnosis, the labelled datasets have an uneven class distribution, with the mutant type as the underrepresented class. If a classification model were built with this imbalance without any further attention given, the model would tend to be biased towards the negative classification, failing to capture the minority class. To tackle this, two strategies were tested:

- **Oversampling** the minority class by applying data augmentation techniques—horizontal and vertical flips, random rotation, and adding Gaussian noise on-the-fly, that is, without physically storing transformed images. Instead, in each mini-batch, the same number of samples for each class is used, allowing repetition of the minority samples and applying transformations to 75% of them;
- **Using a weighted loss function** during training, including in the classification loss an argument with class weights given by $\left[\frac{n_0}{n_0}, \frac{n_0}{n_1}\right]$, where n_0 represents the number of examples in the negative class (the majority class) and n_1 represents the number of examples in the positive class (the minority class) in the training set.

2.5.4. Distribution of Unlabelled Data

A final experiment relates to the percentage of unlabelled data used. To evaluate the performance when different amounts of data without labels are used to build the SSL approach, once a final model was achieved, the number of utilised training samples from the NLST dataset was reduced. Training the model using the entire NLST subset summarised in Table 1 corresponds to a percentage of around 14% of labelled data, as presented in Table 3. To investigate if a variation in the unlabelled dataset size would affect the classification performance and up to which point, different values for the percentage of NLST data used were tested: 100% (the base model), 80%, 60%, and 40%. To provide different percentages of unlabelled data to each model, random splits of the full unlabelled dataset were performed according to the desired proportion. For instance, for a percentage of 80%, the dataset was randomly divided into two groups (of 80% and 20%), giving as input to the model the intended percentage, being, in this case, the remaining 20% of data discarded. The corresponding proportions of labelled and unlabelled data used for developing the models are detailed in Table 4.

Table 3. Percentage of labelled data when using the entire NLST dataset.

Dataset	Available # Samples	# Samples for Training	% Labelled Data
NSCLC-Radiogenomics	117	94	14.07%
NLST	574	574	

Table 4. Proportions of labelled and unlabelled data for different NLST data percentages.

% NLST Considered	% Labelled Data	% Unlabelled Data
100	14	86
80	17	83
60	21	79
40	29	71

3. Results

Numerous experiments were performed, and the best outcomes were provided when using the base architecture for the discriminator represented in Figure 4, with the following combination of loss functions:

- Discriminator—using the $\mathcal{L}_{adversarial}$ depicted in Equation (3) and the $\mathcal{L}_{BCE}$ propagated separately through the network;
- Decoder/generator—applying the $\mathcal{L}_{generator}$ summarised in Equation (13);
- Encoder—optimised with the $\mathcal{L}_{encoder}$ presented in Equation (16).

The base model with the best performance was obtained with the set of hyperparameters presented in Table 5. The model with all these settings defined was trained in the remaining scenarios: tackling the imbalance present in the labelled set by oversampling or by using a weighted loss; rescaling the image depth by interpolation or by selecting 64 linearly separated slices; and adding a manifold regularisation term (Equation (5)) to the discriminator loss.

Table 5. Set of hyperparameters that led to the best-performing SSL model.

Hyperparameter	Values
Mini-batch size	32
Dropout discriminator	0.5
Learning rate discriminator	1×10^{-5}
Learning rate VAE	2×10^{-4}
Optimisers	Adam
Weight decay	1×10^{-4}
l_{dim}	128 [(a)]
n [(b)]	64
n_{FC} [(c)]	512
γ	5
λ	10

[(a)] As no significant differences between the three tested values were found, a size of 128 was selected to reduce the number of trainable parameters. [(b)] Number of filters in the first hidden layer of the discriminator network. [(c)] Number of neurons in the dense layer of the discriminator network.

3.1. Ablation Studies

3.1.1. Imbalanced Data and Rescaling the Depth of the Volumes

Starting with the strategy to overcome the imbalance found in the dataset, Table 6 presents the achieved performances when considering the two tested approaches. The results are also discriminated by the technique used to rescale the depth of the considered volumes. Using a weighted loss function instead of oversampling the minority class provided better results, which was more noticeable when the rescaling of the number of slices per image was performed with interpolation, registering an AUC of 0.701 ± 0.114 averaged over 10 random train-test splits. Thus, in the results provided hereafter, this was the technique applied.

Table 6. Performance results for the *EGFR* mutation prediction when considering different strategies to handle imbalanced data. The evaluation metrics are presented as mean ± standard deviation averaged over 10 random train-test splits.

Strategy Imbalanced	Image Depth Rescaling	AUC	Precision	Sensitivity	Specificity
Weighted loss	Interpolation	**0.701 ± 0.114**	0.425 ± 0.140	**0.660 ± 0.254**	0.742 ± 0.134
	Linearly spaced	0.597 ± 0.106	0.302 ± 0.087	0.500 ± 0.248	0.693 ± 0.121
Oversampling	Interpolation	0.610 ± 0.071	**0.433 ± 0.227**	0.420 ± 0.189	**0.800 ± 0.129**
	Linearly spaced	0.592 ± 0.106	0.420 ± 0.287	0.300 ± 0.241	0.884 ± 0.077

3.1.2. Add a Manifold Regularisation Term

When evaluating the effect of adding a manifold regularisation term to the discriminator loss, the results show that such a term is more advantageous when the strategy used for rescaling is the selection of linearly separated slices, as detailed in Table 7. While with the interpolation method, the model achieved worse performance, using a manifold regularisation with the former strategy increases the average AUC from 0.5974 ± 0.1060 to 0.6274 ± 0.1372.

Table 7. Performance results when considering the addition of a manifold regularisation term to the discriminator loss.

Image Depth Rescaling	AUC	Precision	Sensitivity	Specificity
Interpolation	**0.665 ± 0.125**	**0.367 ± 0.134**	**0.640 ± 0.250**	0.690 ± 0.140
Linearly spaced	0.627 ± 0.137	0.350 ± 0.216	0.460 ± 0.284	**0.795 ± 0.101**

3.1.3. Proportions of Unlabelled Data

Lastly, different percentages for the NLST data were tested. The results for the two models that achieved better results in the previous experiments are detailed in Tables 8 and 9. It should be noticed that a fully supervised model (using exclusively labelled data) is only presented in Table 8, since the addition of the manifold regularisation was added upon the SSL approach. As can be observed, reducing the amount of unlabelled data given as input to the models results in slightly worse performances. The variation between different percentages (80%, 60% and 40%) is almost inexistent when no manifold regularisation is applied, being small when this term is added. Nevertheless, utilising unlabelled data, even if in smaller amounts, provides better models than not using them at all.

Table 8. Performance results for different percentages of the unlabelled dataset used when considering a weighted loss function.

	Percentage Unlabelled Dataset Used				
Metric	**100%**	**80%**	**60%**	**40%**	**0%** [(a)]
AUC	**0.701 ± 0.114**	0.659 ± 0.120	0.651 ± 0.1101	0.655 ± 0.153	0.632 ± 0.063
Precision	**0.424 ± 0.140**	0.388 ± 0.188	0.343 ± 0.124	0.368 ± 0.153	0.286 ± 0.042
Sensitivity	0.660 ± 0.254	0.560 ± 0.265	0.560 ± 0.265	0.620 ± 0.244	**0.800 ± 0.155**
Specificity	0.742 ± 0.134	**0.758 ± 0.158**	0.737 ± 0.125	0.689 ± 0.134	0.463 ± 0.131

[(a)] Fully supervised model—baseline model.

Table 9. Performance results for different percentages of the unlabelled dataset used when considering a weighted loss function and a manifold regularisation term.

	Percentage Unlabelled Dataset Used			
Metric	**100%**	**80%**	**60%**	**40%**
AUC	**0.665 ± 0.125**	0.631 ± 0.120	0.618 ± 0.105	0.613 ± 0.130
Precision	0.367 ± 0.134	**0.388 ± 0.188**	0.348 ± 0.108	0.333 ± 0.115
Sensitivity	**0.640 ± 0.250**	0.560 ± 0.265	0.520 ± 0.223	0.500 ± 0.272
Specificity	0.689 ± 0.140	0.726 ± 0.148	0.726 ± 0.123	**0.726 ± 0.112**

The addition of the unlabeled data does not produce a monotonic increase in the performance; however, there is a more marked increase with the addition of all the unlabeled data, which seems that the final addition of the data brings some variability relevant to the learning model to improve the capability of the prediction.

The combination that achieved the best results was using a weighted loss function to tackle the imbalance of the labelled data without adding a regularisation term, with a mean AUC of 0.701 ± 0.114, as illustrated in the mean ROC curve of Figure 5 with the baseline method.

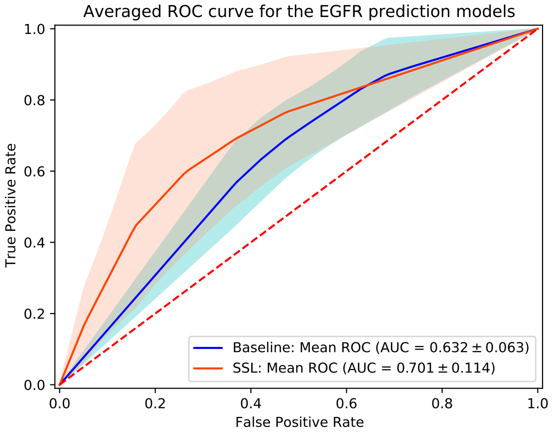

Figure 5. Averaged ROC curve for *EGFR* mutation status prediction with the baseline and best-performing SSL model. For each train-test split, the ROC curve is computed. The blue and orange lines represent the arithmetic average ROC curve for each model, and the shaded areas depict the corresponding standard deviation. The red dashed line illustrates the ROC curve of an at-chance classifier.

4. Discussion

Although this SSL approach contains architectural blocks with generative purposes, the discriminative part was the most exploited here. Not only was the quality of the generated images not expected to be very close to real (mostly due to the choice of $\mathcal{L}_{generator}$), their reality was not aimed at, since that would mean that the concrete imaging manifestations associated with *EGFR* mutation status would be accurately found, which we did not expect in advance given its extreme difficulty. Instead, we focused on finding regularities at a lower level using the feature space extracted by the discriminator. Furthermore, as stated in [33], when it comes to semi-supervised tasks using GANs, good classification performance is not compatible with a realistic generator output. Using feature matching results in better semi-supervised learning performances but, as a drawback, generates worse images.

The results of this research display the difficulty of detecting relevant and significant features that could be related to *EGFR* mutation status. Conditioned by the limited amount of labelled data available, we tried to achieve a more robust classification model by incorporating unlabelled data in a semi-supervised approach. Even with this extra data without annotation, the task has proven to be quite challenging and susceptible to the train-test split variations, as can be observed by the high values of standard deviation across all experiments. This high variation can also be a consequence of the small number of included *EGFR* mutant patients (23 cases), with it being desirable to add extra samples of this class to verify if the variation would be reduced.

When tackling the imbalance present in the data, using a weighted loss function proved to be a better approach when compared to oversampling the minority class. This has the additional benefit of reducing the required training time as the number of images is fewer. Undersampling the majority class was never an option in this work given that one is dealing with the problem of data scarcity, and dismissing valuable data seems counterintuitive.

When using the full unlabelled dataset instead of only a portion of it, the model benefits more from the additional information provided by this data collection. Furthermore, the increased performance is even more notorious when comparing the model with a fully supervised baseline, which uses only the labelled data as input. A direct comparison with related works concerning the effect of such variations cannot be made given that, traditionally, the variation of the labelled-unlabelled data proportions is performed by adding labels to the desired part of the unannotated data and not by removing data without labels, as was tested in this study.

The manifold regularisation term aimed to approximate the information extracted by the discriminator according to how closely the data points were located in the latent space (space mapped by the generative encoder). However, the demonstrated decrease in discriminative performance (although not significant) can be possibly explained by the idea that close latent data points actually belonged to different classes, which could be related to the difficulty of correctly approximating the Bayesian posterior for such a complex task: this approximation implies navigating through explainable factors that are not well-known by clinicians yet and with a vast space of complex structures within the lung that can possibly be wrongly associated with EGFR mutation status.

Regarding related works in *EGFR* assessment in lung cancer CT scans, to the best of our gathered knowledge, no other research has attempted a 3D deep learning approach using the entire lung volumetric region in a semi-supervised fashion. Furthermore, SSL methods are typically tested using extended labelled datasets and by simply removing the label of a significant portion of the data, simulating the existence of a large unlabelled set. Approaches that combine different datasets in a similar way as presented in this work are difficult to find. Additionally, no study was found combining the two datasets used in this task (NSCLC-Radiogenomics and NLST). Silva et al. [34] developed a DL model based on transfer learning methods using 2D CT scan slices from the NSCLC-Radiogenomics dataset. Utilising the analysis of the lung containing the nodule as ROI, a mean AUC of 0.68 ± 0.08 was achieved. Comparing the results, the developed SSL approach was able to slightly increase this performance using the same labelled dataset.

Although the performance results obtained in this work are promising but still not very good, they are aligned with the performance obtained in the previous works. The current work suffers from the inability to find concrete visual manifestations associated with EGFR mutation status, which represents a transversal challenge in any machine learning application. The susceptibility against spurious correlations when trying to extract these very complex relations is often manifested by overfitting and lack of performance stability (here shown by the high variation in test set results), something that is only emphasized when dealing with smaller sets of training data. The possibility of bringing more semantic discriminative information to decisions, which by being connected with the feature extractor will influence what is being captured as relevant or not, might be a further alternative to enhance the generalization power of the system.

Limitations

No direct comparison can be made in terms of state-of-the-art results since some methodologies use feature engineering processes; others that use end-to-end DL models develop them with large, non-public datasets.

Another important aspect relates to the combination of two different datasets that may have different characteristics. These may include distinct stages of cancer that might be translated into different visual manifestations of the target variable. By combining distinct datasets, it is assumed that such manifestations, to exist, are similar in terms of image patterns, though there is still no clear evidence that this happens.

The developed work was constrained by hardware limitations, which, when analysing images in a 3D perspective, may be cumbersome given the density of the networks involved. If this problem could be overcome, it would be helpful to explore, in the future, more complex architectures that might be able to capture more abstract patterns, given the demonstrated complexity of the task. Furthermore, although features extracted from the unlabelled data to discriminate images helped when classifying scans as mutant or wild-type, more samples from the minority class, which included only 23 images, would probably be needed for a more accurate model.

VAE was implemented in adversarial training, and various percentages of unlabeled data were tried to train the model. However, other SSL methods, such as co-training or graph-based methods, can be implemented in the future, and their capacity to predict EGFR mutation status can be compared.

Overfitting is always a concern for small datasets. Some strategies were implemented to mitigate this effect, such as dropout and weight decay; however, with such a small training dataset, overfitting still occurred to some extent.

Despite not reaching remarkable results, the developed work may be seen as a stepping stone from which subsequent works can improve upon the highlighted limitations.

5. Conclusions

A personalised treatment plan presents the opportunity to improve lung cancer patient outcomes. In the era of precision medicine, the identification of driver mutations in lung cancer brought new treatment options and helped increase the overall survival rates. For this reason, a complete cancer characterisation is of the utmost importance to choose the best treatment for each individual. This opens doors to the use of artificial intelligence, which is gaining ground in the medical field as the utilisation of images such as computed tomography scans has already proven to allow the detection of relevant patterns and relations. Despite the success of deep learning models when it comes to analysing such medical images, the lack of labelled data makes their development difficult.

This study aimed to provide an end-to-end lung cancer characterisation by analysing the entire volumetric region of the lung containing the nodule, using CT images, in a semi-supervised approach. The method employed to integrate both labelled and unlabelled data consisted of a combination of a VAE and a GAN. The best-performing classification model achieved a mean AUC of 0.701 ± 0.114. Despite not largely improving the performance results in this task, the utilisation of the additional unlabelled dataset brought more discriminative power to the model. This was further evidenced by the increase in terms of performance when more unlabelled data were used, resulting in an improvement of circa 7 percentual points in the mean AUC compared to a fully supervised model developed with the same labelled set. It should be noticed that the best model was built only with 14% of the data containing a label. Adding an unlabelled dataset, in an SSL fashion, improved the performance of the predictive deep learning model, allowing the development of a better-performing end-to-end model with a reduced amount of labelled data.

Author Contributions: C.P., F.S., T.P., and H.P.O. conceived the study and designed the methodology. F.S. performed data curation. C.P. developed the software, performed the analysis of the results, and drafted the manuscript. All authors provided critical feedback and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT-Foundation for Science and Technology Portugal, within project LA/P/0063/2020, and a PhD Grant Number: 2021.05767.BD.

Data Availability Statement: We acknowledged the National Cancer Institute for the access of National Lung Screening Trial (NLST) dataset, and The Cancer Imaging Archive (TCIA) Public Access for the publicly available Non-Small Cell Lung Cancer (NSCLC)-Radiogenomics Database used in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. The Top 10 Causes of Death. Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed on 30 October 2021).
2. Cancer Today-International Agency for Research on Cancer. Available online: https://gco.iarc.fr/today/home (accessed on 7 March 2022).
3. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer Statistics, 2021. *CA A Cancer J. Clin.* **2021**, *71*, 7–33. [CrossRef] [PubMed]
4. Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.; Grossmann, P.; Cavalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. [CrossRef] [PubMed]
5. Gillies, R.; Boellard, R.; Dekker, A.; Aerts, H.J.W.L. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef]
6. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
7. Bodalal, Z.; Trebeschi, S.; Nguyen-Kim, T.D.L.; Schats, W.; Beets-Tan, R. Radiogenomics: Bridging imaging and genomics. *Abdom. Radiol.* **2019**, *44*, 1960–1984. [CrossRef]
8. Digumarthy, S.R.; Padole, A.M.; Gullo, R.L.; Sequist, L.V.; Kalra, M.K. Can CT radiomic analysis in NSCLC predict histology and EGFR mutation status? *Medicine* **2019**, *98*, e13963. [CrossRef]
9. Pinheiro, G.; Pereira, T.; Dias, C.; Freitas, C.; Hespanhol, V.; Costa, J.L.; Cunha, A.; Oliveira, H.P. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Sci. Rep.* **2020**, *10*. [CrossRef]
10. Morgado, J.; Pereira, T.; Silva, F.; Freitas, C.; Negrão, E.; de Lima, B.F.; da Silva, M.C.; Madureira, A.J.; Ramos, I.; Hespanhol, V.; et al. Machine Learning and Feature Selection Methods for EGFR Mutation Status Prediction in Lung Cancer. *Appl. Sci.* **2021**, *11*, 3273. [CrossRef]
11. Wang, S.; Shi, J.; Ye, Z.; Dong, D.; Yu, D.; Zhou, M.; Liu, Y.; Gevaert, O.; Wang, K.; Zhu, Y.; et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.* **2019**, *53*, 1800986. [CrossRef]
12. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S.; et al. ImageNet Large Scale Visual Recognition Challenge. *CoRR* **2014**, *115*, 211–252.
13. Zhao, W.; Yang, J.; Ni, B.; Bi, D.; Sun, Y.; Xu, M.; Zhu, X.; Li, C.; Jin, L.; Gao, P.; et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Med.* **2019**, *8*, 3532–3543. [CrossRef]
14. Martins, R.A.P.; Silva, D. On Teacher-Student Semi-Supervised Learning for Chest X-ray Image Classification. In *Anais do 15 Congresso Brasileiro de Inteligência Computacional*; Filho, C.J.A.B., Siqueira, H.V., Ferreira, D.D., Bertol, D.W., de Oliveira, R.C.L., Eds.; SBIC: Joinville, Brazil, 2021. pp. 1–6. [CrossRef]
15. Sun, W.; Tseng, T.L.B.; Zhang, J.; Qian, W. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput. Med Imaging Graph.* **2017**, *57*, 4–9. [CrossRef]
16. Al-Azzam, N.; Shatnawi, I. Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. *Ann. Med. Surg.* **2021**, *62*, 53–64. [CrossRef]
17. Das, A.; Devarampati, V.K.; Nair, M.S. NAS-SGAN: A Semi-supervised Generative Adversarial Network Model for Atypia Scoring of Breast Cancer Histopathological Images. *IEEE J. Biomed. Health Inform.* **2021**, 26, 2276–2287. [CrossRef]
18. Xie, Y.; Zhang, J.; Xia, Y. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. *Med Image Anal.* **2019**, *57*, 237–248. [CrossRef]
19. Bakr, S.; Gevaert, O.; Echegaray, S.; Ayers, K.; Zhou, M.; Shafiq, M.; Zheng, H.; Benson, J.A.; Zhang, W.; Leung, A.N.; et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **2018**, *5*, 180202. [CrossRef]
20. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N. Engl. J. Med.* **2011**, *365*, 395–409. [CrossRef]
21. Aberle, D.R. The National Lung Screening Trial: Overview and Study Design. *Radiology* **2010**, *258*, 243.
22. Silva, F.; Pereira, T.; Morgado, J.; Cunha, A.; Oliveira, H.P. The Impact of Interstitial Diseases Patterns on Lung CT Segmentation. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; pp. 2856–2859. [CrossRef]
23. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
24. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2016**, arXiv:1606.05908.
25. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.

26. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 2015; Volume 37, pp. 448–456.
27. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In Proceedings of the Advances in Neural Information Processing Systems; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
28. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
29. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
30. Larsen, A.B.L.; Sønderby, S.K.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of The 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2015.
31. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Series in Telecommunications and Signal Processing; Wiley-Interscience: Hoboken, NJ, USA, 2006.
32. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
33. Dai, Z.; Yang, Z.; Yang, F.; Cohen, W.W. Good semi-supervised learning that requires a bad gan. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3272.
34. Silva, F.; Pereira, T.; Morgado, J.; Frade, J.; Mendes, J.; Freitas, C.; Negrão, E.; De Lima, B.F.; Silva, M.C.D.; Madureira, A.J.; et al. EGFR Assessment in Lung Cancer CT Images: Analysis of Local and Holistic Regions of Interest Using Deep Unsupervised Transfer Learning. *IEEE Access* **2021**, *9*, 58667–58676. [CrossRef]