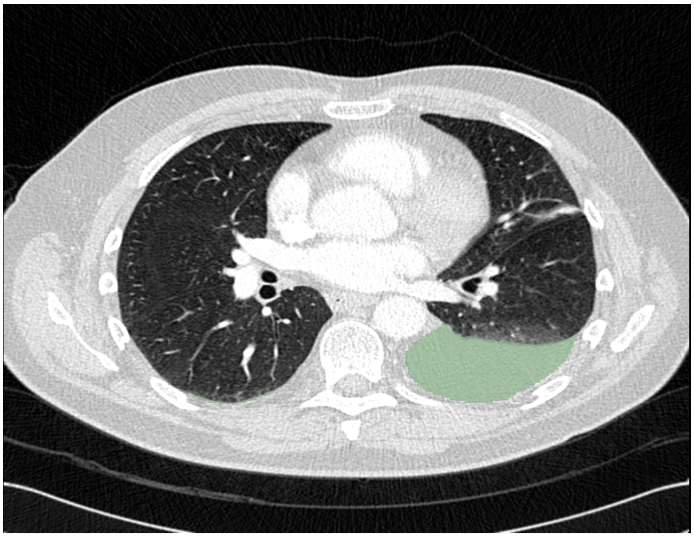
Survival prediction using Convolutional Neural Networks based on lung CT scans of patients with Malignant Pleural Mesothelioma

Information Science
University of Amsterdam
Bachelor thesis

Isa-Ali Kirca - 12014672

June 24, 2021



Supervisor UvA: Dr. T.R. Langerak
Co supervisor UvA: Dr. S. van Splunter

Supervisor NCI: K.B.W. Groot Lipman, MSc
Co supervisors NCI: Dr. S. Trebeschi
Z. Bodalal, MD MSc

Scope of the study: 18 European Credits
Number of words: 4981

Abstract

Tumors and cancers are cells with accelerated cell divisions that are masses and are worldwide one of the most common contributors to death. Malignant Pleural Mesothelioma (MPM) is an aggressive lung cancer and is caused by exposure to asbestos. Late-state patients are treated with non-surgical treatments like radiation, chemotherapy, targeted or immunotherapies. In this study two state-of-the-art Artificial intelligence models are trained and evaluated with three experiments on CT scans of patients with MPM. To get better metrics for determining tumor sizes over time, predict survival rate and possibly save a patient from a lot of suffering, this study analysed whether CT scans had predictive values for survival. Three experiments have been conducted. The whole CT scan, the segmented lungs from the whole CT scan and the segmented abnormalities from the segmented lungs were pre-processed and trained on Residual Networks and 3D CNNs using data augmentation operations, all in the same way. The total volume per CT scan was calculated. The result shows that the Residual Network with 6 filters for the first convolutional layer, has shown the best performance on the test set, slightly better than a random classifier with ($AUC = 0.549$, $p = 0.047$) 55% area under curve of the receiver operating characteristic and an validation accuracy of 64%, performing better than all other models on the various experiments. Although, a weak negative correlation ($r =$ -0.199, $p < 0.001$) has been found between the difference in days between death date minus scan date and the volume of abnormalities in mm3. Hence, based on the experiments conducted in this study, can be determined that CT scans of patients with MPM, do not show any predictive values for survival with the conducted methods.

Keywords: Convolutional Neural Network, CT scans, Lung segmentation, Malignant Pleural Mesothelioma, Residual Network, Survival prediction.

Contents

1 List of Abbreviations **4**

2 Introduction **5**

3 Objective and relevance of the study **7**
3.1 Scientific relevance: . . . 7
3.2 Social/Societal relevance: . . . 7

4 Theoretical framework **8**
4.1 Convolutional Neural Networks (CNN) . . . 8
4.1.1 Feature extraction . . . 8
4.1.2 Classification . . . 9
4.2 Residual Neural Networks (ResNet) . . . 10

5 Methods **11**
5.1 Dataset . . . 11
5.2 Labels and expected output . . . 11
5.3 State-of-the-art models and loss function . . . 11
5.3.1 3D Convolutional Neural Network . . . 12
5.3.2 Residual Network (ResNet) . . . 12
5.3.3 Loss function . . . 12
5.4 Experiments and image pre-processing . . . 13
5.4.1 Experiment 1: Whole CT scan . . . 13
5.4.2 Experiment 2: Lungs including abnormalities . . . 14
5.4.3 Experiment 3: Abnormalities . . . 15
5.5 Experiments evaluation and used hard/software . . . 16
5.6 Post-processing . . . 16

6 Results **17**
6.1 Experiment 1: Whole CT scan . . . 17
6.2 Experiment 2: Lungs including abnormalities . . . 19
6.3 Experiment 3: Abnormalities . . . 20
6.4 Post-processed experiment . . . 21

7 Discussion **22**
7.1 Interpretation results . . . 22
7.2 Limitations . . . 23
7.3 Future perspectives . . . 23

8 Conclusion **24**

References **25**

1 List of Abbreviations

Various abbreviations are used throughout this study. The following table shows the explanation of each abbreviation with the corresponding page where the abbreviation is used for the first time.

Abbreviation	Explanation	Page
AI	Artificial Intelligence	6
AUC	Area Under Curve of the Receiver Operating Characteristic	16
CT	Computed Tomography	5
FPR	False positive rate	18
MPM	Malignant Pleural Mesothelioma	5
NaN	Not a number	21
ResNet	Residual Network	10
ROC	Receiver Operating Characteristic	16
TPR	True positive rate	18

2 Introduction

Tumors and cancers are cells with accelerated cell divisions that are masses and are worldwide one of the most common contributors to death (Ferlay et al., 2020). The term cancer covers a variety of malignant tumors, each with various subcategories, since cancer can originate in any type of cell (Cooper & Hausman, 2000). For example, there are various forms of lung, breast and brain tumors. These various subcategories of tumors can come in both benign and malignant forms, whereas cancer is the malignant form.

Malignant Pleural Mesothelioma (MPM) is an aggressive lung cancer and is caused by exposure to asbestos. After inhalation, asbestos fibers embed in the pleura of the lungs and cause inflammations and scarring (Manning et al., 2002). These inflammations and scarrings could lead to mesothelioma tumors, with a latency period of around 40 years between the exposure of the asbestos fibers and symptoms occurring (Bibby et al., 2016).

MPM is usually diagnosed through multiple tests, scans and biopsies. Once the disease is diagnosed, it is often treated with chemotherapy, surgery, radiation therapy and/or immunotherapy. However, surgical treatment is omitted if a patient has late-stage disease. These tumors are treated with non-surgical treatments like radiation, chemotherapy, targeted, or immunotherapies. The prognosis of a patient varies depending on their individual case, since the biological behavior of mesothelioma is unpredictable (Opitz & Weder, 2018). In addition to patients' individual risk factors, the lack of correlation between clinical and pathological staging makes it difficult to allocate the best treatment for each individual case (Opitz & Weder, 2018).

Monitoring radiographic changes is crucial to determine therapy response of tumors over time. Pass et al. (1997), Frauenfelder et al. (2011), Gill et al. (2012) and Rusch et al. (2016) found that measurement of tumor volume assessed on Computed Tomography (CT) scans can predict treatment outcome. Clinical response assessment criteria, such as RECIST, analyze follow-up imaging using simple size-based measurements, such as the axial diameter of tumors (Eisenhauer et al., 2009).

Quantitatively assessing MPM tumor burden showed prognostic significance in tumor volume measurement (Murphy & Gill, 2017). These measurements are good prognostic metrics of treatment response (Plathow et al., 2008). Murphy & Gill (2017) has shown that it is useful to compare these measurements to RECIST and Modified RECIST (mRECIST) (Lencioni & Llovet, 2010) in patients undergoing chemotherapy and tumor volume measurement on CT to assess treatment response. RECIST determines the longest diameter of the tumor in several CT scans at the same location by a radiologist in order to monitor the growth. However, RECIST is suboptimal for MPM, since there are several ways to determine a diameter (Figure 1), which results into higher inter- and

intra-observer variability. mRECIST are relatively new guidelines to circumvent this problem (Tsao et al., 2011), but suffers from the same problems, although less. Determining the total volume is more accurate, but labor-intensive. By automatically segmenting the total tumor volume, the change in tumor volume can be determined in a cost-effective and reproducible way (Tsao et al., 2011).

Liu et al. (2010) assessed this treatment response in patients with MPM by measuring tumor volume on CT before and after therapy, which suggested that there was a significant association between the reduction in tumor volume after treatment and improved survival. Frauenfelder et al. (2011) found that RECIST is a less reliable measure of response to chemotherapy and a predictor of outcomes than measuring tumor volume on CT scans.

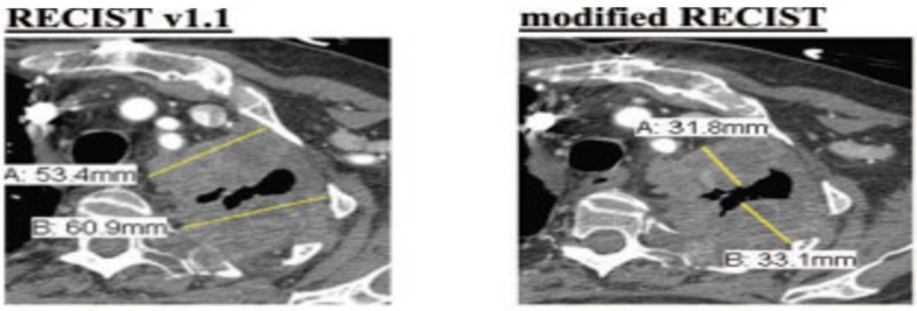


Figure 1: Multiple ways to track growth of tumor through diameter, total segmentation would be more accurate (Tsao et al., 2011)

Artificial Intelligence (AI) has had a major impact on medical imaging as machines keep getting better at representing and interpreting complex data, which allows for a quantitative assessment of radiographic tumor characteristics. In particular, Machine Learning and Deep Learning models are becoming widely used models in medical image-recognition tasks. These models are able to match and even surpass humans in task-specific applications (Hosny et al., 2018).

State-of-the-art models on predicting survival of the patient are amongst others based on Convolutional Neural Networks (CNN). These networks could automatically extract imaging features and identify nonlinear relationships in complex data (Xu et al., 2019). The research question investigated in this study is: To what extent can the survival rate of a patient diagnosed with Malignant Pleural Mesothelioma be predicted on the basis of CT scans?

The rest of this study is organized as follows. First, the objectives and relevance of the study are described. Second, overall introductions about Convolutional Neural Networks and Residual Neural Networks are given. Third, the methods and experiments to classify whether a patient lives for another year or not are described, followed by the results of testing. Finally, discussions in addition with recommendations for future work and a conclusion are presented.

3 Objective and relevance of the study

Chemotherapy and immunotherapy are tough for patients suffering from Malignant Pleural Mesothelioma. Investigating whether changes in the volume of the tumor in CT scans correlate with the survival of a patient can save a patient a lot of suffering. Instead of tracking growth of tumor through diameter, total segmentation of the tumor volume would be more accurate (Tsao et al., 2011). Segmentation of the tumor can potentially improve the performance of predicting survival rate, thus inclusion and evaluation of this segmentation plays an important role in the investigation of this correlation.

3.1 Scientific relevance:

Current literature assessed the change in tumor volume by diameter in different CT scans, but little to no research is performed on the change in tumor volume based on the total volume of the lungs or abnormalities, training an AI model with this data and then having an AI model automatically predict survival for each patient with various CT scans. The reason for this study is to narrow this gap and to get better metrics for determining tumor sizes over time and predicting survival based on this data. The results of this study can provide insight into whether the change in tumor volume correlates with the survival rate of a patient.

3.2 Social/Societal relevance:

If better estimates can be made of the patient's survival, empowerment of the doctors' and the patients' therapy choices will be made (therapy is tough on patients). This could lead to the avoidance of unnecessary suffering for the patient, while the patient can be treated in a more cost-effective way.

4 Theoretical framework

4.1 Convolutional Neural Networks (CNN)

A regular CNN is a neural network and typically consists of single or multiple layers which can often be subdivided into input layers, convolution layers, pooling layers, fully-connected layers and an output layer. As figure 2 shows, a simple CNN for classification has two parts. In the first part features are extracted and thereafter either classification or regression tasks are carried out.

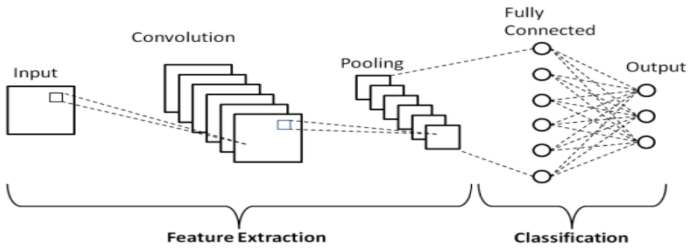


Figure 2: Block diagram of a simple CNN (Phung & Rhee, 2018)

4.1.1 Feature extraction

The feature extraction part of a CNN always consists of an input layer and often consists of multiple convolutional and pooling layers. For volumetric data the input layer of a CNN consists of the matrix with voxels of the image as input. The shape of this input is specified by a fixed size, where the images may need to be resized accordingly before being fed into the network. Once these images are fed into the network, the following layers are part of the feature extraction:

1. *Convolutional layers*
   The most important components of a CNN architecture are the convolutional layers. Linear and nonlinear combinations are combined, i.e. convolutional operations (see figure 3) and activation functions. These operations are computed using kernel filters with the same dimension as the input image but with a smaller size, resulting in feature maps. Feature maps are results of estimations of the dot product between weights and kernel filters of each step the filter makes over the entire input image (Singh et al., 2020).

   The most common activation functions are ReLU and Tanh and are applied directly to the feature maps. The main difference between these activation functions is that ReLU returns 0 in case the input value is negative, otherwise the value is returned. Tanh in contrast takes any real value as input and returns values from -1 to 1. Hence larger input values are closer to an output of 1, while smaller input values are closer to an output of -1.

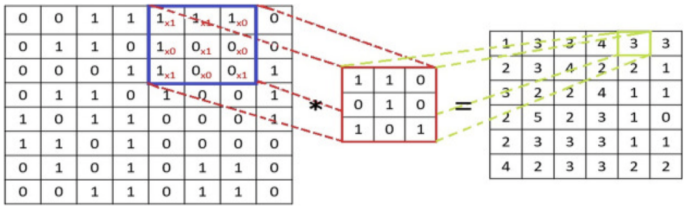


Figure 3: Example of a convolutional operation (Singh et al., 2020)

2. *Pooling layers*
   Pooling layers are layers which are often added after each convolutional layer. Specifically, after the application of the activation function on the feature map which is returned by the convolutional layer. It is basically down sampling of an image (Sultana et al., 2018). The size of the pooling filter is smaller than the size of the feature map, hence the most used size is 2x2. The most common pooling operations are:

   (a) *Max Pooling* The maximum pixel value of the feature map is selected, see figure 4a.

   (b) *Average Pooling* The average pixel value of the feature map is calculated, see figure 4b.

(a) Max pooling **(b)** Average pooling

Figure 4: Most common pooling operations

4.1.2 Classification

After feature extraction, the features are weighted and combined in fully connected or dense layers.

1. *Fully connected layers*
   Fully connected or dense layers are the last component of the CNN architecture and are connected to all the information acquired in the previous layers.

2. *Output layer*
   The output layer of a CNN with a classification task often has either Softmax or Sigmoid as an activation function. The softmax is used for a multi-class classification task (sum of all probabilities has to be 1), while the sigmoid is used for binary classification tasks.

4.2 Residual Neural Networks (ResNet)

ResNets have shown impressive performance on image classification tasks (Bello et al., 2021), hence these ResNets comprise of multiple Residual Units as displayed in figure 5. While a typical ResNet also consists of multiple convolution layers, it has the choice of skipping connections, which enables identity mapping (He et al., 2016).

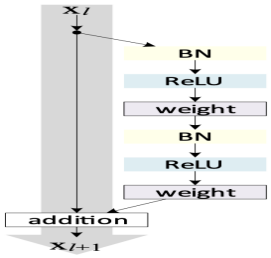


Figure 5: Residual unit proposed by He et al. (2016)

As shown in figure 6 there are various usages in different orders of Residual units which can be tested. However, the principle stays the same. Figure 6a shows the original ResNet, while figures 6b, 6c, 6d and 6e show additions proposed by He et al. (2016).

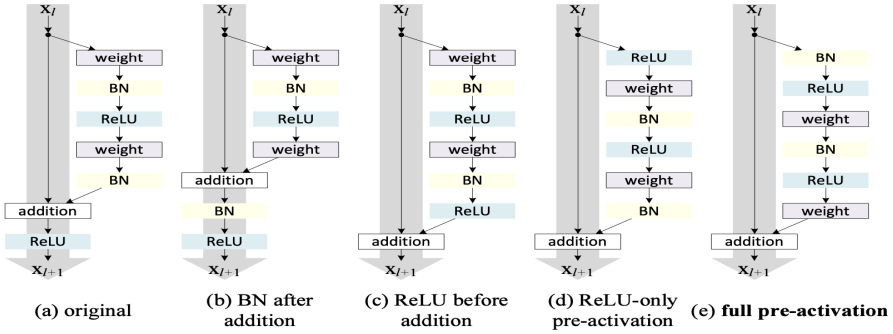


Figure 6: Various usages of Residual units proposed by He et al. (2016)

The identity mappings obtained from the convolutional layers are added to the information which is passed through the skip connection. Hence, by relying on the skip connections direct identity mappings can be learned. After the addition of the identity mappings and the information which is passed through the skip connection, it could be passed through the activation function ReLu or as proposed in figures 6c, 6d and 6e the activation function could be used in the indirect mapping.

5 Methods

5.1 Dataset

The dataset used in this study, consisted of 2347 CT scans of advanced stage Malignant Pleural Mesothelioma patients undergoing immunotherapy or chemotherapy. These CT scans were collected at the Netherlands Cancer Institute (NKI, Amsterdam), but acquisition was at hospitals throughout the Netherlands. Each patient's CT scan at the time of diagnosis as well as their follow up CT scans with dimensions 512 x 512 x slices were included (i.e. height x width x depth). CT scans of patients that are still alive were also included to prevent a bias (consisting of only patients which passed away) and each follow up scan with the biggest file size was chosen, i.e. with the most slices (depth).

The dataset was split in a train, validation and test set. However, since both CT scans at time of diagnosis as well as their follow up CT scans were included, the split was based on the number of patients (357), to prevent that one patient's CT scans were in multiple sets and guarantee the independence of each set. The train set comprised 249 patients with 1606 CT scans, whereas the validation comprised 53 patients with 350 CT scans and the test set comprised 55 patients with 391 CT scans.

5.2 Labels and expected output

Besides the dataset of CT scans, a dataset of each patient's diagnosis date, therapy start date and end date was provided by the Netherlands Cancer Institute. The subtraction of the patient's therapy start date from the scan's date led to the labelling of each scan. A treshold of one year has been set to obtain a 50/50 distributed soft labeled data. Patients alive, with an end date in June 2021, were labeled with a soft label (labels from 0 up to 1) of at least 0.5. CT scans with a difference lower than 182 days were labeled as 0, a difference between 182 and 548 days were labeled with soft labels in the range of [0,1], whereas CT scans with a difference greater than 548 days were labeled with label 1.

In this study survival is defined as whether the patient lives for another year or not. Hence the output was rounded to 1 if it was greater than 0.5 and rounded to 0 if it was lower than 0.5. This cutoff was based on the soft labeling process, where patients who were still alive (end date in June 2021) and patients who lived longer than a year (difference greater than 365 days) were labeled with at least 0.5.

5.3 State-of-the-art models and loss function

This study evaluated two state-of-the-art models, namely a 3D Convolutional Neural Network and a Residual Network. To learn presentations from data which contain volume, 3D CNNs are powerful networks to use for classifcation

tasks (Zunair et al., 2020). The original Residual Network is a 2D model, but the one used in this study is a 3D implementation.

5.3.1 3D Convolutional Neural Network

A specific form of a 3D CNN proposed by (Zunair et al., 2020) is a 3D CNN consisting of 17 several layers such as 3D convolutional layers, max pooling layers, batch normalization layers and a fully connected layer (figure 7). Both the 17 layer deep 3D CNN as well as the small additions made model were evaluated.

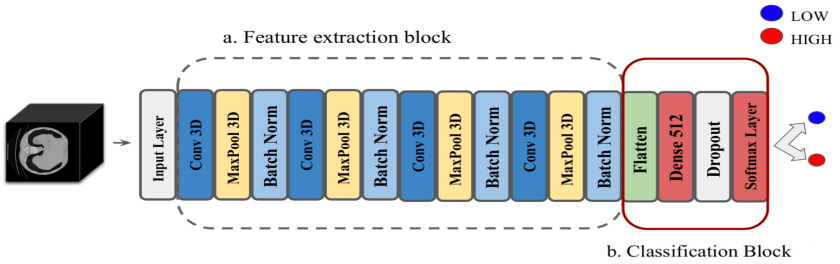


Figure 7: 3D CNN with 17 layers proposed by (Zunair et al., 2020)

5.3.2 Residual Network (ResNet)

ResNet is proposed by (He et al., 2016) and is evaluated through hyperparameter tuning, since every problem leads to different implementations of this model, specifically different number of filters. This model is trained and evaluated with several number of filters, specifically with 2, 4 and 6 filters for the first convolutional layer, with an increase of a multiplication of 2 per convolutional block.

5.3.3 Loss function

The loss function used for all models is binary cross-entropy with activation function Sigmoid. Hence the models were trained with 250 epochs and a patience of 50 epochs monitoring validation loss. Individual learning rates are automatically computed for different parameters using the Adam optimizer for stochastic optimization.

Outputs generated by the neural network indicate the probability of belonging to one label. The image is passed through the convolutional layers and is multiplied with the kernel filters, i.e. forward propagation. Backpropagation in contrast, explained by Werbos (1990), is an algorithm where the loss function's partial derivatives are calculated for each trainable weight of the neural network. These partial derivatives iteratively adjust the trainable weights to result in a lower loss model (Ho & Wookey, 2019).

5.4 Experiments and image pre-processing

The original CT scans included were scans with dimension 512 x 512 x the most slices for that date. Figure 8 displays an example of a 512x512 slice.

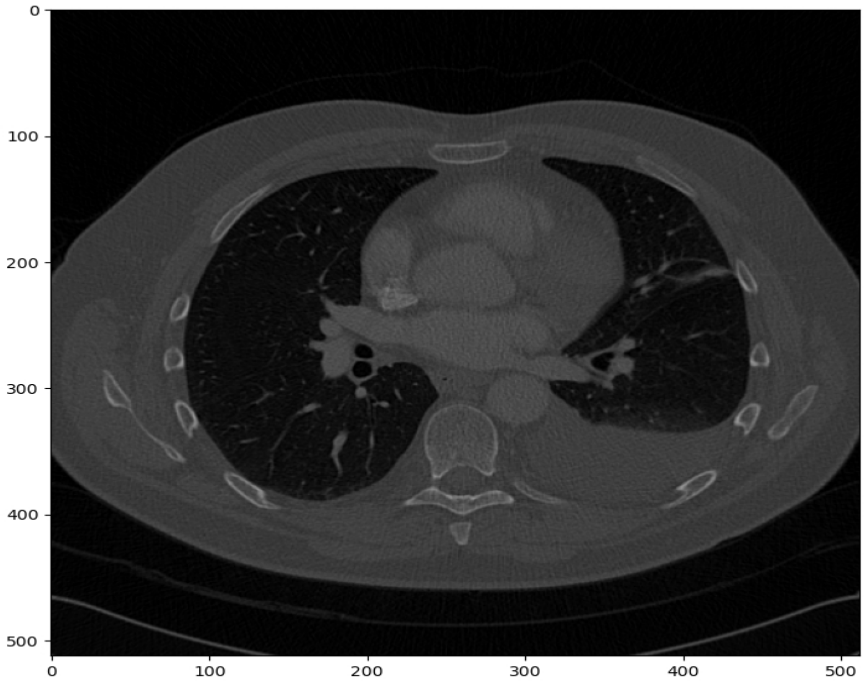


Figure 8: Example of a 512x512 slice (anonymized data)

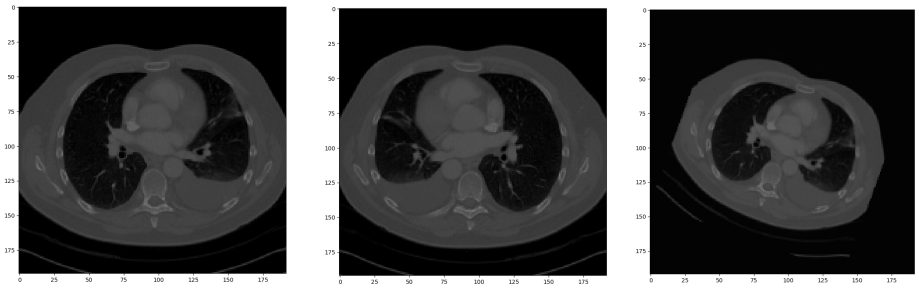
All CT scans consisted of a Hounsfield scale, which is a quantitative scale used for describing radiodensity. First, an Hounsfield Unit treshold window of [-1024, 3072] was set to normalize all the values between 0 and 1. Thereafter the CT scan was cropped to ignore all the irrelevant data. Due to memory limitations the CT scans were resized to 192x192x96. Although not every CT scan had a height and width of 192 and a depth of 96, this requirement was met by using padding, i.e. added spaces/slices which contained only zero values.

The following three experiments were evaluated with the previously introduced state-of-the-art models.

5.4.1 Experiment 1: Whole CT scan

Figure 9a displays the cropped and resized 192x192 slice of the example showed in figure 8. Once the pre-processing was done, the models were trained using generators to reduce the amount of data which had to be loaded at once. Due

to memory limitations, these generators were trained with a maximum batch size of 8 (number of training examples used in one iteration). Only the training set was trained using data augmentation operations such as flipping the image, i.e. figure 9b and rotating it randomly in a range of [-20°, 20°], i.e. figure 9c.



(a) Resized and cropped 192x192 slice

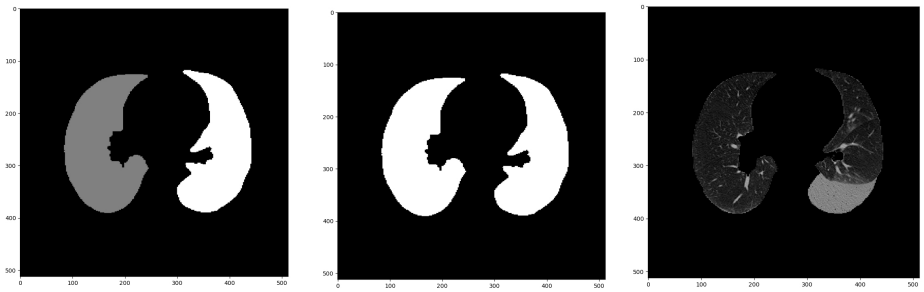
(b) Data augmentation: flipping whole CT

(c) Data augmentation: rotation whole CT

Figure 9: Pre-processing and data augmentation whole CT scan

5.4.2 Experiment 2: Lungs including abnormalities

This experiment mainly focused on automatically segmenting the lungs including the abnormalities to train the two state-of-the-art models with. U-net(R231) proposed by Johannes et al. (2020) was used to segment the lungmask for each CT scan. Once the lungmask was segmented, as seen in figure 10a, the two lungs had different values, 0 and 1 for the right and left lung, respectively. In order to obtain the original lungs including the abnormalities, the two lungs had to be the same value, i.e. see figure 10b. After the lungs were brought to one value, the lungmask was multiplied by the original CT scan to obtain the original lungs including the abnormalities, i.e. see figure 10c.



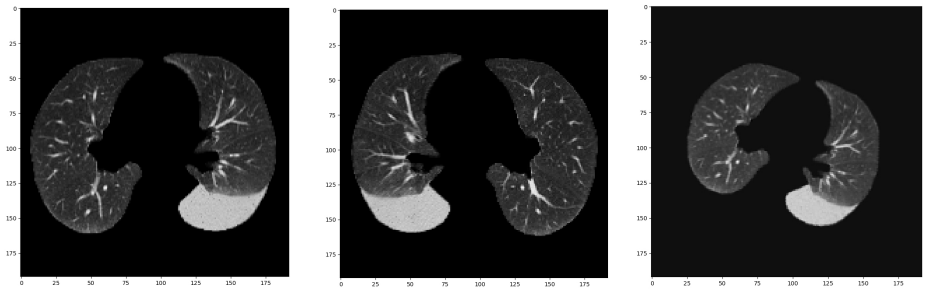
(a) Lungmask of lungs figure 8

(b) Lungmask same value lungs inclusive abnormalities

(c) Lungmask multiplied by the original ct

Figure 10: Pre-processing lungs

The same procedure as in experiment 1 is applied to crop and resize the lungs including the abnormalities (figure 11a) and thereafter to train the models using the same data augmentation operations as previously introduced, i.e. see figures 11b and 11c.



(a) Cropped and resized lungs

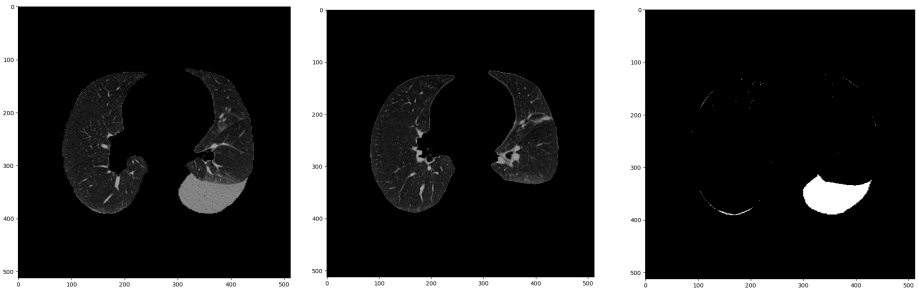
(b) Data augmentation: flipping lungs

(c) Data augmentation: rotation lungs

Figure 11: Pre-processing and data augmentation of the lungs

5.4.3 Experiment 3: Abnormalities

This part of the study focused primarily on calculating the difference between the lungs including the abnormalities and the healthy lungs (without the abnormalities). This was done by subtracting the healthy lungs from the lungmask which is previously introduced. To segment the healthy parts only, SegCaps proposed by LaLonde & Bagci (2018) was used. Once the healthy parts were segmented, i.e. see figure 12b, the healthy parts were subtracted from the lungmask which contained the healthy parts (lungs) as well as the abnormalities. This led to a segmentation of the abnormalities only, as shown in figure 12c.



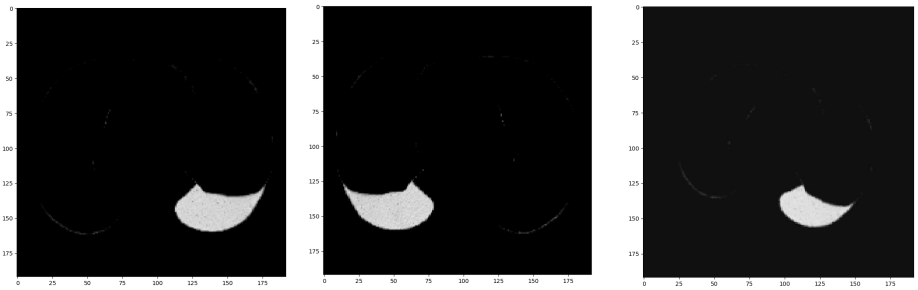
(a) Lungmask multiplied by the original ct

(b) Healthy lungs obtained through SegCaps

(c) Abnormalities

Figure 12: Pre-processing abnormalities

The same procedure as in experiments 1 and 2 is applied to crop and resize the abnormalities (figure 13a) and thereafter to train the models using the same data augmentation operations, see figures 13b and 13c.



(a) Cropped and resized abnormalities

(b) Data augmentation: flipping abnormalities

(c) Data augmentation: rotation abnormalities

Figure 13: Pre-processing and data augmentation abnormalities

5.5 Experiments evaluation and used hard/software

All models were evaluated using the same pre-processing and data augmentation operations to be able to evaluate their performance fairly. Evaluation was based on the validation loss, validation accuracy, Receiver Operating Characteristic (ROC), Area Under Curve (AUC) of the Receiver Operating Characteristic and a p-value (statistically significant if $p < 0.05$), sensitivity and specificity, all evaluated using the test set except the first two. Hence, the formulas of sensitivity and specificity are given in equation 1.

$$\textbf{Sensitivity (TPR)}: \frac{TP}{TP+FN} \qquad \textbf{Specificity (FPR)}: \frac{TN}{TN+FP} \tag{1}$$

where:
TP: true positive TN: true negative FP: false positive FN: false negative

Making use of the GPU Quadro RTX 8000 with 48 GB memory provided by the Netherlands Cancer Institute, all models were trained and evaluated.

5.6 Post-processing

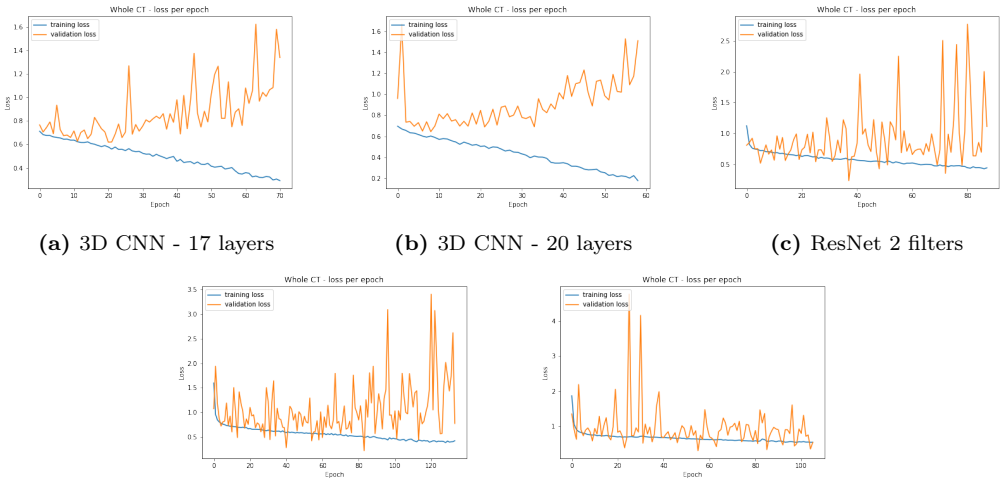
Investigation whether the volume of the abnormalities were generally decreasing, the total volume of the segmented abnormalities (figure 12c) were multiplied with the space directions of the CT scans. Since CT scans had varying slice thicknesses, this multiplication led to a total volume per CT scan which could be evaluated. Examination of the generated lungmask, i.e. figure 10a and SegCaps, i.e. figure 12b, has shown that various CT scans were not segmented. These lungmasks and SegCaps contained NaN values, which resulted in NaN losses during training. Hence, these CT scans were excluded.

6 Results

The difference in days between the death date minus scan date ($M = 655$, $SD = 772$) and the volume of abnormalities in mm3 ($M = 1018478$, $SD = 879999$) varied for the total 2347 CT scans used in this study.

6.1 Experiment 1: Whole CT scan

Training and validation losses on the whole CT scans are shown in figure 14. The ResNets with various filters have shown better performances regarding the validation losses, see figure 14c, 14d and 14e. The 3D CNNs indicate overfitting, while the ResNets indicate that it was learning from the data. However, a validation loss less than 0.4 is not achieved.



(a) 3D CNN - 17 layers **(b)** 3D CNN - 20 layers **(c)** ResNet 2 filters

(d) ResNet 4 filters **(e)** ResNet 6 filters

Figure 14: Whole CT - Loss per epoch

As shown in table 1 the highest achieved validation accuracy was 0.674 with the ResNet consisting of 6 filters. In addition, the AUC and the p-value are also shown, which indicates whether almost all cases are predicted under one class if it is not significant ($AUC = 0.513$, $p = 0.332$), followed by a sensitivity of 0.01 and a specificity of 0.99.

Networks	Val. acc.	Sensitivity	Specificity	AUC	p-value
3D CNN - 17 layers	0.663	0.42	0.61	0.496	0.453
3D CNN - 20 layers	0.654	0.40	0.59	0.522	0.229
ResNet 2 filters	0.657	0.18	0.78	0.473	0.186
ResNet 4 filters	0.663	0.42	0.59	0.520	0.258
ResNet 6 filters	0.674	0.01	0.99	0.513	0.332

Table 1: Performance of different networks on whole CT scans

Furthermore the ROC curves of the various networks are shown in figure 15, which contains the FPRs (False Positive Rate) on the x-axis and the TPRs (True Positive Rate) on the y axis.

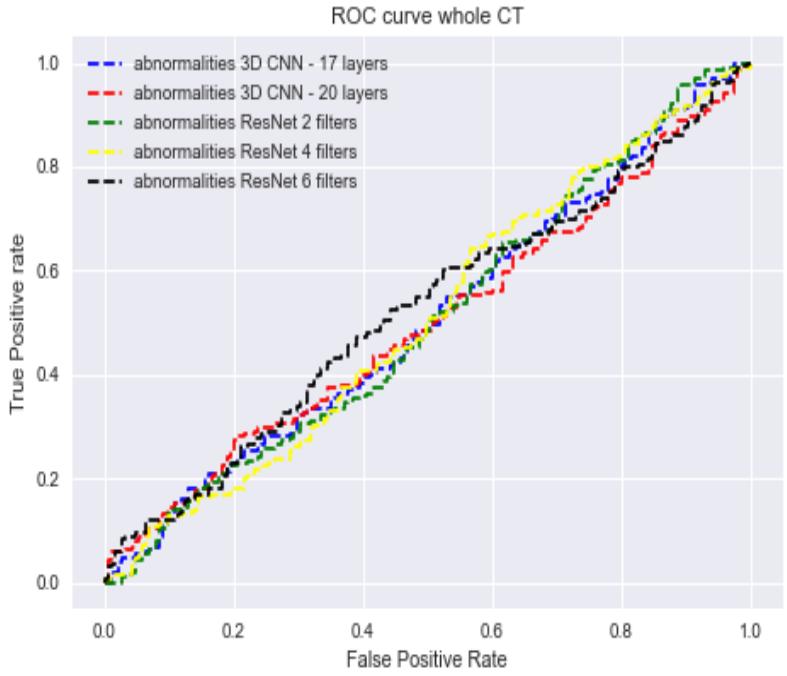
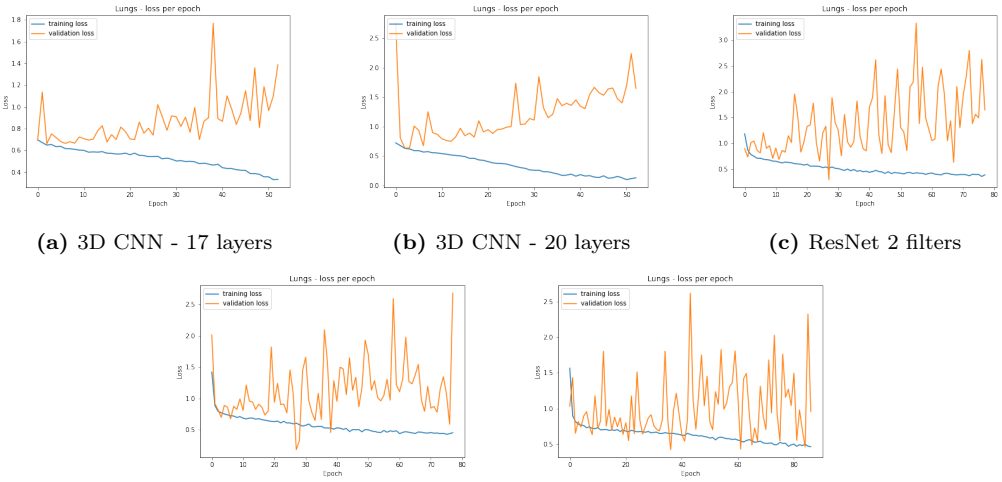


Figure 15: ROC curves whole CT

While random classifiers usually have a straight diagonal line, as shown in figure 15, no model has statistically outperformed a random classifier.

6.2 Experiment 2: Lungs including abnormalities

For this experiment the training and validation losses are shown in figure 16. The two 3D CNNs are again indicating overfitting, while the three other ResNets have learned from the data, i.e. see figures 16c, 16d and 16e.



(a) 3D CNN - 17 layers **(b)** 3D CNN - 20 layers **(c)** ResNet 2 filters

(d) ResNet 4 filters **(e)** ResNet 6 filters

Figure 16: Lungs - Loss per epoch

The highest achieved validation accuracy is 0.674 with the 3D CNN consisting of 20 layers, as shown in table 2. However, the AUC is less than the AUC of the ResNet with 6 filters. The p-value is also not significant, while the p-value of the ResNet with 6 filters is ($AUC = 0.549$, $p = 0.047$). The sensitivity and specificity of the ResNet with 6 filters are 0.16 and 0.87.

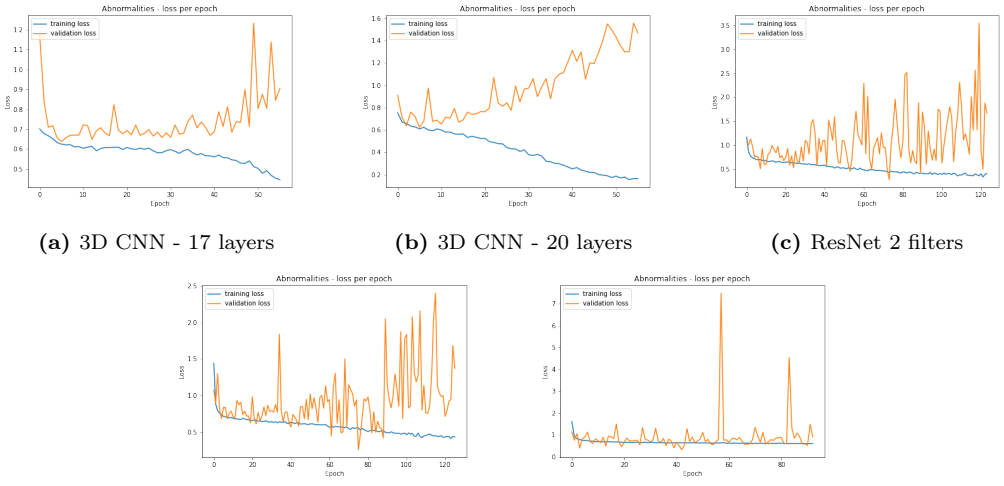
Networks	Val. acc.	Sensitivity	Specificity	AUC	p-value
3D CNN - 17 layers	0.651	0.07	0.90	0.494	0.420
3D CNN - 20 layers	0.674	0.06	0.96	0.508	0.402
ResNet 2 filters	0.651	0.40	0.57	0.478	0.219
ResNet 4 filters	0.642	0.13	0.90	0.541	0.078
ResNet 6 filters	0.640	0.16	0.87	0.549	0.047

Table 2: Performance of different networks on the segmentation of the lungs

This experiment has shown that only the ResNet with 6 filters performed better than a random classifier with a statistically significant p-value. However, it is still very close to a random classifier.

6.3 Experiment 3: Abnormalities

The training and validation losses of the abnormalities are shown in figure 17. The 3D CNNs are indicating overfitting again, while the other three ResNets indicate that learning from the data has been achieved.



(a) 3D CNN - 17 layers **(b)** 3D CNN - 20 layers **(c)** ResNet 2 filters

(d) ResNet 4 filters **(e)** ResNet 6 filters

Figure 17: Abnormalities - Loss per epoch

For this experiment the highest achieved validation accuracy is 0.686 with the 3D CNN consisting of 17 layers, as shown in table 3. However, the AUC is less than the AUC of the ResNet with 6 filters. The p-value is also higher than the p-value of the ResNet with 6 filters ($AUC = 0.524$, $p = 0.212$). The sensitivity and specificity of the ResNet with 6 filters are 0.12 and 0.89.

Networks	Val. acc.	Sensitivity	Specificity	AUC	p-value
3D CNN - 17 layers	0.686	0.35	0.64	0.506	0.419
3D CNN - 20 layers	0.669	0.38	0.62	0.500	0.495
ResNet 2 filters	0.648	0.43	0.54	0.504	0.437
ResNet 4 filters	0.668	0.18	0.79	0.510	0.371
ResNet 6 filters	0.674	0.12	0.89	0.524	0.212

Table 3: Performance of different networks on the segmentation of the abnormalities

No model has shown better performance than a random classifier, since all the AUCs are fluctuating around the 0.51 and no p-value is significant.

6.4 Post-processed experiment

The volumes in mm3 of the patients varied in the whole dataset. As figure 18 shows, the volume of the abnormalities of a specific anonymized patient could first decrease and thereafter increase again.

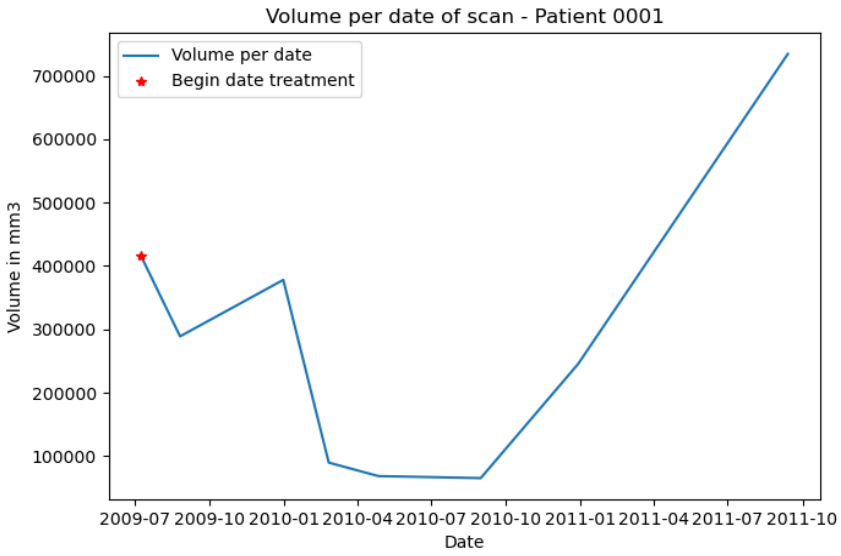


Figure 18: Volume per date - anonymized patient 0001

However, patients with less than three scans were excluded to calculate the correlation between the difference in days between death date minus scan date and the volume of abnormalities in mm3, since patients with two CT scans led to a correlation of 1 and patients with one CT scan led to a correlation of NaN .

A weak negative correlation (r = -0.199, $p < 0.001$) was found between the difference in days between death date minus scan date and the volume of abnormalities in mm3.

7 Discussion

This study analysed whether CT scans had predictive values for survival. Three experiments have been conducted. The whole CT scan, the segmented lungs from the whole CT scan and the segmented abnormalities from the segmented lungs were pre-processed and trained using data augmentation operations, all in the same way. From there the total volume per CT scan was calculated. The main research question was: To what extent can the survival rate of a patient diagnosed with Malignant Pleural Mesothelioma be predicted on the basis of CT scans?

7.1 Interpretation results

First, the experiment with whole CT scans will be discussed, which showed better validation loss graphs for the ResNets with 2, 4 and 6 filters, than the two 3D CNNs. However, low AUC scores were achieved. This could be due to irrelevant information available in the whole CT scan for MPM, since the regions of interest were the abnormalities. The CT scans had to be resized due to memory limitations to fed the data generator with, which could possibly have caused loss of details.

The second point of discussion is the experiment with the lungs. For this experiment, the 3D CNNs were outperformed by the ResNets with 2, 4 and 6 filters based on the graphs of validation losses. While the 3D CNN with 20 layers has had a higher validation accuracy, the ResNet with 6 filters performed better on the sensitivity, specificity, AUC and p-value, indicating not all cases were predicted under one class. However, the performance is still very close to a random classifier. One possible reason for this is resizing, since all models were pre-processed and trained with the same procedures.

Third, the experiment with only the abnormalities will be discussed, which showed comparable performances on the validation losses as the other two experiments. While the validation accuracy of all models were comparable with each other, no model statistically outperformed a random classifier. This could be due to the segmentation of the lungmasks and healthy lungs including the abnormalities. Some cases showed that no precise segmentations were made (i.e. segmenting fluid), which led to segmentations of more volume for the abnormalities. Resizing could also possibly have caused loss of details.

Fourth, the correlation between the difference in days between death date minus scan date and the volume of abnormalities in mm3 will be discussed, which showed a weak negative correlation. This could be due to the inaccurate segmentations, which led to larger total volume of the abnormalities. One other possible reason could be that the patient's treatment-response was positively, which could cause smaller volumes of the abnormalities.

7.2 Limitations

This study has had several limitations. First, the end date of the patient (date of death) was not exact. The end date consisted of the last date the radiologists were told the patient had a CT scan. Thus, after consideration, in this study the end date served as the death date of the patient. Hence, the soft labels are also not precise.

Second, the third experiment consisted of the volume of the abnormalities instead of the tumor. Due to absence of expert segmentations of MPM, the model was limited to all abnormalities. The difference between the lungmask and the SegCaps model was calculated, which had some limitations regarding the exact volume of the abnormalities. Since the segmentations of the two models were not exactly the same (regarding the healthy lungs and the healthy lungs including the abnormalities), the difference consisted of the abnormalities, but also for example of pixels around the edges of the lungs, which were not per definition abnormalities.

Third, the hypothese was that abnormalities would perform better than the other two experiments, but in some cases due to inaccurate segmentations, volumes of abnormalities were much bigger than the volumes of previously made CT scans. In general, it is expected that the volume will descend in the follow-up CT-scans if the therapy is successful.

Fourth, due to memory limitations, batch sizes equal to 8 were used and resizing was done before loading the generator instead of resizing during generating.

Fifth, manually inspecting the pre-processed and segmented CT scans was not done, due to time limitations. Inspecting all the pre-processed and segmented CT scans, could give an insight if mistakes were made. Mistakes could be that not every experiment, had for example comparable normalized values and did not contain any outliers in for example volume of segmentations.

Sixth, due to the time it took to train each model, i.e. approximately 18 hours, the models could not be trained with a greater number of epochs and for the ResNets it was not feasible to train with more filters.

Finally, inclusion of all the patient's CT scans was not feasible, since the last part of the CT scans were added after all the pre-processing. Due to time limitations and computational power limitations to add and pre-process these CT scans within a short time, these CT scans were excluded.

7.3 Future perspectives

Future work should be focused on, manually inspecting the pre-processed and segmented data. Hence, it could prevent the model from performing poorly.

Also, larger batch sizes with more epochs could potentially improve the performance of the models (Narin & Pamuk, 2020). As proposed by (Van Gerwen et al., 2019), clinical factors could be predictive for longer survival, thus inclusion of these factors could potentially improve the performance of the model. Furthermore, as stated by (Xu et al., 2019), to capture changes in tumor volume, follow-up time points are key to predict survival of a patient. One possible method to obtain better performance is to combine a CNN with a RNN, since several time points could be combined. The model would still learn, even if a patient's CT scan was missed at a certain time point, which is inevitable in studies like this one (Xu et al., 2019). The change in CT scans over time, could possibly contain more predictive values than only a CT scan.

8 Conclusion

CT scans of patients with Malignant Pleural Mesothelioma were pre-processed, trained and evaluated on two state-of-the-art models with three different experiments. All experiments showed that no model statistically outperformed a random classifier, except one which was slightly better. Patient's volumes in mm3 varied in the whole dataset. Hence, a weak negative correlation was found between the difference in days between death date minus scan date and the volume of abnormalities in mm3.

Based on the experiments conducted in this study, can be determined that CT scans of patients with Malignant Pleural Mesothelioma, do not show any predictive values for survival with the conducted methods.

References

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., & Zoph, B. (2021). Revisiting resnets: Improved training and scaling strategies. *arXiv preprint arXiv:2103.07579*.

Bibby, A. C., Tsim, S., Kanellakis, N., Ball, H., Talbot, D. C., Blyth, K. G., Maskell, N. A., & Psallidas, I. (2016). Malignant pleural mesothelioma: an update on investigation, diagnosis and treatment. *European Respiratory Review*, 25(142):472–486.

Cooper, G. M. & Hausman, R. E. (2000). The development and causes of cancer. *The cell: A molecular approach*, 2.

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2020). Global cancer observatory: cancer today. lyon: International agency for research on cancer; 2018.

Frauenfelder, T., Tutic, M., Weder, W., Götti, R., Stahel, R., Seifert, B., & Opitz, I. (2011). Volumetry: an alternative to assess therapy response for malignant pleural mesothelioma? *European Respiratory Journal*, 38(1):162–168.

Gill, R. R., Richards, W. G., Yeap, B. Y., Matsuoka, S., Wolf, A. S., Gerbaudo, V. H., Bueno, R., Sugarbaker, D. J., & Hatabu, H. (2012). Epithelial malignant pleural mesothelioma after extrapleural pneumonectomy: stratification of survival with ct-derived tumor volume. *American Journal of Roentgenology*, 198(2):359–363.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. *CoRR*, abs/1603.05027.

Ho, Y. & Wookey, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8:4806–4813.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.

Johannes, H., Jeanny, P., Sebastian, R., Helmut, P., & Georg, L. (2020). Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1).

LaLonde, R. & Bagci, U. (2018). Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*.

Lencioni, R. & Llovet, J. M. (2010). Modified recist (mrecist) assessment for hepatocellular carcinoma. In *Seminars in liver disease*, volume 30, pages 052–060. © Thieme Medical Publishers.

Liu, F., Zhao, B., Krug, L. M., Ishill, N. M., Lim, R. C., Guo, P., Gorski, M., Flores, R., Moskowitz, C. S., Rusch, V. W., et al. (2010). Assessment of therapy responses and prediction of survival in malignant pleural mesothelioma through computer-aided volumetric measurement on computed tomography scans. *Journal of thoracic oncology*, 5(6):879–884.

Manning, C. B., Vallyathan, V., & Mossman, B. T. (2002). Diseases caused by asbestos: mechanisms of injury and disease development. *International immunopharmacology*, 2(2-3):191–200.

Murphy, D. J. & Gill, R. R. (2017). Volumetric assessment in malignant pleural mesothelioma. *Annals of translational medicine*, 5(11).

Narin, A. & Pamuk, Z. (2020). Effect of different batch size parameters on predicting of covid19 cases. *arXiv preprint arXiv:2012.05534*.

Opitz, I. & Weder, W. (2018). Pleural mesothelioma: is the surgeon still there? *Annals of Oncology*, 29(8):1710–1717.

Pass, H. I., Kranda, K., Temeck, B. K., Feuerstein, I., & Steinberg, S. M. (1997). Surgically debulked malignant pleural mesothelioma: results and prognostic factors. *Annals of surgical oncology*, 4(3):215–222.

Phung, V. H. & Rhee, E. J. (2018). A deep learning approach for classification of cloud image patches on small datasets. *Journal of information and communication convergence engineering*, 16(3):173–178.

Plathow, C., Klopp, M., Thieke, C., Herth, F., Thomas, A., Schmaehl, A., Zuna, I., & Kauczor, H.-U. (2008). Therapy response in malignant pleural mesothelioma-role of mri using recist, modified recist and volumetric approaches in comparison with ct. *European radiology*, 18(8):1635–1643.

Rusch, V. W., Gill, R., Mitchell, A., Naidich, D., Rice, D. C., Pass, H. I., Kindler, H. L., De Perrot, M., Friedberg, J., Ginsberg, M., et al. (2016). A multicenter study of volumetric computed tomography for staging malignant pleural mesothelioma. *The Annals of thoracic surgery*, 102(4):1059–1066.

Singh, S. A., Meitei, T. G., & Majumder, S. (2020). 6 - short pcg classification based on deep learning. In Agarwal, B., Balas, V. E., Jain, L. C., Poonia, R. C., & Manisha, editors, *Deep Learning Techniques for Biomedical and Health Informatics*, pages 141–164. Academic Press.

Sultana, F., Sufian, A., & Dutta, P. (2018). Advancements in image classification using convolutional neural network. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 122–129. IEEE.

Tsao, A. S., Garland, L., Redman, M., Kernstine, K., Gandara, D., & Marom, E. M. (2011). A practical guide of the southwest oncology group to measure malignant pleural mesothelioma tumors by recist and modified recist criteria. *Journal of Thoracic Oncology*, 6(3):598–601.

Van Gerwen, M., Alpert, N., Wolf, A., Ohri, N., Lewis, E., Rosenzweig, K. E., Flores, R., & Taioli, E. (2019). Prognostic factors of survival in patients with malignant pleural mesothelioma: an analysis of the national cancer database. *Carcinogenesis*, 40(4):529–536.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R. H., & Aerts, H. J. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, 25(11):3266–3275.

Zunair, H., Rahman, A., Mohammed, N., & Cohen, J. P. (2020). Uniformizing techniques to process ct scans with 3d cnns for tuberculosis prediction. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 156–168. Springer.