university of groningen

Exploring Vector Attributes of Max tree, Component tree using Supervised Machine Learning Technique

Under Supervision of :

Prof. Dr. Kerstin Bunte

Dr. Michael H.F. Wilkinson

Presented by:

Pooja Gowda

S4410963

CONTENTS

university of groningen

university of groningen

1 Introduction

Introduction

university of groningen

Lung cancer is the leading cause of cancer-related deaths in Europe as per European Commission.

Europe's Beating Cancer Plan is (1) **Prevention**; (2) Early **detection**; (3) **Diagnosis and treatment**; and (4) **Quality** of life of cancer patients and survivors

Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) scans are used to detect lung tumors since FDG-PET can provide information on active inflammatory lesions.

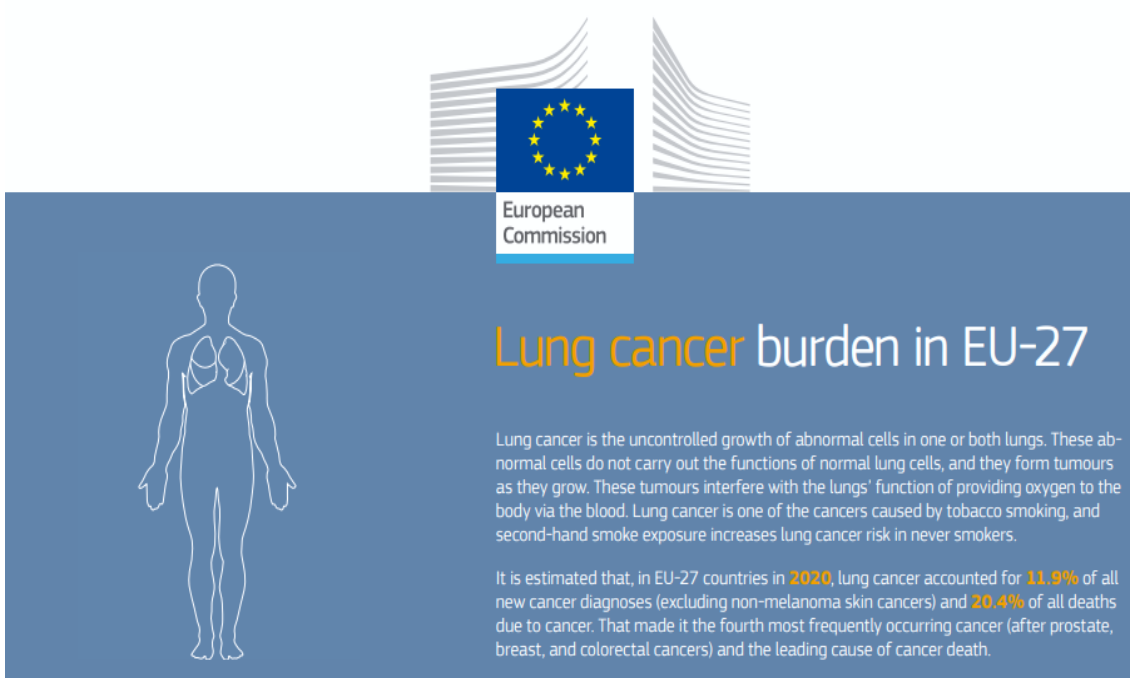


Fig. Lung Cancer Statistics provided by European Commission [1].

Introduction

university of groningen

Deep learning, machine learning and Image filtering, segmentation based solutions are existing for early detection of lung tumours.

H. Gan et al. presented **a hybrid technique** for detecting lung tumors, using Unsupervised Machine Learning Technique **Self Organization Maps** (SOMs) and **Image Segmentation filtering**.

FDG-PET scans were used in this work to explore vector attributes of these 3D-volume, where the **Max Tree was built for each FDG-PET Scan using DISSCOFAN (Distributed Component Filtering analysis)**

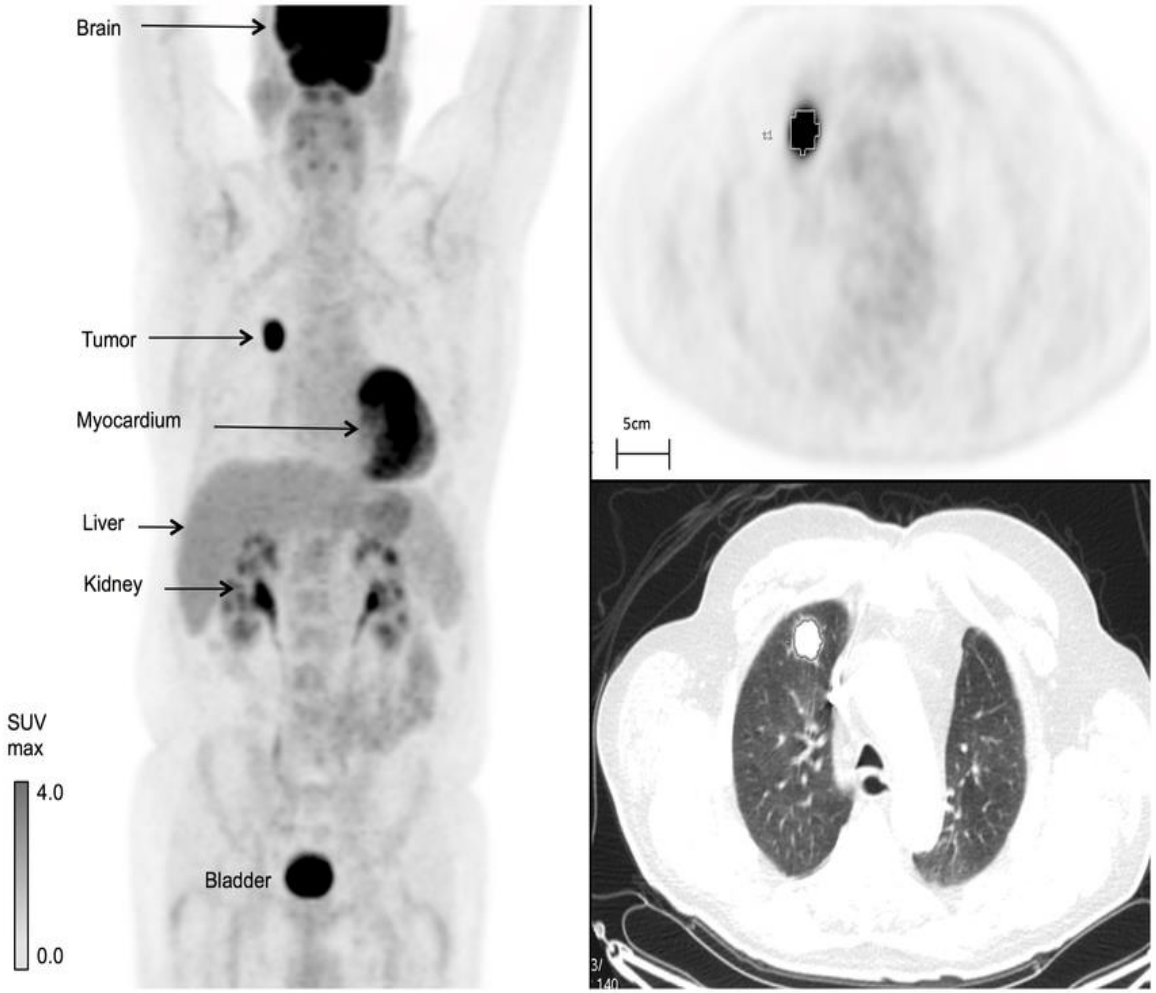


Fig. PET scan indicating the lung tumour [2].

Introduction

university of groningen

Connected filters, essential morphological mathematical tool - selection or elimination of connected components.

The **attribute filter**, which is a subset of connected filters that enables the filtering of images based on specific size and shape characteristics

Efficient on Multi scale representation of images, like Max tree , component Tree.

Non effective on noisy images (less discriminatory power), requires manually setting discriminatory threshold, select attributes manually to describe data.

Filtering an image using attribute filters [3].

Introduction – Thesis Objective

university of groningen

This thesis proposes a general-purpose data science tool using a Supervised Machine learning technique, Random Forest to analyse the significance of vector attributes to classify tumors effectively without manually thresholding the set of the vector attributes.

Introduction – Research Questions

university of groningen

Q1. - Can we apply the supervised machine learning techniques to explore vector attributes of hierarchical region-based data structures of an image to better characterize objects in an image in terms of size and shape?

Q2. - What are the relevant feature vectors from PET scan data that enables the supervised machine learning technique to classify tumors effectively?

Q3. - How does the threshold choice of high tumor fraction from the PET scan volumes affect the classifier's performance and the feature's relevance for the classification?

university of groningen

2 Methodology

MAX TREE

university of groningen

The connected filters are not applied to the individual pixels. Instead, it operates on the connected components where the image is constant.

The component tree data structure uses the hierarchical parent affinities between data nodes and organizes the connected components of different threshold levels of an image

One of the compact component tree representations is known as the Max tree. In Max trees, the leaves are the regional maxima. The max trees are built using (DISCCOFAN).

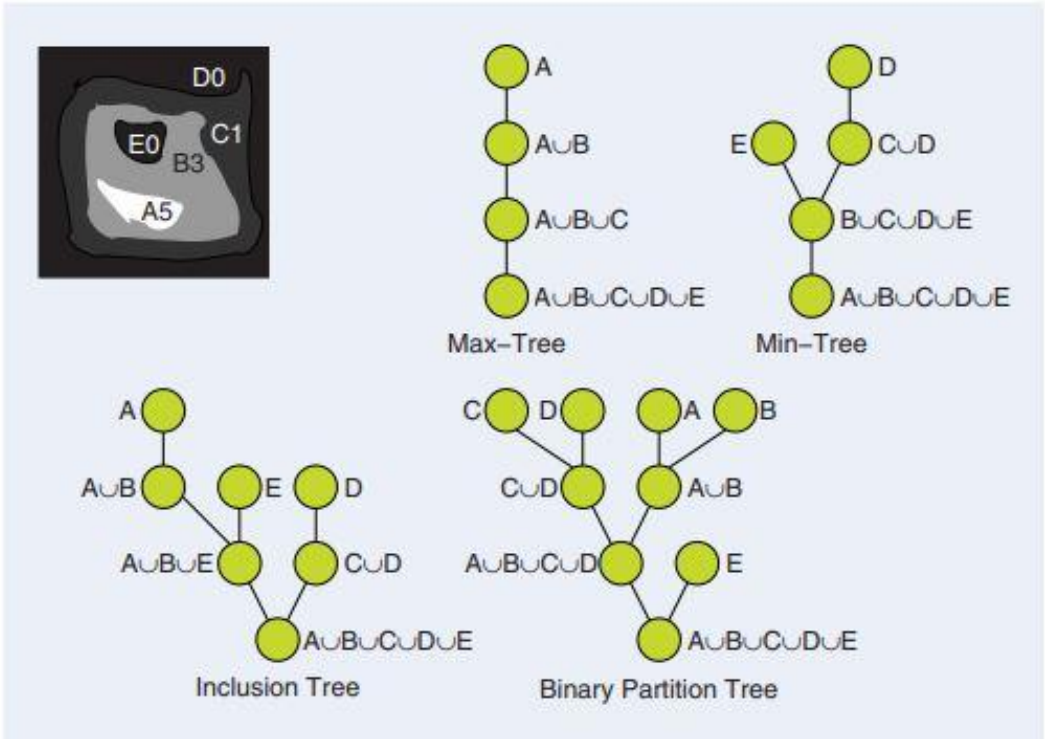


Fig. Tree representations(Max, Min, inclusion) built from the original images with flat nodes: A, B, C, D, E and the corresponding gray level values [4].

Vector-Attribute Filters

university of groningen

The attributes filters, filters from the images attributes like the image's shape and size properties.

We examine different sizes and shape structures in 3D PET scan medical images to improve the discriminating power between images.

X-extent, Y-extent, Z-extent – size based Elongation, flatness, Non-compactness – shape based attributes selected based on central image moments (Centroids). The intensity mean, variance and power are also calculated based on central image moments from the max tree data structure.

Self Organizing Maps (SOMs),

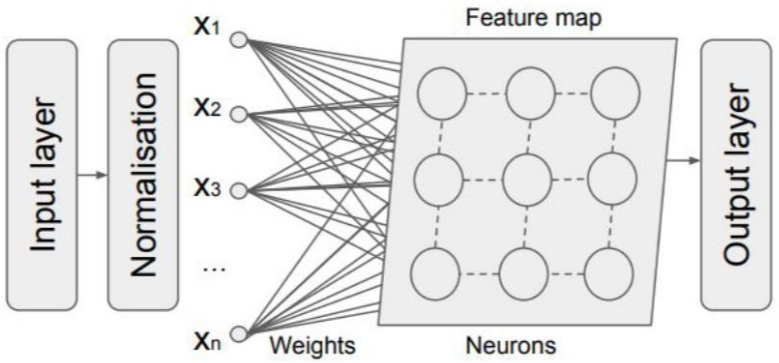
Unsupervised Machine Learning Technique

university of groningen

Algorithm 1 Self Organisation Map Algorithm

Input: input data vector X;
Initialisation : weight vectors W_k;
N ← Iteration count;

1. for i = 1 to N do
2. d_k ← the distance between the input and weight vector;
3. c ← The winning neuron(BMU), that is the neuron with the shortest distance from the input;
4. b_k ← The topological neighborhood of the winning neuron;
5. Update the weight vector of each neuron.
6. end for



Random Forest (Bagging)

Supervised Machine Learning Technique

university of groningen

Require: To generate c classifiers:

for i=1 to c **do**
  Randomly samples the training data D with replacement to produce D_i.
  Create a root node, N_i containing D_i.
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly selects x% of the possible splitting features in N.
  Select the feature F with the highest information gain to split on
  Create f child nodes N, $N_1, .., N_f$, where F has f possible values $(F_1, ...F_f)$

  for i=1 to f do **do**
    Set the contents of N_i to D_i, where D_i is all instances in N that match F_i
    Call BuildTree(N_i)
  end for
end if

3D volume visualization Tool: MTdemo

university of groningen

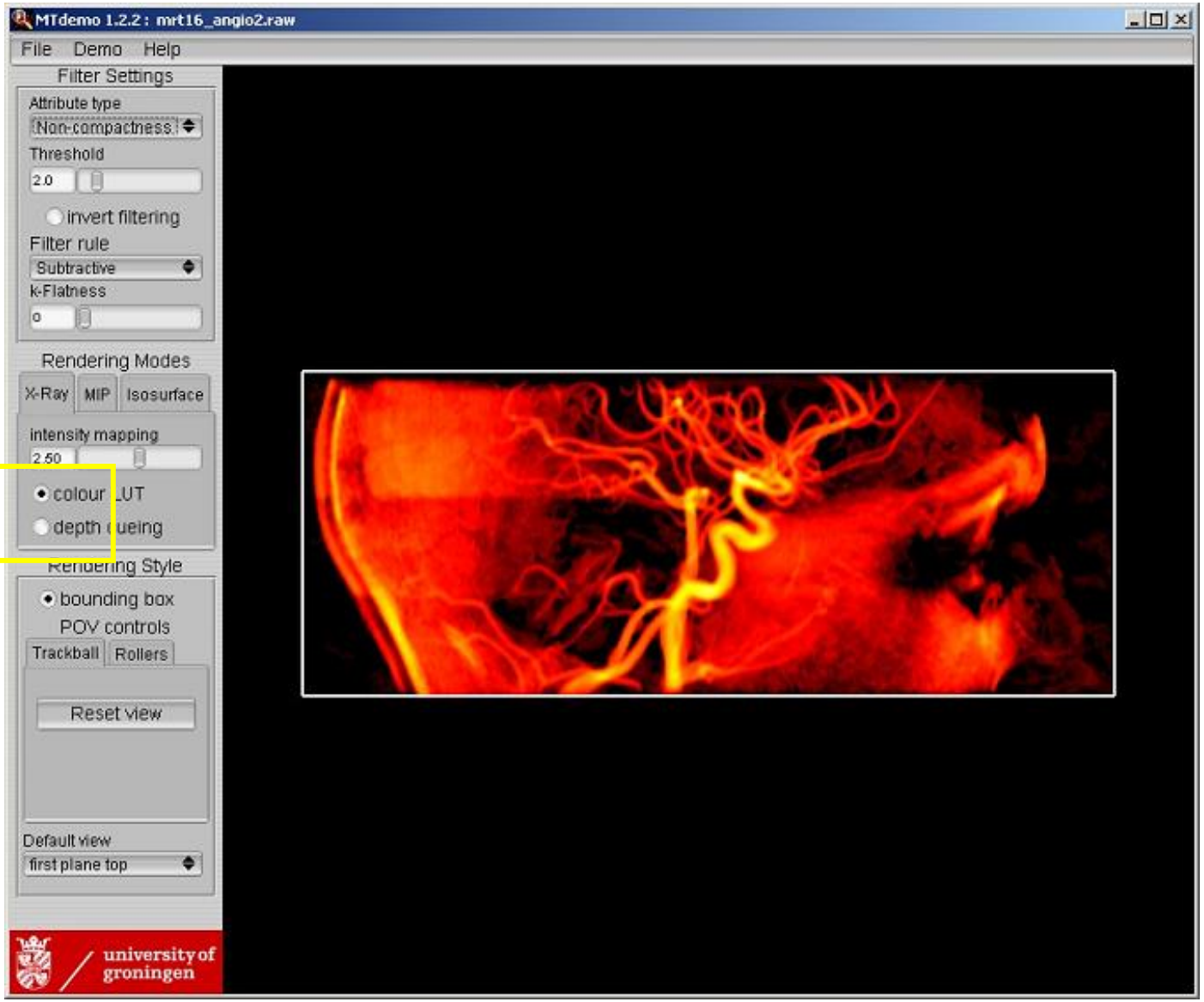


Fig. X-ray visualization mode using MTdemo visualization tool [5].

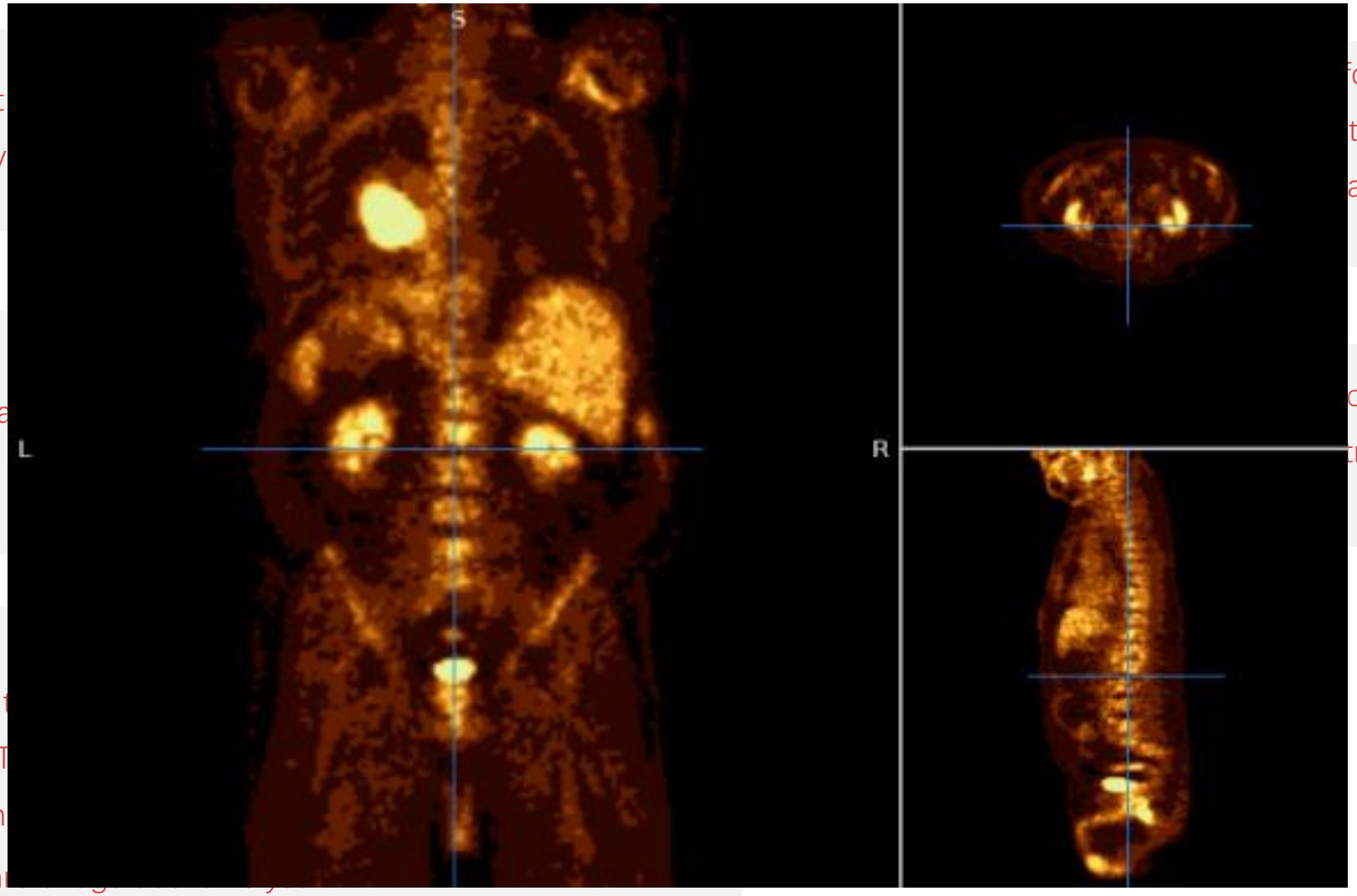
university of groningen

3 Implementation

Data Set

university of groningen

The dataset
tomography



for each FDG-PET scan using
t filtering and analysis
are computed from the Max

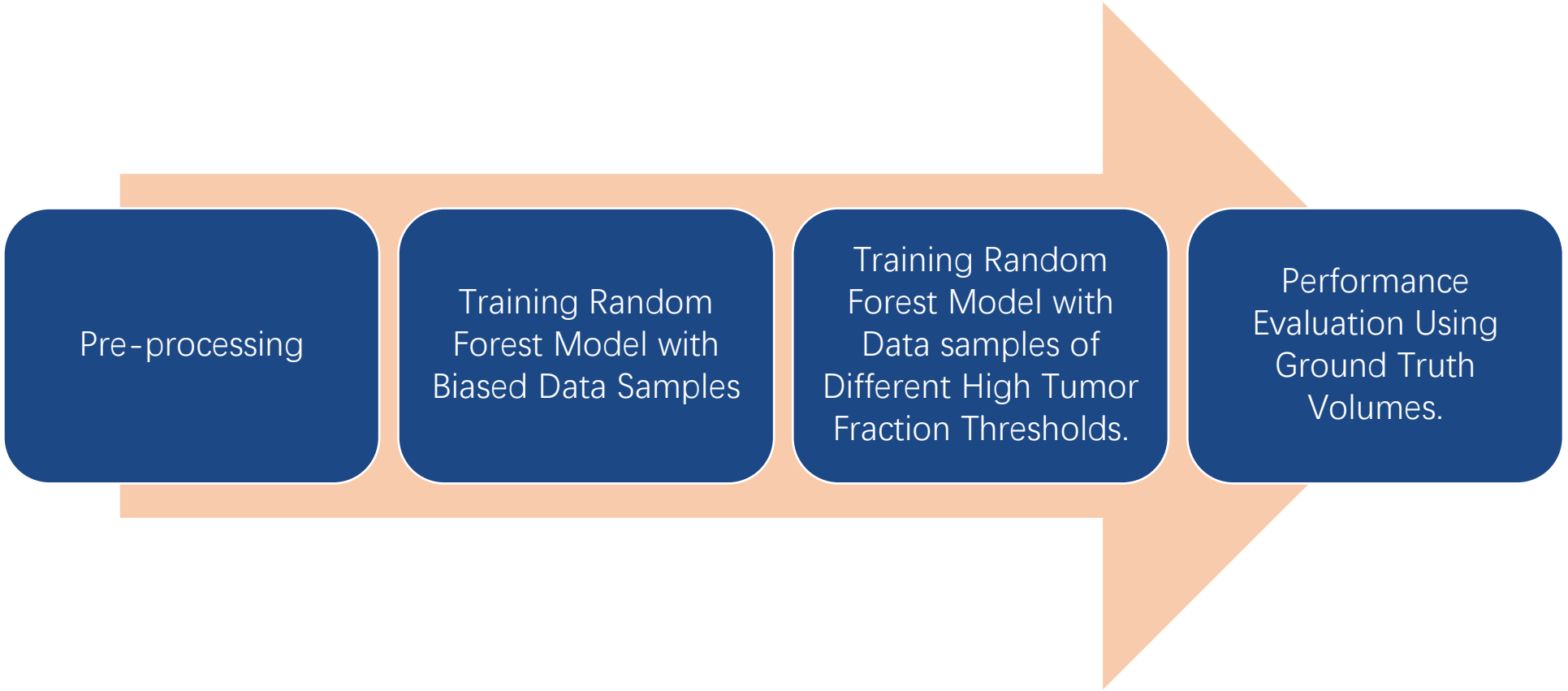
The analysis wa
ground truths (

d and distributed techniques
trees, user-defined attribute

A multi-center
this thesis by
ACRIN) and RT
FDG-PET which
patient's data a

Experimentations Performed

university of groningen



Pre-processing

university of groningen

In the PET scan volumes, it is observed that the lung tumor nodes occupied a significantly smaller portion of the PET scans (< 5%).

The Nodes with High tumor Fraction Threshold (a node with a tumor volume ⩾ to 90% of the total volume) are considered. Such nodes are labelled as ‘High’ and rest nodes below the threshold as ‘Low’

We split the 202 Data samples – 180 datasets as Training Data and 22 volume datasets are kept aside as unseen data or Test set to evaluate the model’s generalizing ability.

Training Random Forest Model with biased data samples

university of groningen

High tumor nodes account only for ~ 0.21% of the total input nodes from the 180 dataset volume. Hence, four different samples are used for training random forest using majority under sampling technique.

Filter out the least important features and retrain the model with rest of the features.

We apply 10-fold cross-validation to notice overfitting in the model.

Test the trained model with different data samples to evaluate the generalization on 22 rest of the data sets as a whole.

We use random forest algorithm with “bagging” as the method with 100 Decision Trees.

		ur nodes	High tumour percentage
1	5000000	10.285	0.2%
		55	
		55	
		55	

Table 3.1: Four different samples were used for training Random Forest Model.

Training Random Forest Model with Data samples of Different High Tumor Fraction Thresholds.

university of groningen

Change the High tumor fraction Thresholds, nodes with tumo ain the volum

We a ples to the m ets as a

We use random forest algorithm with "bagging" as the method with 100 Decision Trees.

Sample	Total nodes	High tumour nodes	Tumour Fraction Threshold	High tumour percentage
1	866862	433431	$\geq 10\%$	50.0%
2	670512	335256	$\geq 30\%$	50.0%
3	540792	270396	$\geq 50\%$	50.0%
4	435136	217568	$\geq 70\%$	50.0%
5	396816	198408	$\geq 80\%$	50.0%
6	356710	178355	$\geq 90\%$	50.0%

Table 3.2: Six different samples labeled with different Tumor Fraction Threshold that were used for training the Random Forest Model

Performance Evaluation Using Ground Truth Volumes

university of groningen

Validate the performance by comparing classified volumes from the test set and the training set concerning two ground truth volumes (Lung tumor images from PET scans provided by the two experts).

These computations are performed with classified volume from every classifier that is trained on datasets with 180 datasets concerning different thresholds.

Calculate True Positive, True Negative, False Positive, False Negative, F-score, Precision, Recall, and Accuracy.

First filter out the volumes and assign tumor labels(0 or non-zero) according to the classification results to each volume.

Finally, The average F-score and classification accuracy on the test and training set is computed.

Compare the volumes with ground truth.

university of groningen

4 Results & Discussion

Training Random Forest Model with biased data samples

university of groningen

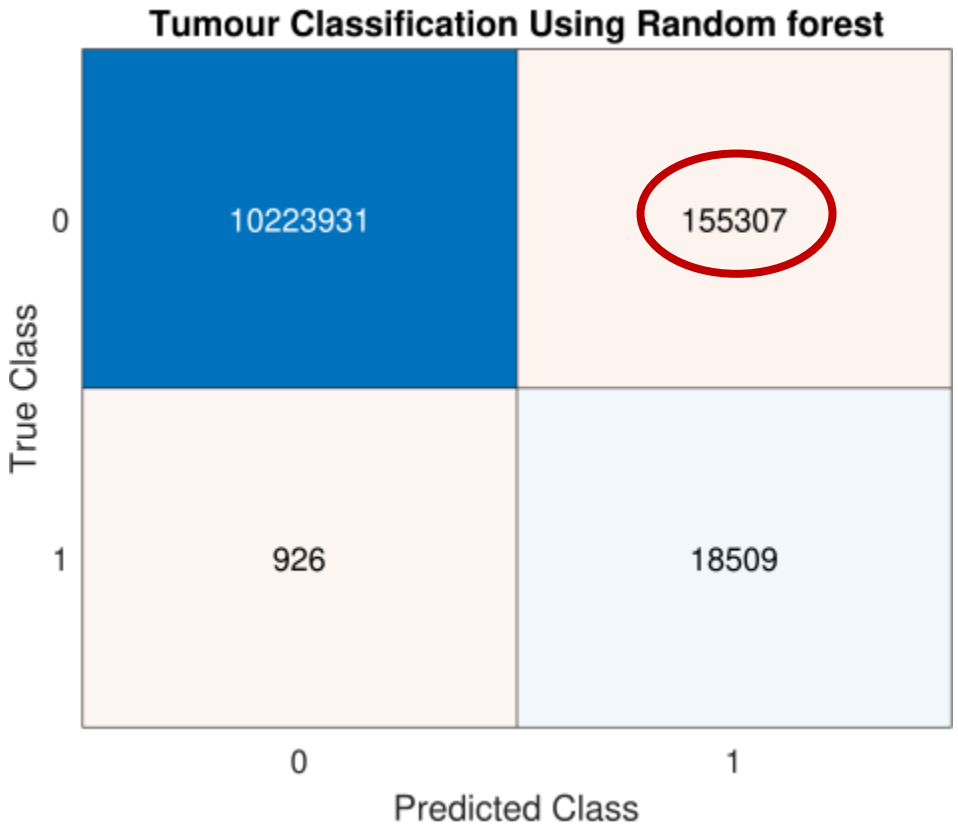


Fig. Classifier's results on the test set, trained with data sample 1 (0.2% hight tumor percentage)

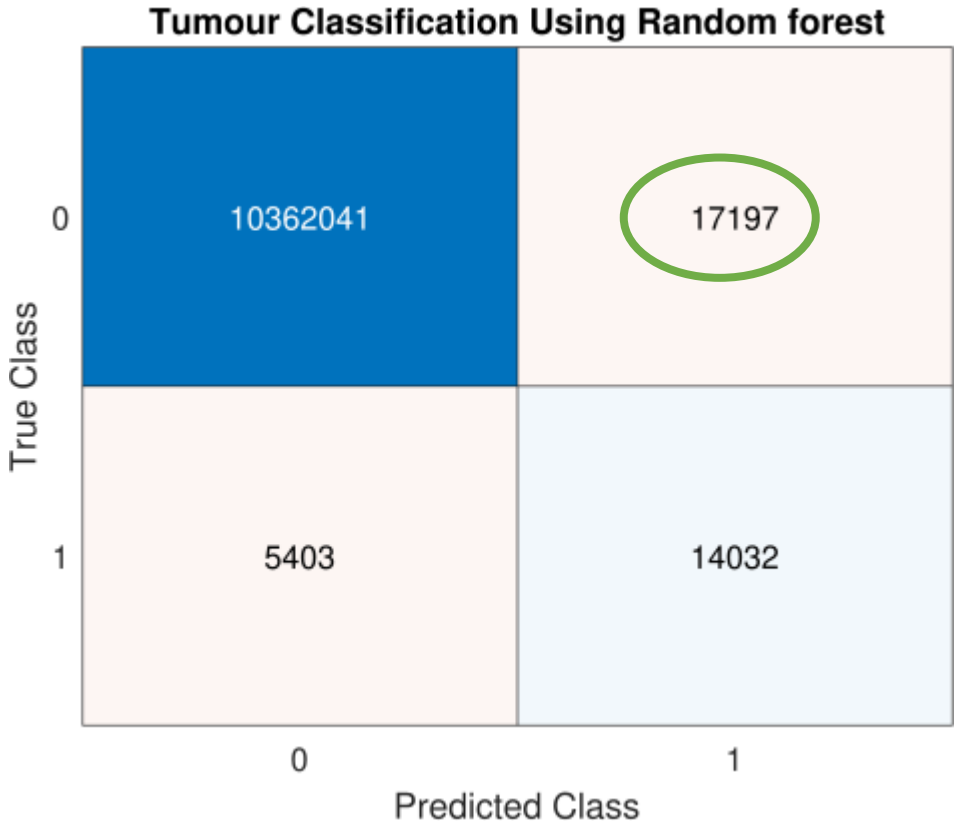
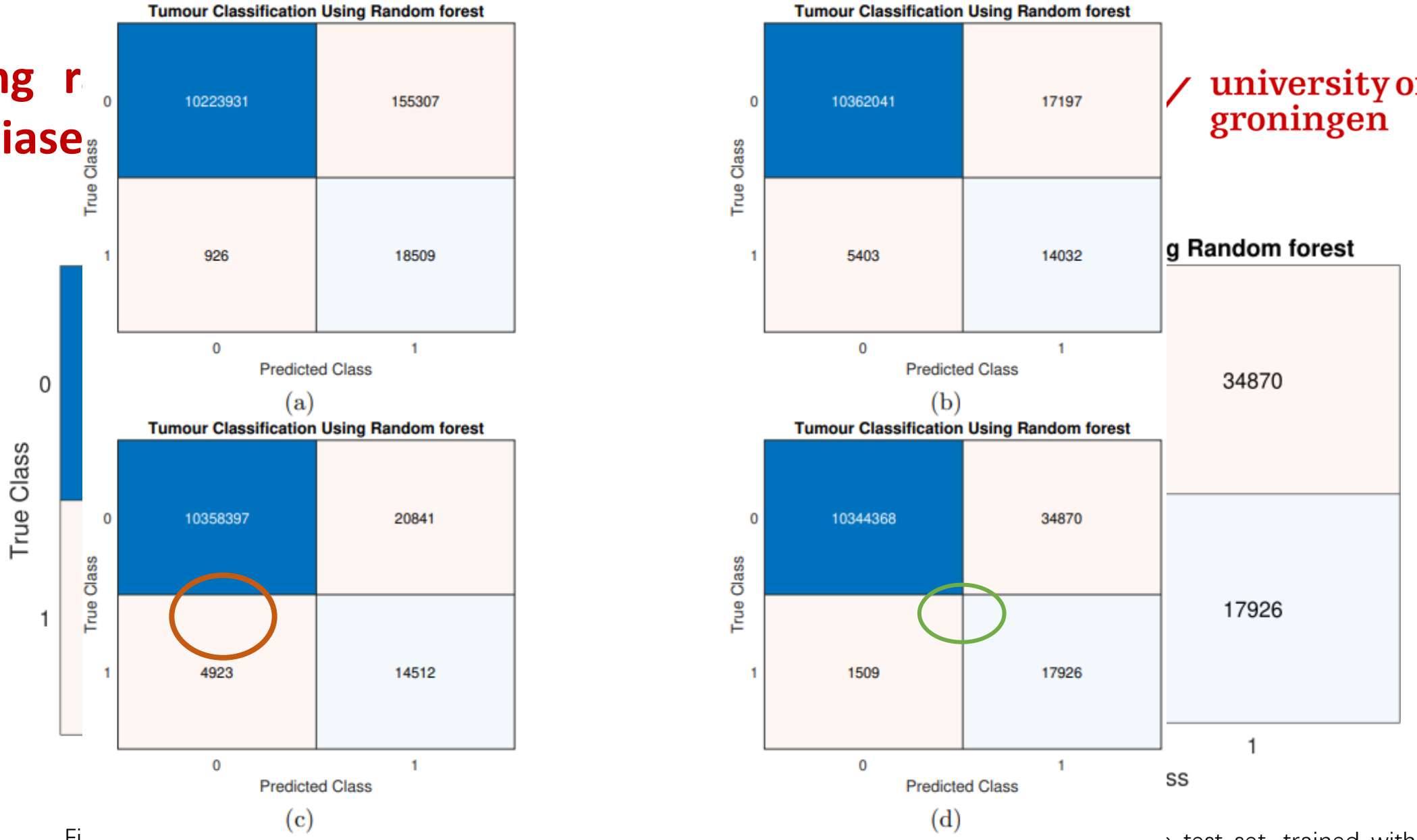


Fig. Classifier results on the test set, trained with data sample 2 (10% hight tumor percentage)

Training r
with biase

university of groningen



Fi e test set, trained with
da nor percentage)

Figure 4.1: The Sub Figures 4.1a, 4.1b, 4.1c, 4.1d are the classification results on the test set, the classifier trained on biased data samples 1 to 4 with tumor volume accounting $\geq 90\%$ of total volume respectively.

Training random forest model with biased data samples – Feature's relevance

university of groningen

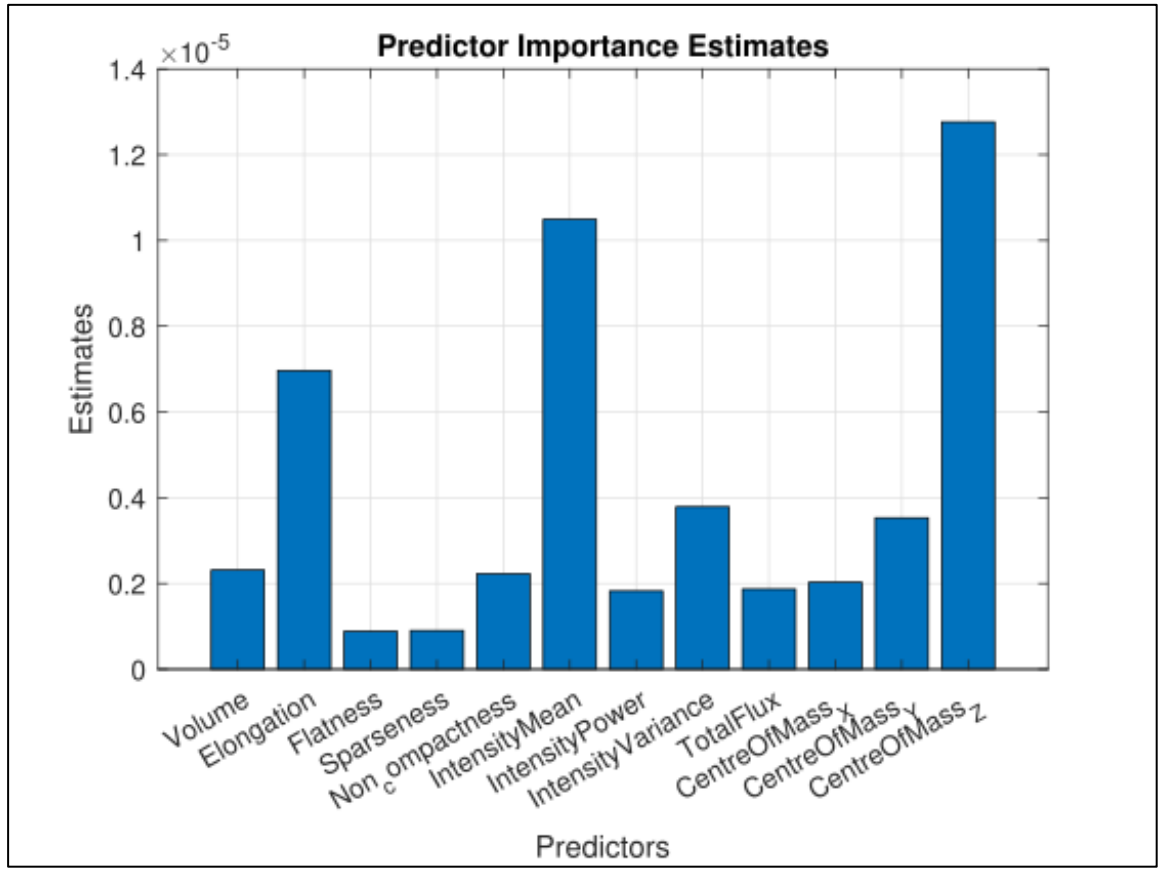


Fig. Feature's relevance results of the classifier, trained with data sample 1 (0.2% hight tumor percentage)

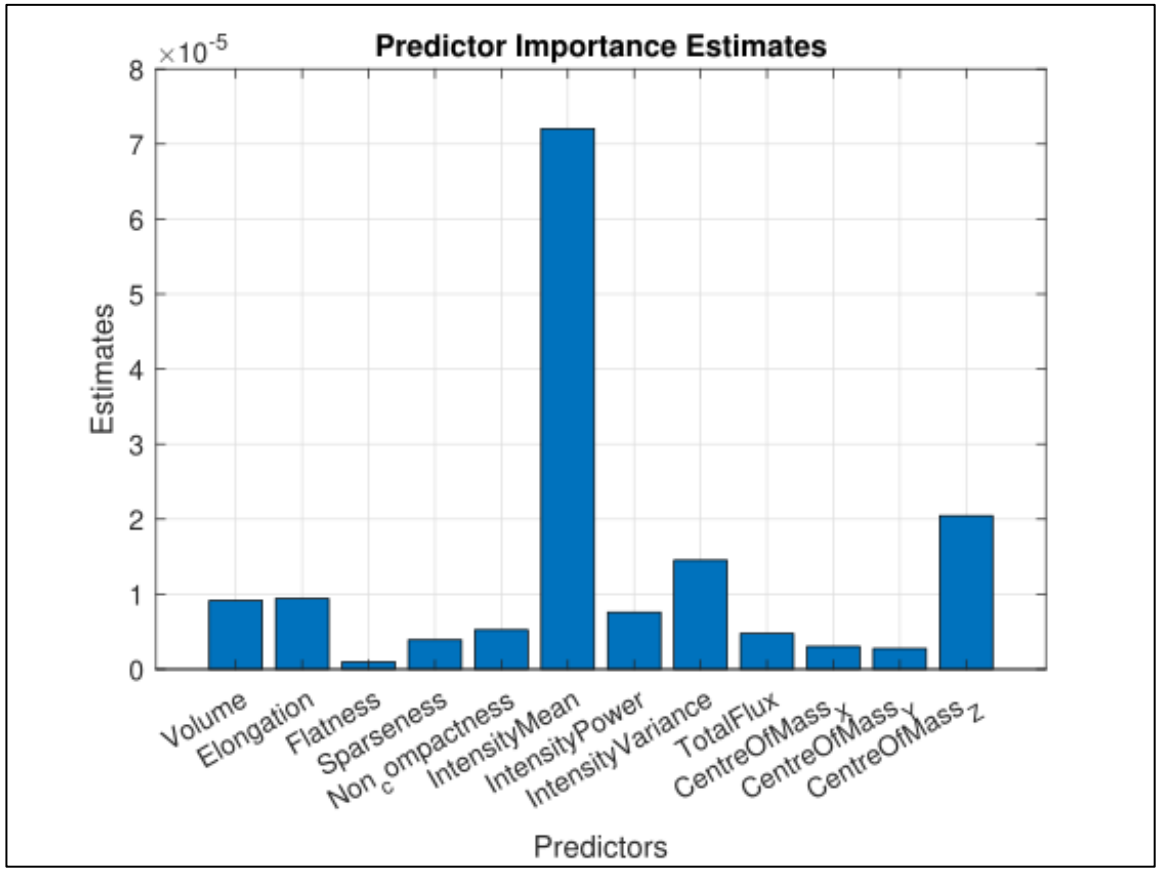


Fig. Feature's relevance results of the classifier trained with data sample 2 (10% hight tumor percentage)

Training data sam

university of groningen

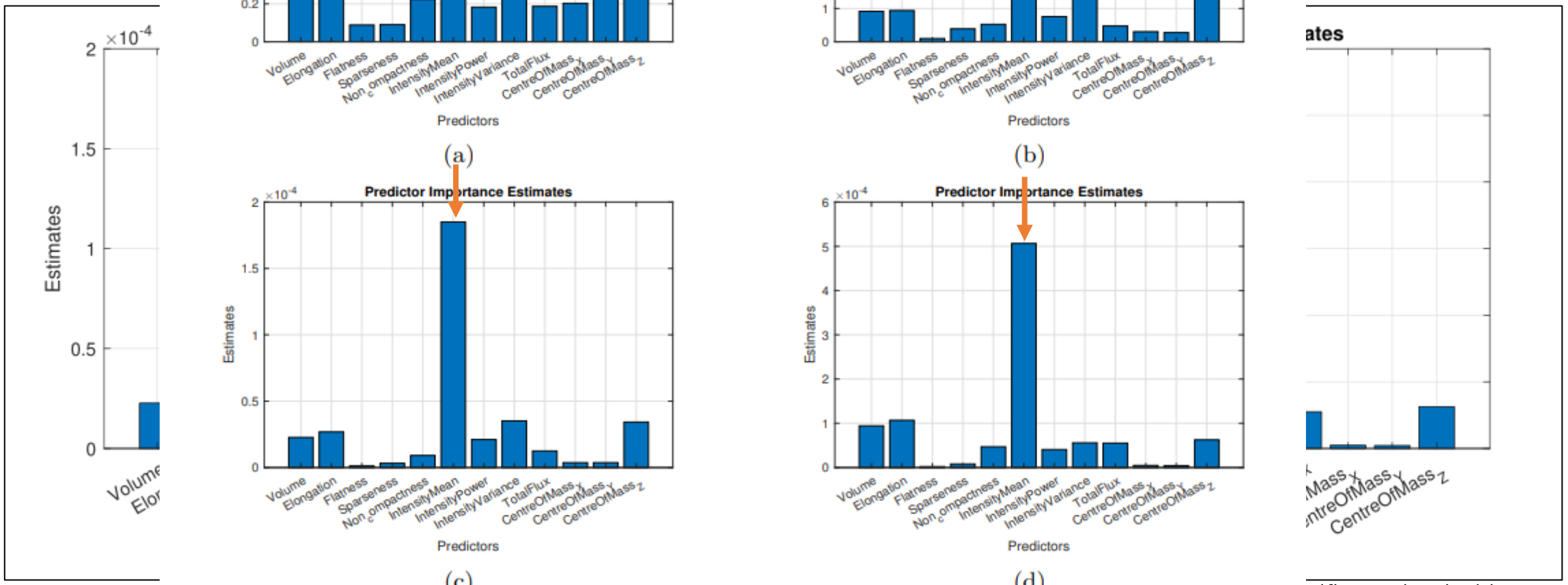


Fig. F … sifier, trained with data … age)

Figure 4.2: The Sub Figures 4.2a, 4.2b, 4.2c, 4.2d show the feature vector ranking by the Random Forest classifier, which is trained with biased data samples 1 to 4 with tumor volume accounting $\geq 90\%$ of total volume respectively.

Training random forest model with data samples of different high tumor fraction thresholds

university of groningen

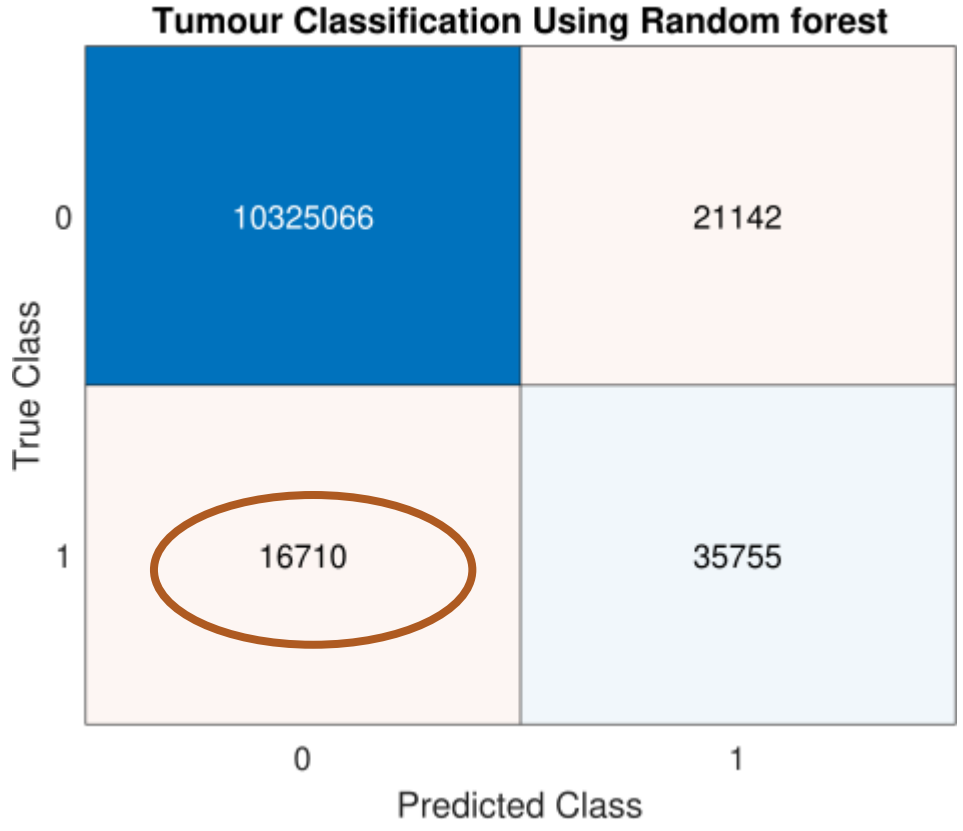


Fig. Classifier's results on the test set, trained with data sample 1 (<=10% High tumour fraction Threshold)

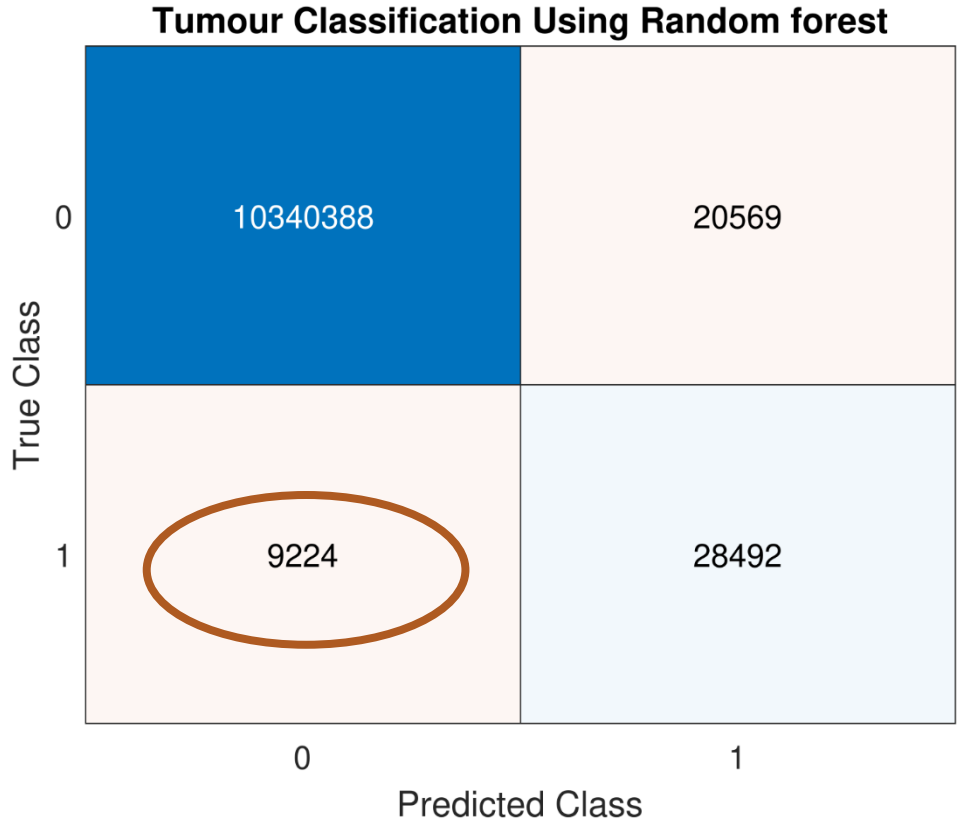


Fig. Classifier's results on the test set, trained with data sample 2 (<=30% High tumour fraction Threshold)

Training random forest model with data samples of different high tumor fraction thresholds

university of groningen

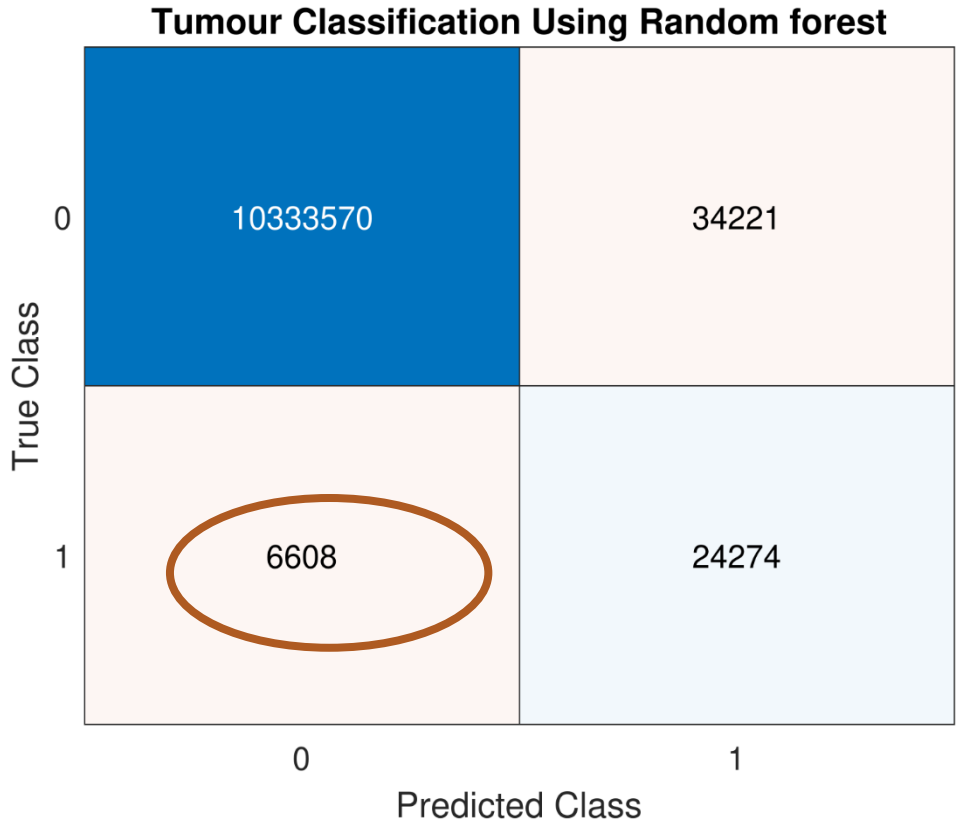


Fig. Classifier's results on the test set, trained with data sample 3 (>=50% High tumour fraction Threshold)

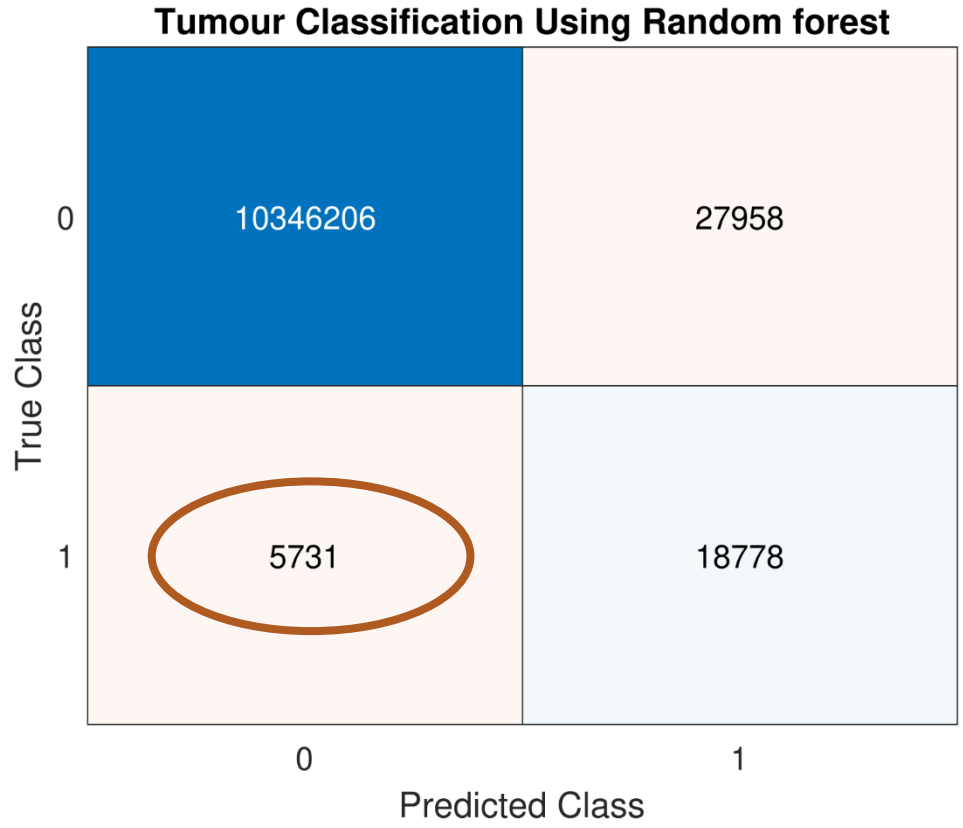


Fig. Classifier's results on the test set, trained with data sample 4 (>=70% High tumour fraction Threshold)

Training random forest model with data samples of different high tumor fraction thresholds (80% and 90%)

university of groningen

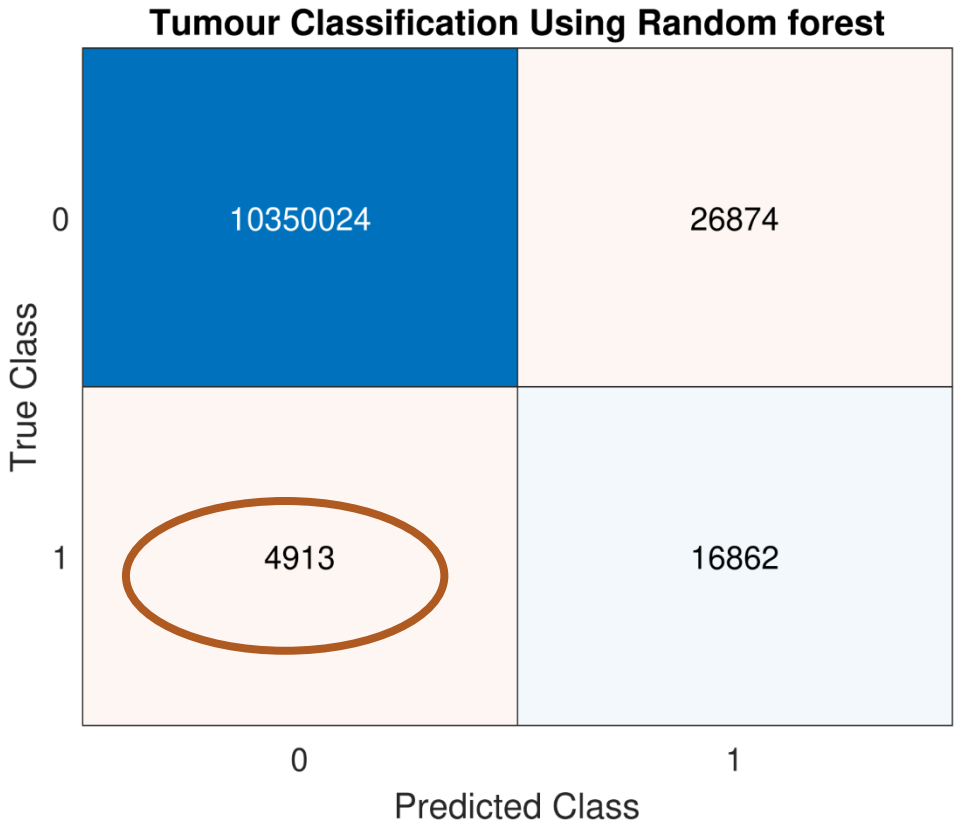


Fig. Classifier's results on the test set, trained with data sample 5 (>=80% High tumour fraction Threshold)

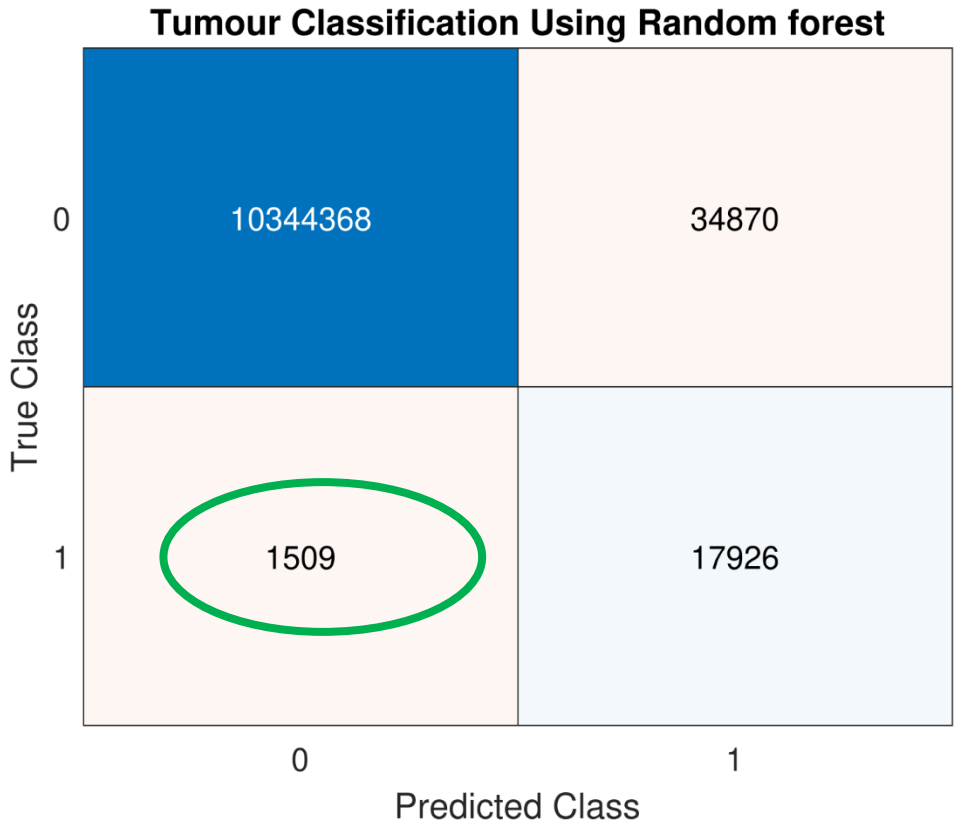


Fig. Classifier's results on the test set, trained with data sample 6 (>=90% High tumour fraction Threshold)

Training random forest model with data samples of different high tumor fraction thresholds

university of groningen

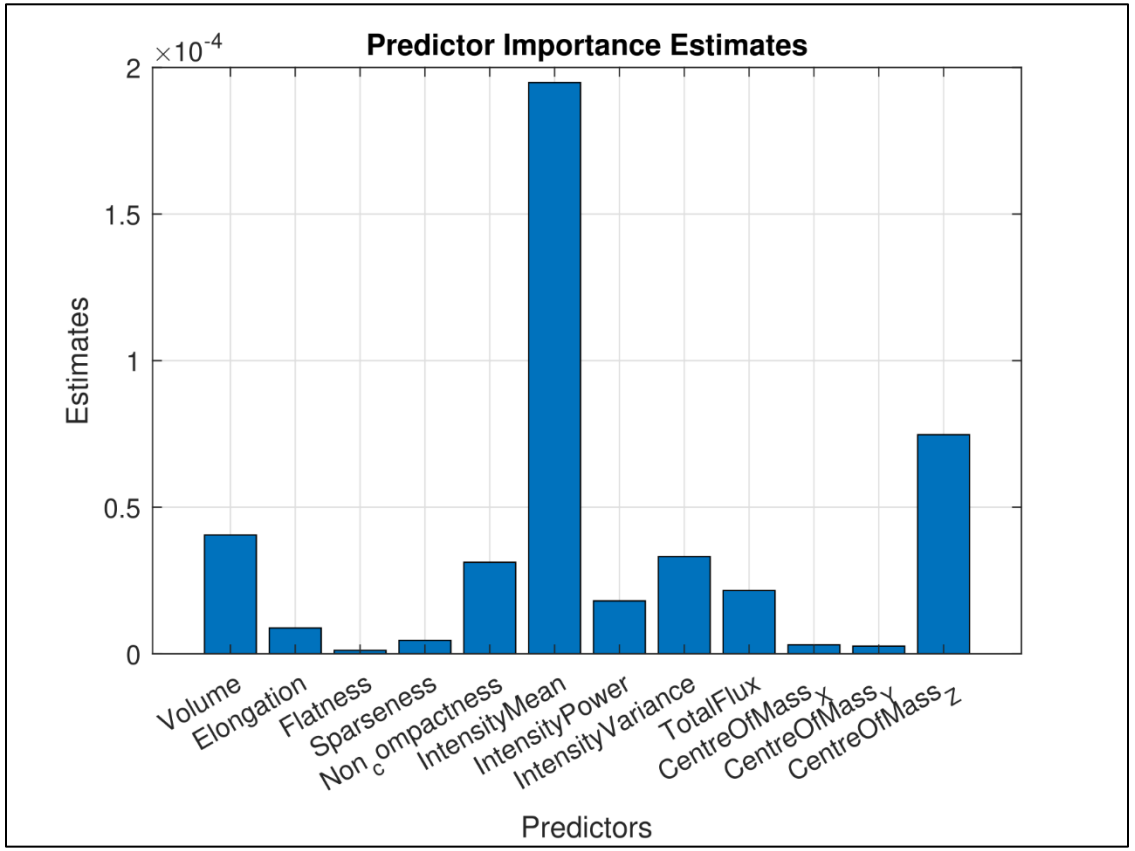


Fig. Feature's Relevance of the classifier trained with data sample 1 (>=10% High tumour fraction Threshold)

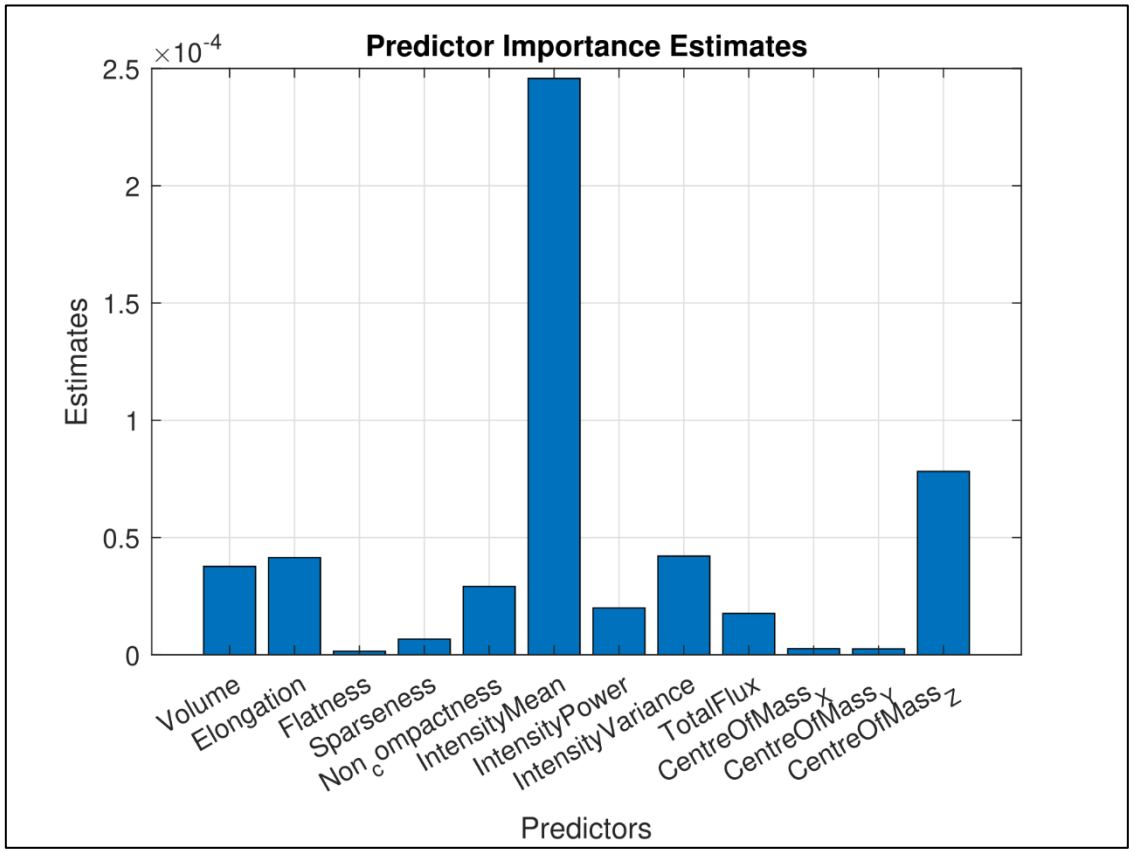


Fig. Feature's Relevance of the classifier trained with data sample 2 (>=30% High tumour fraction Threshold)

Training random forest model with data samples of different high tumor fraction thresholds

university of groningen

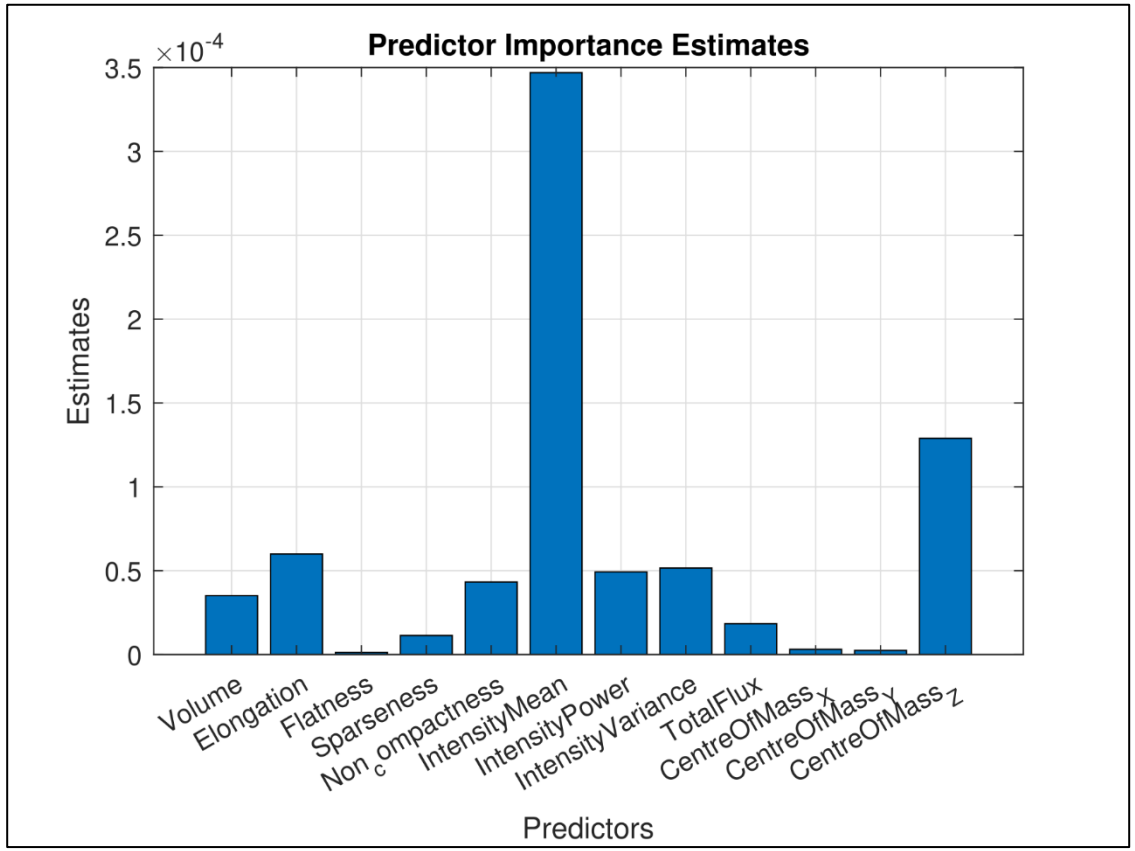


Fig. Classifier's results on the test set, trained with data sample 3 (>=50% High tumour fraction Threshold)

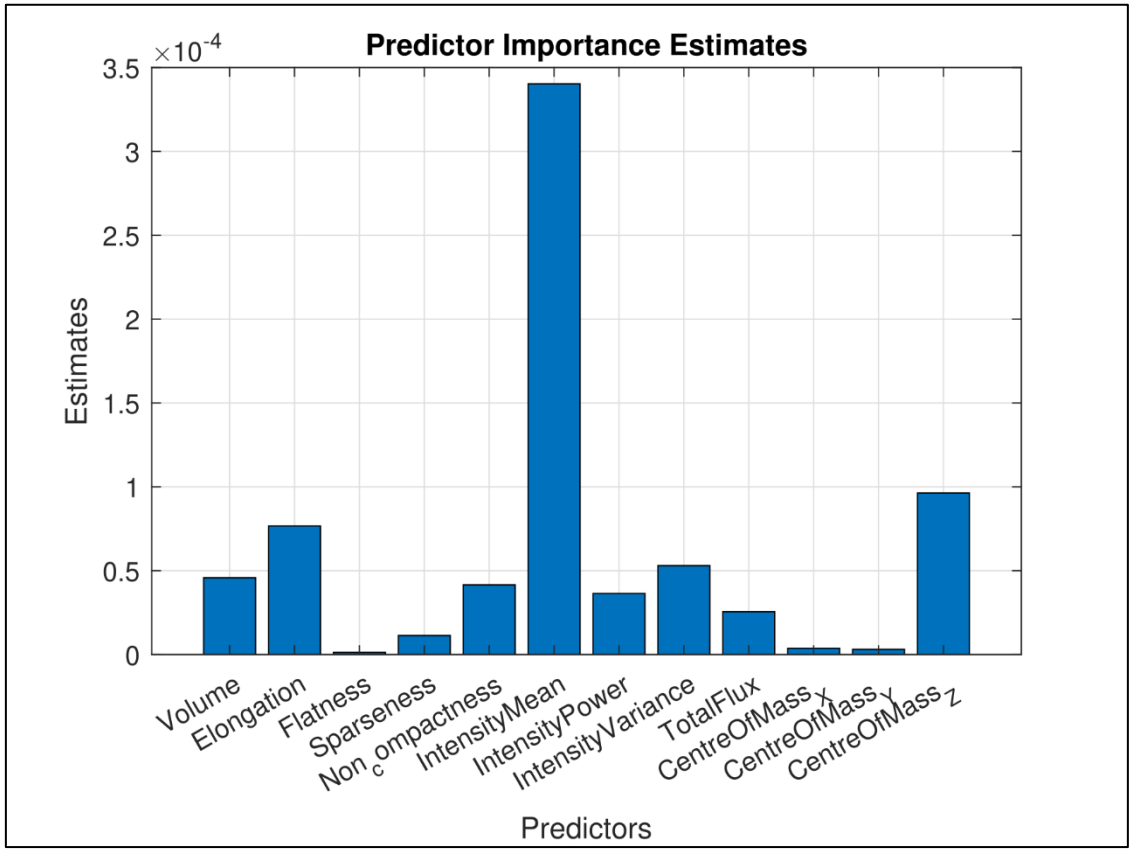


Fig. Feature's Relevance of the classifier trained with data sample 4 (>=70% High tumour fraction Threshold)

Training random forest model with data samples of different high tumor fraction thresholds.

university of groningen

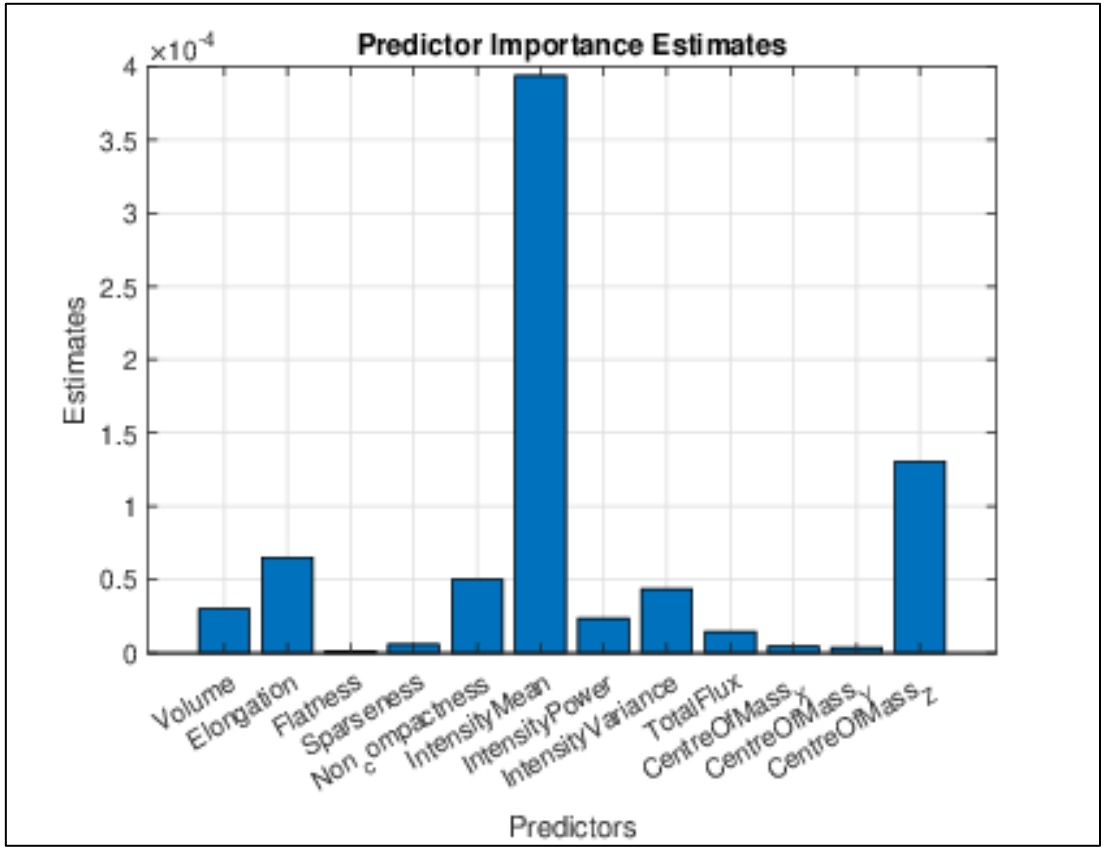


Fig. Feature's Relevance of the classifier trained with data sample 5 (>=80% High tumour fraction Threshold)

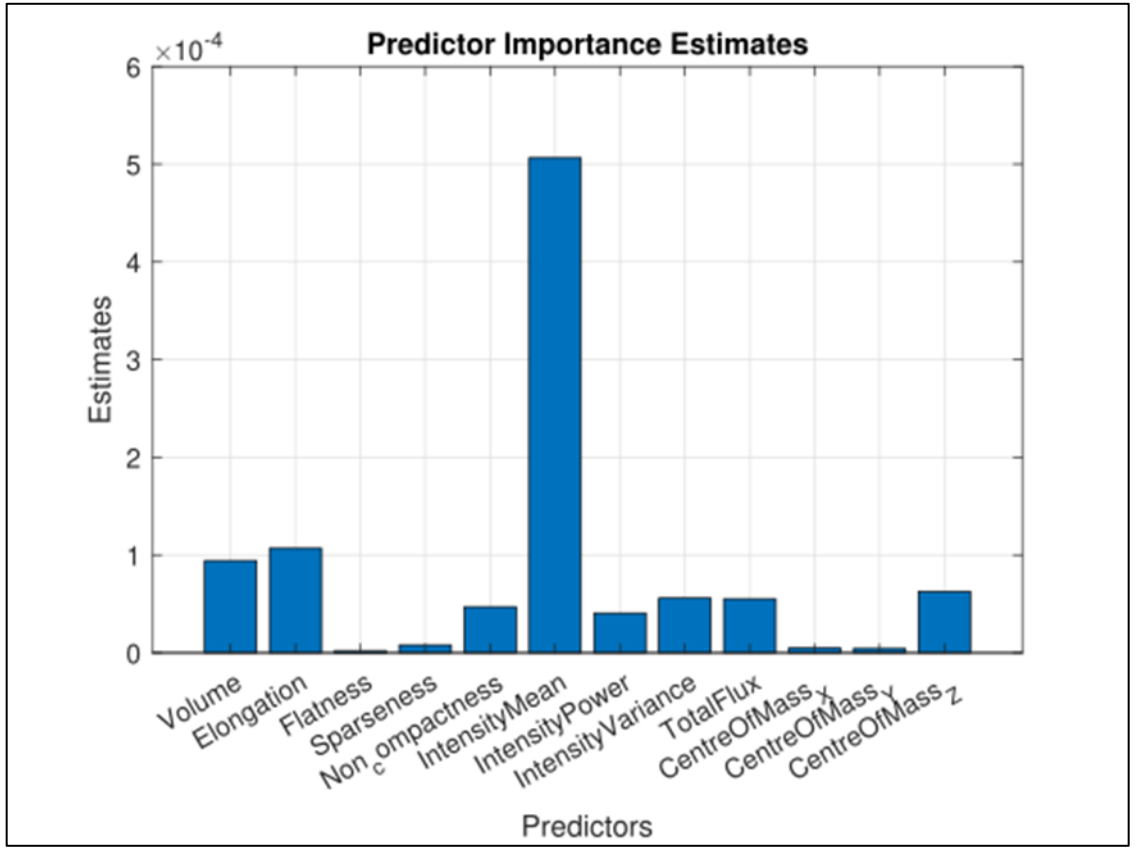


Fig. Feature's Relevance of the classifier trained with data sample 6 (>=90% High tumour fraction Threshold)

Performance Evaluation using ground truth volumes

university of groningen

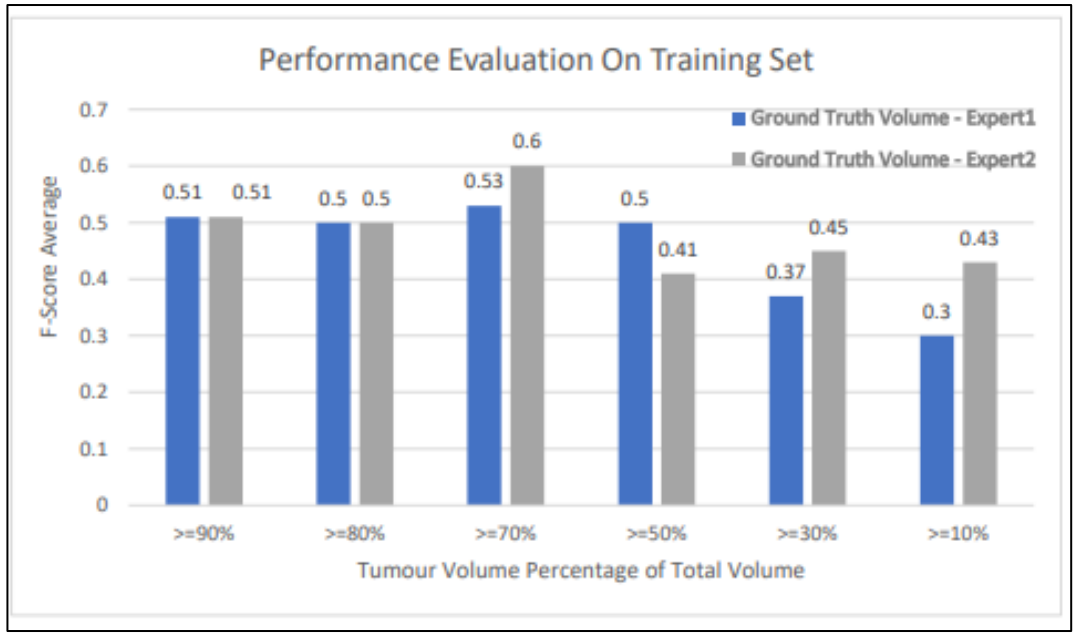


Fig. Performance Evaluation on filtered training set volumes with respect to two ground truth volumes respectively. The X-axis depicts the classifiers trained with training data samples of different thresholds of high tumor fraction, and Y-axis depicts the average F-scores obtained.

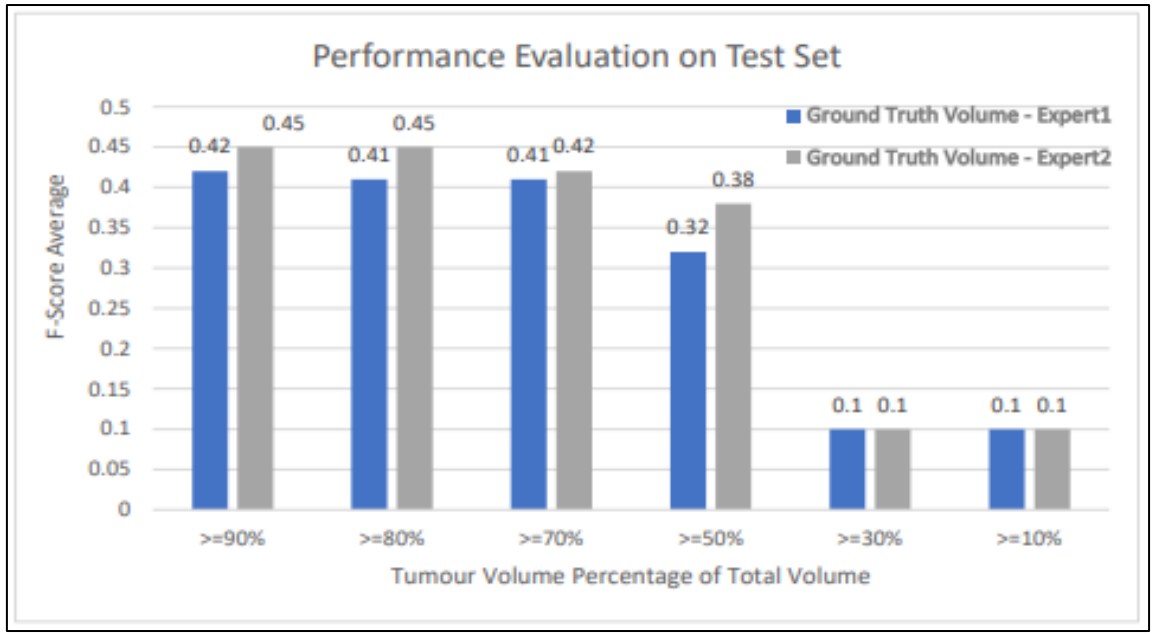
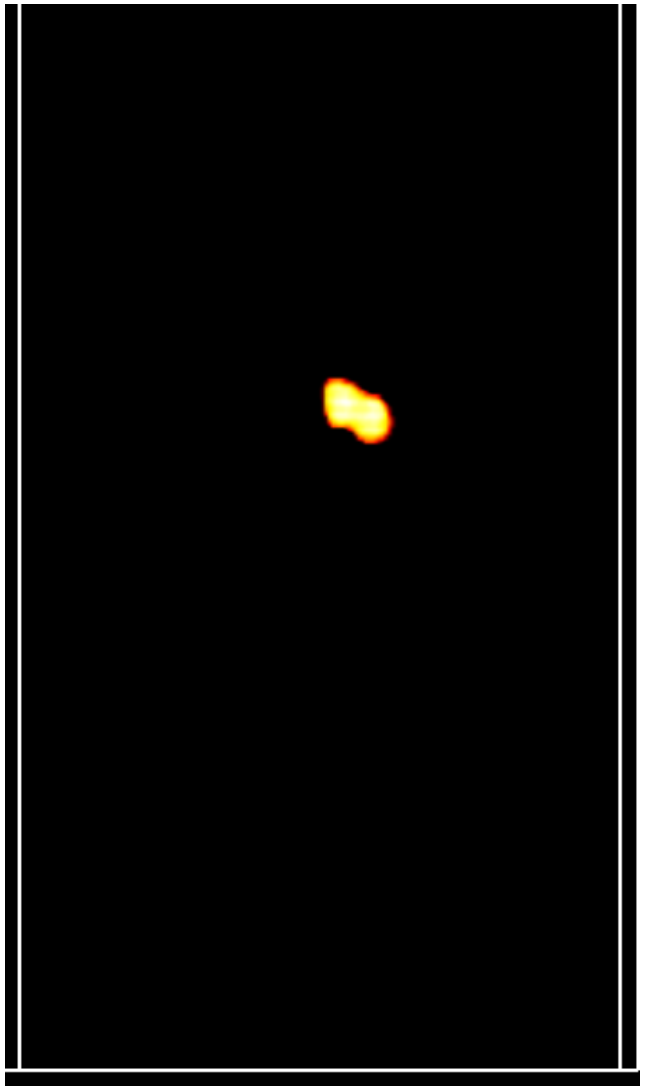


Fig. Performance Evaluation on test set volumes with respect to two ground truth volumes respectively. The X-axis depicts the classifiers trained with training data samples of different thresholds of high tumor fraction, and Y-axis depicts the average F-scores obtained.

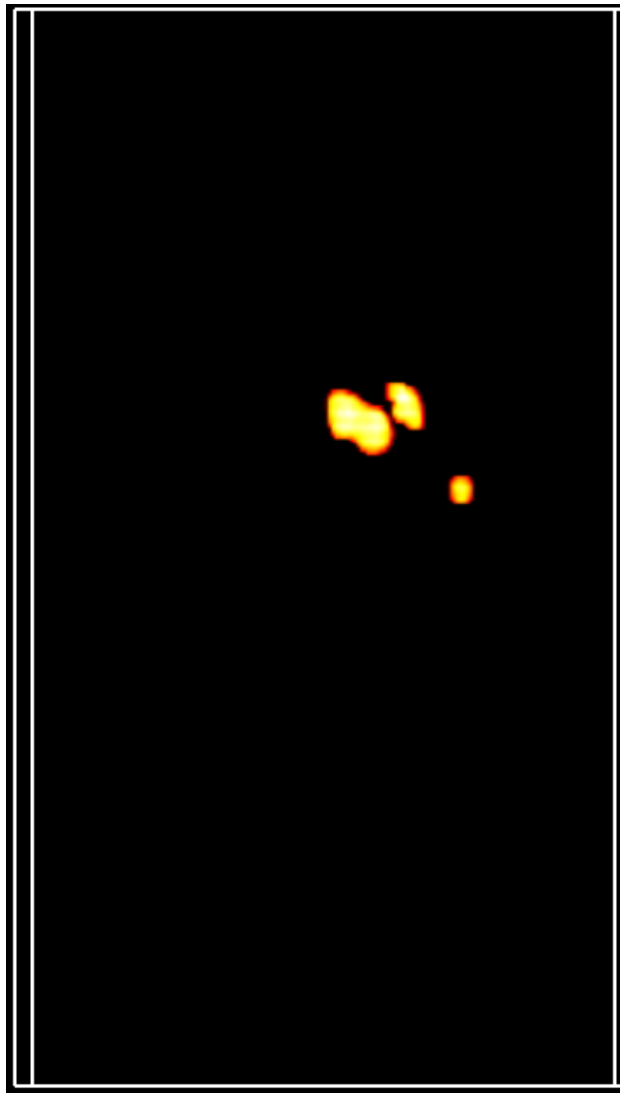
Performance validation using ground truth volumes – 228 volume with F-scores 0.85 and 0.88

university of groningen

Ground truth Volume 228 – expert 1



Ground truth Volume 228 – expert 2

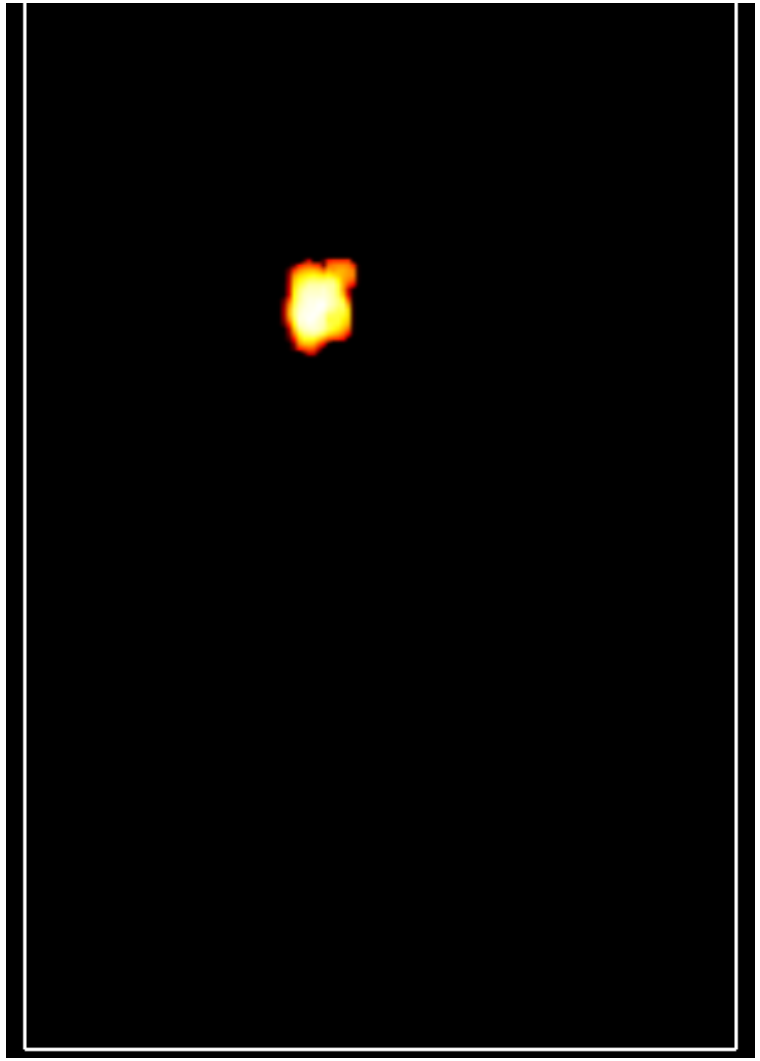


Classified Volume 228

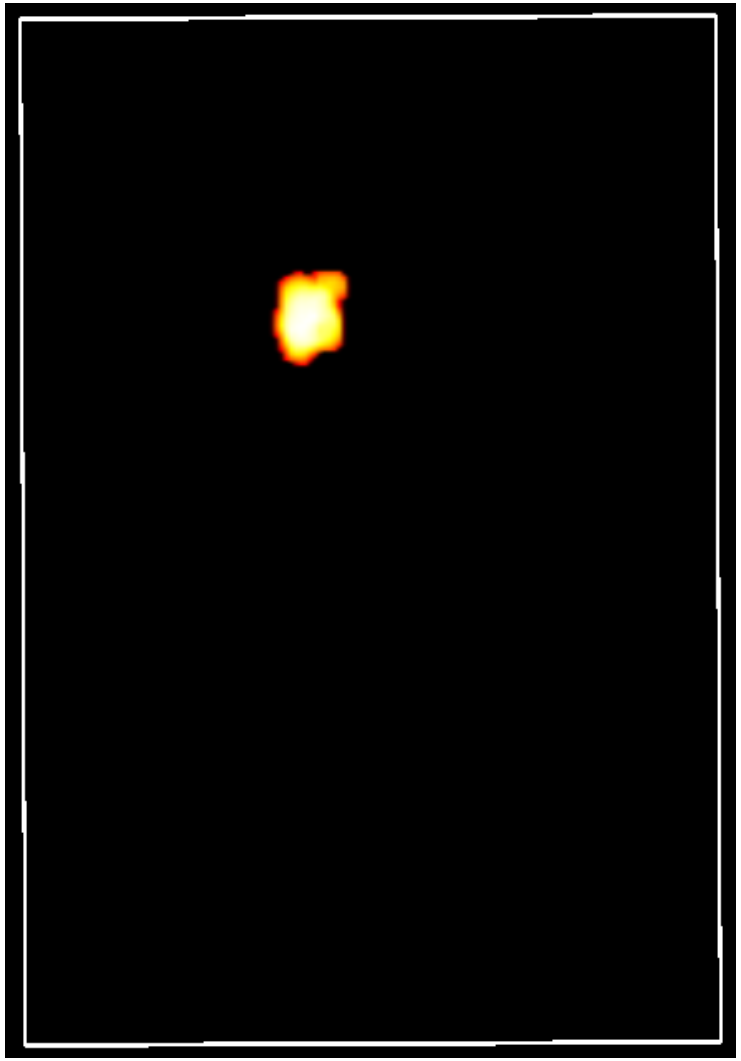
Performance validation using ground truth volumes – 238 volume with F-scores 0.74 and 0.76

university of groningen

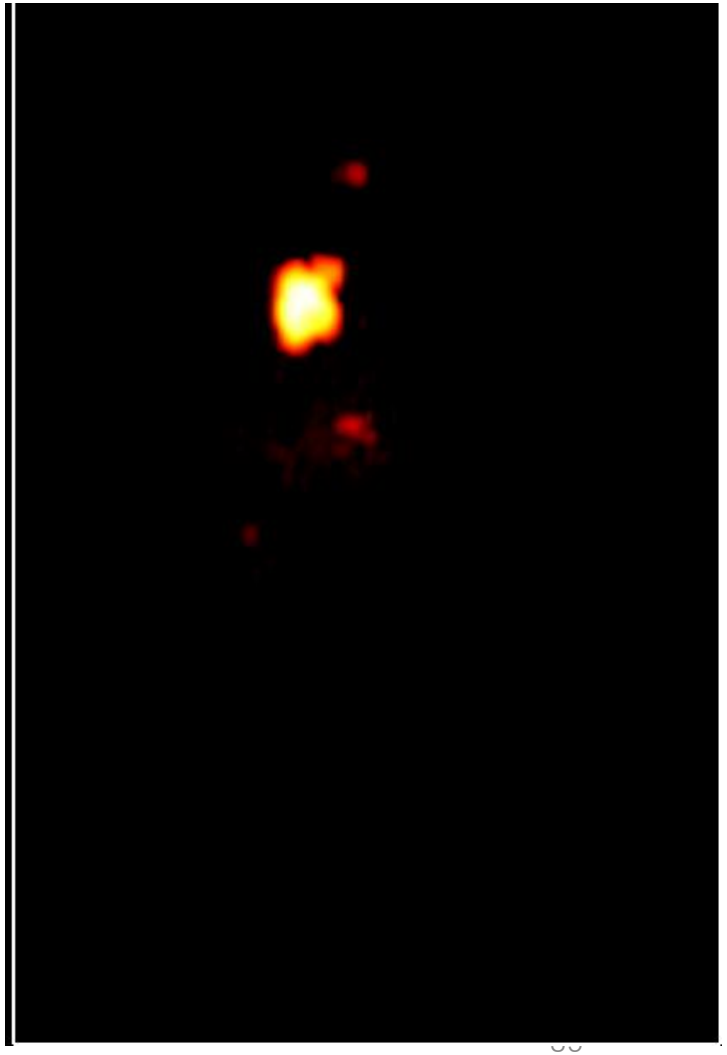
Ground truth Volume 238 expert 1



Ground truth Volume 238 expert 2



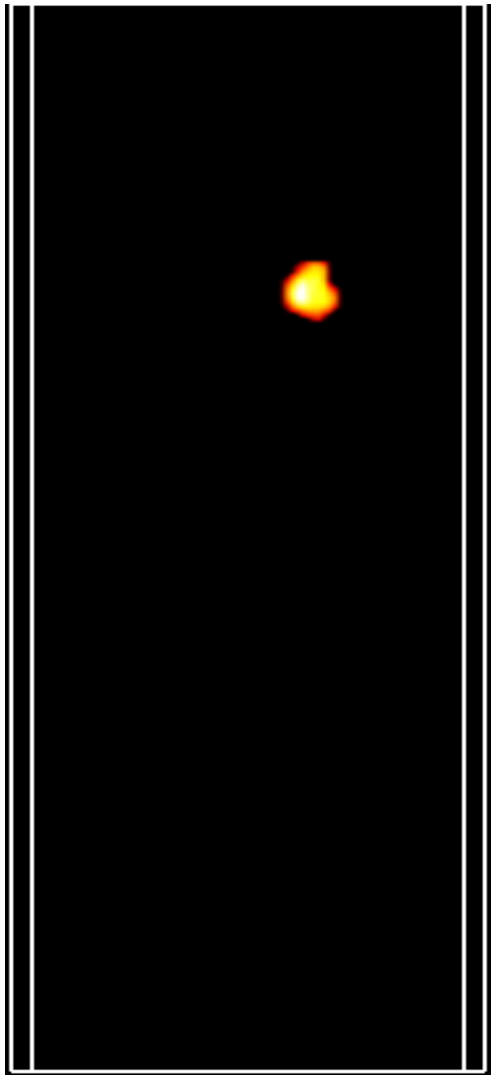
Classified Volume 238



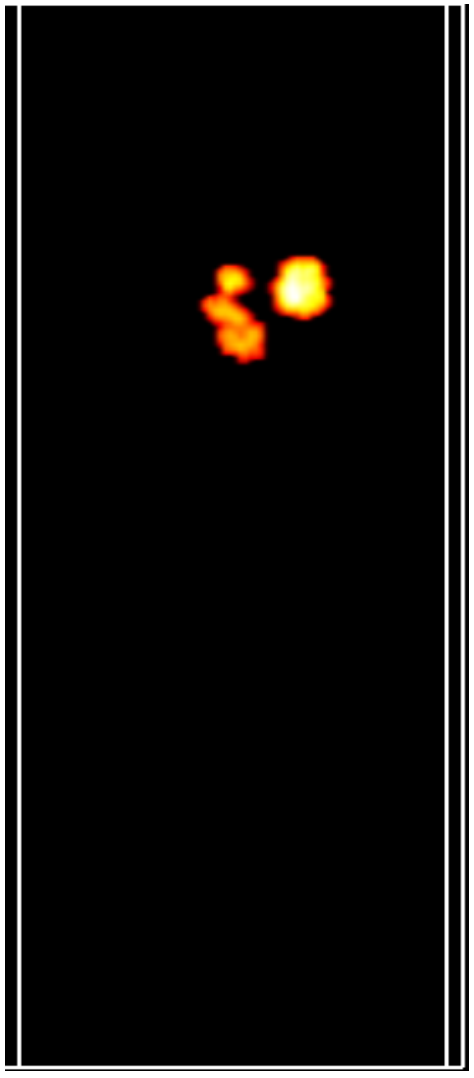
Performance Validation using Ground Truth Volumes – 230 Volume with F-scores 0.3 and 0.2

university of groningen

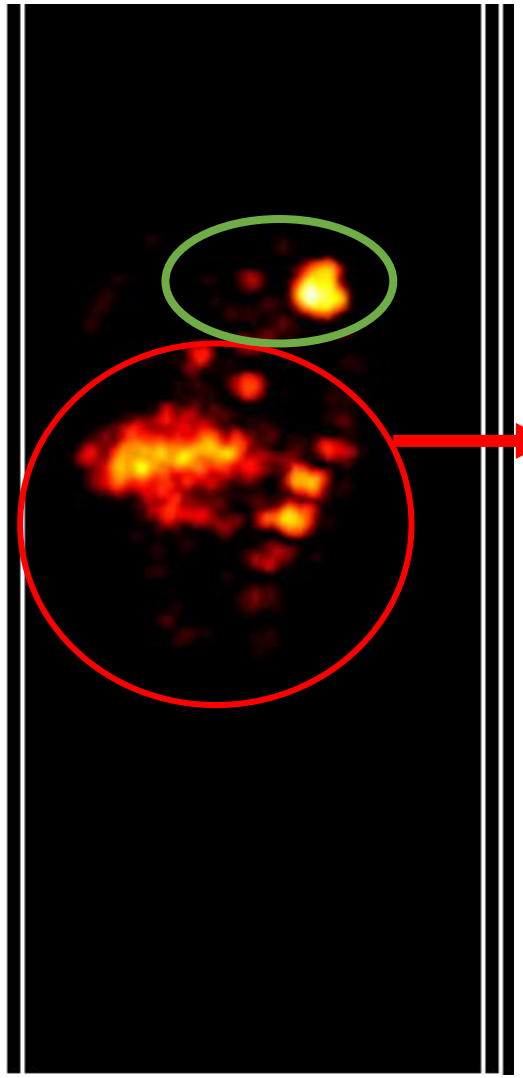
Ground truth Volume 230 – expert 1



Ground truth Volume 230 – expert 2



Classified Volume 230

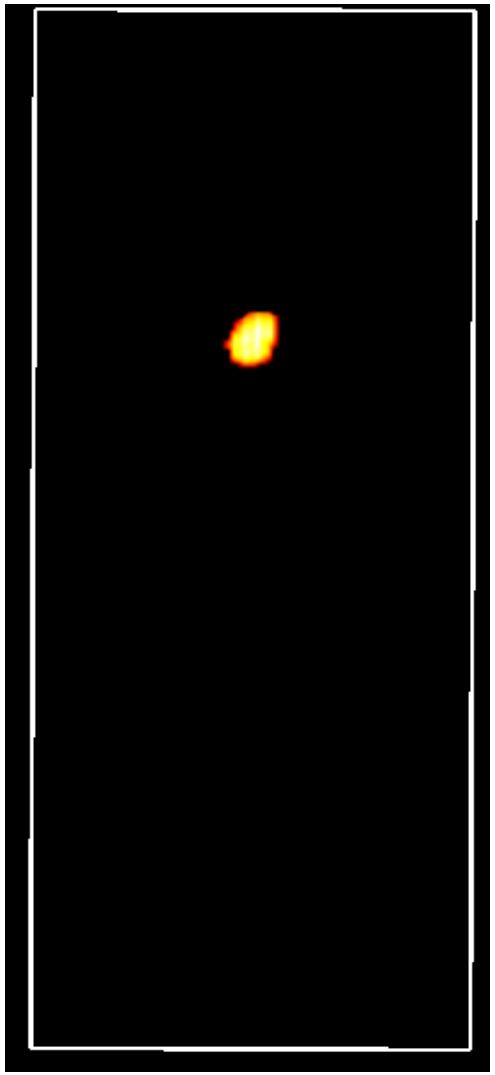


False positive – Indicating the classifier identifying vertebral column and Large Intestine.

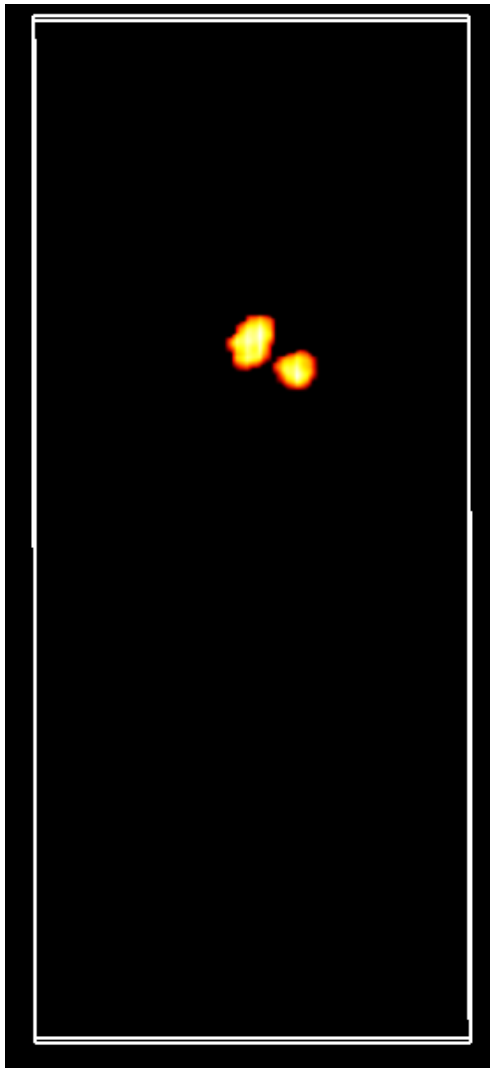
Performance Validation using Ground Truth Volumes – 236 Volume with F-scores 0.3 and 0.4

university of groningen

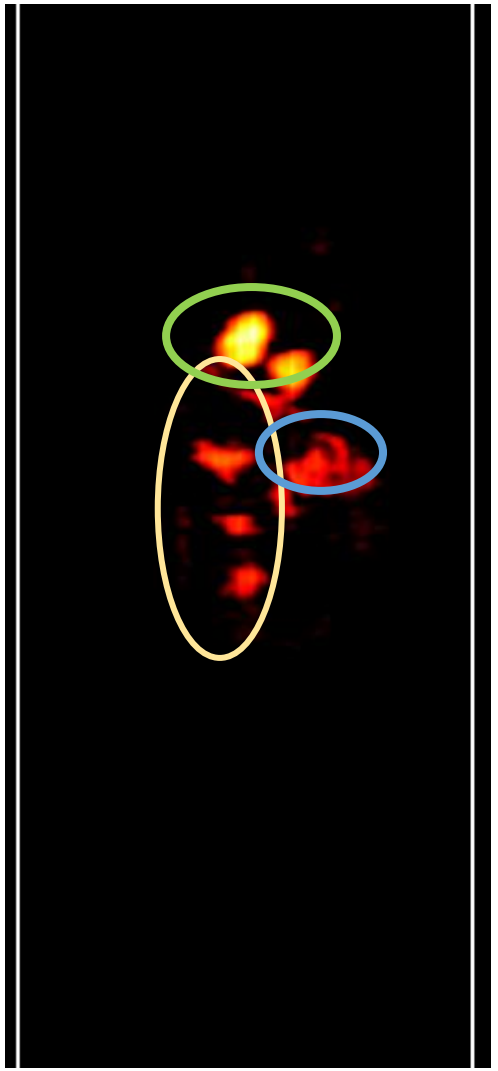
Ground truth Volume 236 – expert 1



Ground truth Volume 236 – expert 2



Classified Volume 236



False positive – Indicating the classifier identifying vertebral column and Large Intestine.

university of groningen

5 Conclusion & Future Work

Conclusion

university of groningen

In H. Gan et al. presented an exploratory tool using unsupervised machine learning technique to detect the lung tumour. This thesis implements supervised learning technique successfully.

The classifier trained with a higher tumor fraction threshold performs better.

The classifier trained with the training set with a significant number of high fraction tumor nodes (> 10% of the total nodes) provided a better performance.

The feature 'Intensity Mean'- most significant feature ranked by the method to detect the lung tumor effectively.

The false negative resulted to be the least for 50% high tumor percentage of the total nodes.

Similarly to that H. Gan et al. work, the classification method's sensitivity intensity mean is observed, led to identification of organs as tumor with higher intensity mean value.

Future Work

university of groningen

Using CT scans rather than the FDG-PET scans.

- FDG-PET scans can capture other inflammatory cells along with lung tumours.

Process concerning gray levels of the structures, as the gray level is not same as the PET scans.

References

[1] Manola, B. E. T. T. I. O., Giorgia, R. A. N. D. I., Raquel, N. D. C., Del Carmen, M. J. M., Artur, D. T., Nicholas, N. I. C. H. O. L. S. O. N., & Sandra, L. C. (2021). Lung cancer burden in EU-27.

[2] SNMMI, https://www.snmmi.org/AboutSNMMI/Content.aspx?ItemNumber=949.

[3] R. Jones, "Connected filtering and segmentation using component trees," Computer Vision and Image Understanding, vol. 75, no. 3, pp. 215–228, 1999

[4] M. Westenberg, J. Roerdink, and M. Wilkinson, "Volumetric attribute filtering and interactive visualization using the max-tree representation," IEEE Transactions on Image Processing, vol. 16, no. 12, pp. 2943–2952, 2007.