POLITECNICO
MILANO 1863

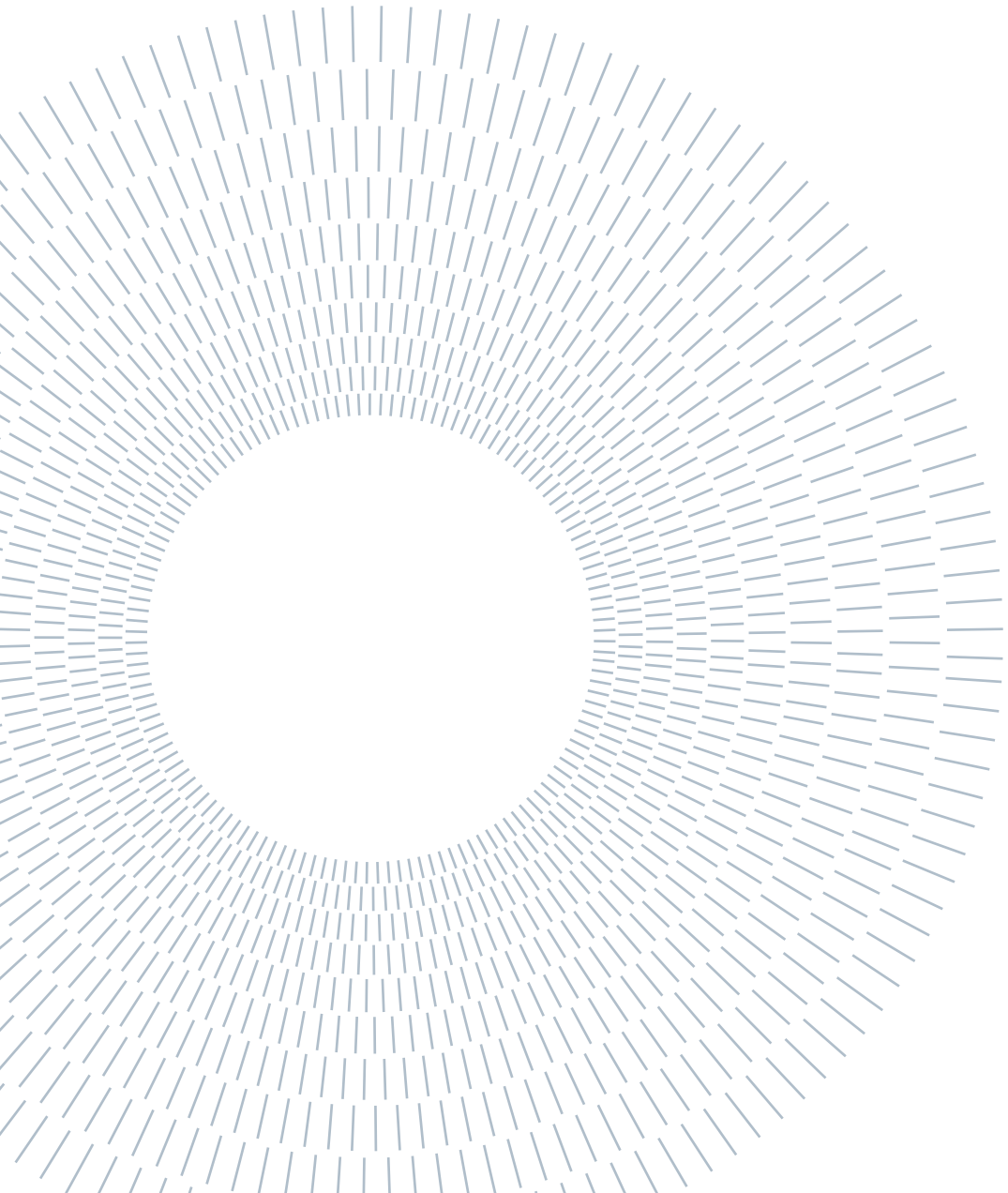
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

NAML Project

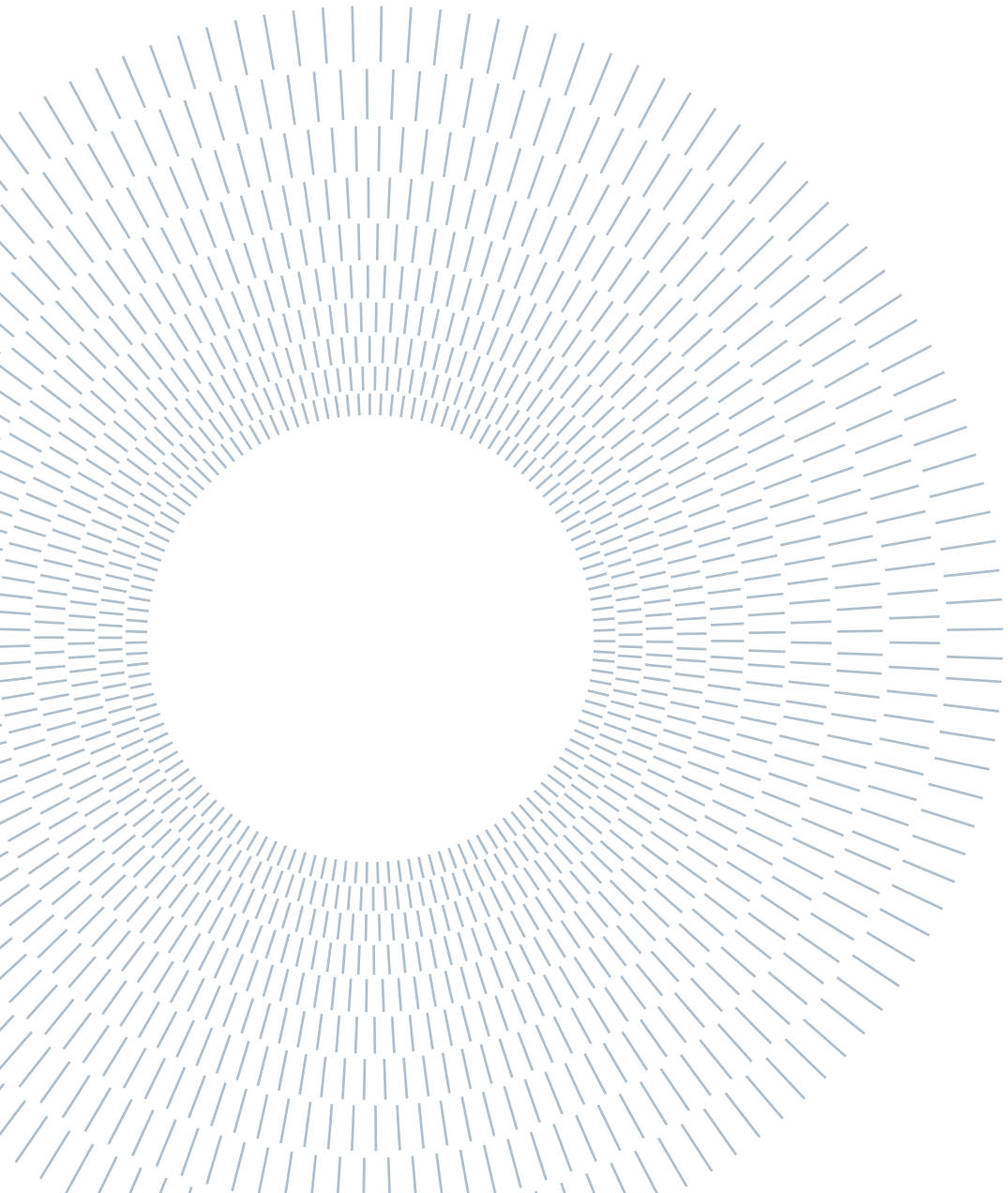
Author(s): **Tommaso Giordano (10723193)**

Luca Olivieri (10723789)

Academic Year: 2023-2024



Contents



1 Introduction

1.1. Scope

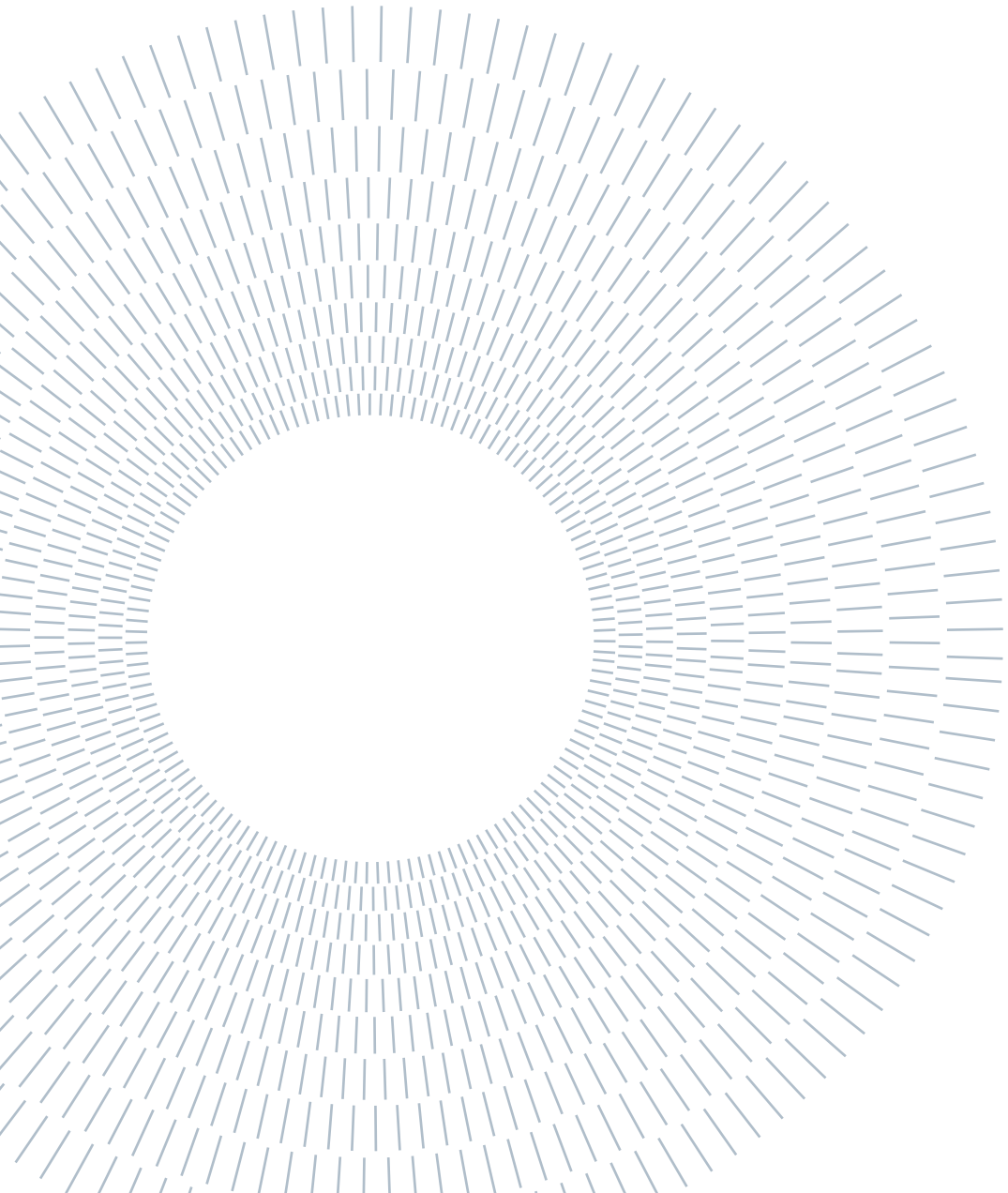
The purpose of the study was to replicate an article which attempts to face the problem of pancreatic tumour recognition in microscope images, through image segmentation, that is, identifying in the photo the regions of tumour nests with respect to sane ones.

This problem, traditionally, has been solved in a manual way, by pathologists: experts would analyse the image and manually segment the image in the tumour and non-tumour portions. However, this approach, while being in general convenient, leads to some issues:

1. **Difficulties of explainability**: experts are, usually, able to explain the reasoning behind their decision. Their operating is mostly based on their studies, the knowledge, and their expertise on the topic, but still they fail to explain some basic rule and assumptions to effectively solve the segmentation problem in an unambiguous and rational way. In short, they know how to do it, but they cannot explain why they do so.
2. **Variance of the results**: Following directly from the last point, the experts' results are quite subjective and tend to differ from one view to another. Experts of different knowledge and experience might come to different conclusion, and no absolute method other than bare consensus has been found. In such an environment, it is difficult to face the problem in a objective and rigorous way.

To deal with this issues as a whole, scientists have attempted to leverage imaging and AIA methods, however all these techniques did not completely succeeded to completely address this issue.

Eventually, data scientists have attempted to solve the problem through supervised learning: treating it as an image segmentation problem, sophisticated specialised deep neural networks can be trained on labeled data, in order to predict the segmentation of the image, delegating to the network the burden of extracting some solid and reusable knowledge from the data.



2 | Article Review

We are replicating the article *"Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning"*, by Muhammad Khalid Khan Niazil, Thomas Erol Tavolaral, Vidya Arole, Douglas J. Hartman, Liron Pantanowitz, Metin N. Gurcan.

2.1. Dataset

As written on the original paper, the database consists of 33 whole slide images of Ki67 stained neuroendocrine tumor biopsies acquired from 33 different patients. All slides were anonymized and digitized at 20x magnification using a high-resolution scanner (Aperio ScanScope, Leica Biosystems) at 0.2437 microns per pixel squared.

Portions of whole slide images were annotated for tumor positive/negative and non-tumor positive/negative regions by an expert pathologist A. Each annotation was sampled for 64x64 pixel tiles at 20x magnification.

In total, this procedure lead to the generation of 138,056 tiles of 64x64 pixels each, divided in 129,024 tumor and 9,024 non-tumor tiles. As is it noticeable, the resulting dataset is very unbalanced: for each non-tumor tile, there are approximately 14 tumor tiles.

2.2. Models

The models leveraged for the problem were two convolutional deep neural networks, both pre-trained on Imagenet dataset and achieved positive results on testing evaluations.

- **Inception v3**: in the study, it is the main model used for the study, more sophisticated and promising, adopting a transfer learning approach.
- **AlexNet**: this model is used for a baseline comparison, trained through fine-tuning.

2.2.1. AlexNet

AlexNet is a large, deep convolution neural network trained on the Imagenet large Visual Recognition Challenge dataset from 2012. A standard dataset in computer vision classification tasks, consisting of 1000 classes. It consists of five ReLU convolutional layers followed by three fully-connected layers and a final softmax for classification.

2.2.2. Inception v3

Inception-v3 is a large, deep convolution neural network trained also trained on Imagenet. Inception-v3 is distinguished from conventional CNNs in four respects:

1. **1x1 convolutions**: they reduce computation through dimensionality reduction,
2. **Inception modules**: they allow the network to choose which size convolution at each layer is best by performing smaller, parallel convolutions of different sizes, whose filters are concatenated as a final output. Conventional CNNs are limited by fixed convolution sizes.
3. **Label smoothing**: it is a regularization method that replaces target vector 0s and 1s used for classification of k different classes with ε/k and 1-ε(k-1)/k, respectively, where ε is the estimated proportion of mislabeled training samples.
4. **Auxiliary classifiers**: Finally, Inception v3 contains two auxiliary softmax classifiers, connected to the outputs of two intermediary Inception modules. In a sense, these allow the network to choose at which Inception module output it classifies, rather than propagating to the end.

2.3. Training

- **AlexNet** (fine-tuning): It is trained with stochastic gradient descent with momentum and decay rate of 0.9, a mini-batch size of 100 for 10 epochs, a learning rate of 0.0001 with an exponential decay of 0.9, and employs cross-entropy for loss. The first decay rate reduces the effect of momentum by a factor of 0.9 every epoch, and the second decay rate reduces the learning rate by a factor of 0.9 every epoch. Rather than training solely the final softmax layer, the error is allowed to back-propagate through the entire network, allowing for fine-tuning of each weight.
- **Inception v3** (transfer learning): Inception v3 is trained with stochastic gradient descent with momentum and decay rate of 0.9, a learning rate of 0.045 with an exponential decay of rate of 0.94, and employs cross- entropy to measure loss. The first decay rate essentially reduces the effect of momentum by a factor of 0.9 every epoch, and the second decay rate reduces the learning rate by a factor of 0.94 every epoch. Inception v3 has learnt succinct features to successfully categorize data into 1000 classes. We use transfer learning to exploit these rich set of features, i.e., we used Inception v3 as a feature extractor and trained solely its softmax classifiers (auxiliary and principal) on our two class (tumor and non-tumor) dataset. The learning rate was set to 0.01, and a mini-batch size of 100 was used over 3000 iterations.

2.4. Validation

Validation was performed on a validation set, generated through a training-validation split. For both of the models, 10% of the training data was utilized for validation.

2.5. Testing

2.5.1. Prediction Procedure

The generation of the image segmentation prediction follows a structured methodology.

To start with, the testing dataset is created by extracting a set of 2100×1300 pixels images (referred to as HPF images in the article) from one of whole slide images specifically left out from the training phase of the model. These images are processed in the following way: a 64×64 pixels sliding window passes over the test input image with a step size of 8 pixel, producing a set of partially overlapping tiles.

As each tiles are fed to the model, a map keeps track of the number of times a certain pixel is passed over, while two other maps are updated with number of times a pixel is classified as non-tumor and tumor. In such a framework, the number of tumor hits maps averaged by the third map results in a **probability map**, showing, according to the model, how likely a pixel is to be a tumor pixel. E.g., if a certain pixel is hit 7 times (thus, in 7 different, but overlapping, tiles), and in 4 cases this resulted in a tumor classification, then $\frac{4}{7}$ is an estimate of the tumor classification probability.

Finally, the probability is thresholded with respect to the value 0.5. That is, the final **binary mask** counts as tumors the pixel that were hit as such in the majority of the cases.

2.5.2. True Labeling Procedure

While pathologist A took part in the training generation procedure, in order to maintain the unbiasedness of the study, two other pathologists, B and C, have been involved in the making of the testing process. Both of them have been instructed, individually and separately, to operate this way:

1. Observe the model's prediction decision boundaries of the testing data.
2. Leave unchanged the prediction's portions if they agreed with the model's annotations.
3. Edit and draw new decision boundary if they did not agree with the predicted annotations.

The model's annotations were considered accurate when both pathologists left computer annotations unchanged. However the agreement between the two experts drops to as low as 83% in non-tumor regions. To deal with this problem two consensus methods were proposed: the first one by treating pathologists edits by means of logical `AND` operation (referred as C1 reading) and the second one using the logical `OR` operation (referred as C2 reading).

Ultimately, the true mask derived from the correct classifications, untouched by the B and C, and those which have been considered as such according to the two consensus reading C1 and C2. An average of 3909 tumor and 274 non-tumor tiles were used for testing.

Remarks

Despite the apparent simplicity of the consensus approaches adopted, still it is very uncertain how the operation are actually implemented, especially with respect to opposite edits (in which the experts edit the same portion but assign different classes). This issue is, in particular, significant when we consider that the article's authors stated that some pixels were discarded from the validation process, as pathologists disagreed on their classification. Unfortunately, no further explanation is provided.

2.6. Experimental Results

- **Inception v3**: This model proved to be an excellent one for this segmentation problem: the accuracies are close to 99%, but still losing some reliability regarding non-tumor classifications. The difference of results between C1 and C2 is quite negligible, proving that the model achieved high quality predictions that left little space for errors. The advantage of this model over AlexNet is that, through transfer learning, scientists were able to rely on the model's complexity while reducing the risk of exposure to over-fitting, as this learning approach affects only how the model combines the features extracted from input images (as sort of high level reasoning), without altering the actual pattern recognition capabilities and, thus, the generalisation ability (on a lower level).
- **AlexNet** This model showed positive results, especially regarding tumor classification. It instead had much lower reliability on non-tumor classifications. Hence, this model's prediction capabilities are negatively biased towards non-tumor. This is most likely due to the fact that fine-tuning an entire network often leads to over-fitting, which leads to a significant degradation of the quality of the prediction results.

3 Methodologies

While we tried to adhere as closely as possible to the methodologies described by the paper, we encountered instances where deviations were necessary either because of the **lack of specific details** in the original paper, **unavailability of all the necessary** data, or of **access to expert pathologists**. In this section we are going to focus on the differences between our approach and the one proposed on the paper and outline how this affects the results.

3.1. Dataset

The dataset published by the authors consists of 30 images with a resolution of 2100×1300 px.

However, this dataset represents only **a portion of the complete dataset** used in the original paper and, in particular, we have reasons to believe that it consists of the data used to perform testing, as its size matches the original testing one. In our replication effort, due to the unavailability of the entire dataset, we were constrained to utilize this subset for training, validation, and testing purposes.

This dataset is **highly unbalanced** between tumor and non-tumor cells. Tumor cells are about 10 times more present than non-tumor ones. For this reason we opted to train our models using a **balanced** and **non-augmented version** of this dataset, obtained by **clipping** the considered amount of images to the size of the minority class (in our case, the non-tumor class), in order to achieve a 50/50 dataset. Furthermore, **training images have been normalised** (each channel has been transformed separately). We also tried, in some training scenarios, to employ simple data processing techniques, as an attempt to address unbalances while minimising the amount of discarded majority images.

The ground truth data shared by the authors is also different from the one used in the paper. While the authors utilized input from two different pathologists and merged them with two different consensus criteria, we only had access to a single ground truth mask of unspecified origin and consensus criterion.

3.2. Models

We used **default pre-trained models** for both AlexNet and Inception-V3, described in section 2.2, previously trained on ImageNet dataset and found to be successful in the image classifi-

cation and segmentation task. We changed the classifier of both networks to adapt them to a binary classification problem: **both classification layers were augmented** with additional progressively shrinking layers in order to ease the features abstraction and classification process as a whole.

The following is the classifier that we used for AlexNet:

Layer Type	Input Dimension	Output Dimension
Linear	4096	1000
Linear	1000	512
Linear	512	128
Linear	128	32
Linear	32	2

Table 3.1: AlexNet classifier architecture

And the next one is the classifier that we used for Inception-V3:

Layer Type	Input Dimension	Output Dimension
Linear	2048	512
Linear	512	128
Linear	128	32
Linear	32	2

Table 3.2: Inception-V3 classifier architecture

We used ReLu activation functions for both classifiers.

3.3. Training

The training has been performed, as in the original paper, leveraging **fine-tuning** with AlexNet, and **transfer learning** with Inception-V3.

Given the size and unbalance issues that undermined the usability of the dataset, many optimisers, loss functions, training parameters, data collection and processing techniques have been employed to achieve optimal training results.

However, many of these approaches found to be ineffective and thus discarded early on. In total, 4 version of models have been cross-validated: their training settings are illustrated in the following table.

ID	Model	Optimiser	Loss function	Batch size	Epochs
A1	AlexNet	Adam(lr $= 10^{-4}$)	XEn	100	5
A2	AlexNet	AdamW(lr $= 10^{-4}$, $\lambda = 0.1$)	XEn	100	5
I1	Inc.-V3	Adam(lr $= 5 \cdot 10^{-4}$)	FL($\alpha = 1, \gamma = 2$)	100	5
I2	Inc.-V3	AdamW(lr $= 5 \cdot 10^{-4}$)	XEn	100	5

Table 3.3: Training parameters

Here is a list of methods that were employed during some of the training attempts but then disregarded for their ineffectiveness:

1. **Dataset generation**
   (a) **Full dataset**: leveraging the full dataset for the training delivered negative results due to its unbalance. The models trained through this were severely biased towards tumors (as expected due to the greater presence of tumor samples), leading to predictions which presented many more tumor labels than necessary.
   (b) **Over-sampled dataset**: We over-sampled the minority (non-tumor) class in order to equalise the frequency of image classes taken into account during the training. Reasonably, this helped with the bias but led to significant over-fitting.
2. **Data processing**
   (a) **Data Augmentation**: All images of both classes, with some independent probabilities, have been subject to:
      - random horizontal flipping,
      - random vertical flipping,
      - random rotation by 20° degrees.

      These transformation have been set in addition to the baseline normalisation. Unfortunately, training evaluation metrics showed no improvement: small transformation probabilities showed no significant results; too high probabilities worsened visibly the training as it was not able to converge to a optimal minimum, without actually scoring a much better unbiasedness.
   (b) **Image Equalisation**: Images have been fed to an histogram equalisation filter provided by the library ComputerVision2 (CV2), which automatically enhances the contrast and the details visibility. This approach, with the hope to provide more visual separation from tumor and non-tumor areas, led to no significant results.
3. **Loss function definition**:

(a) **Custom class weighting**: we increased the weight assigned to non-tumor training points so as to incentivise the optimiser to prioritise minority samples. This lead to beneficial results with some predictions and negative results with others. Due to this unrobustness, this approach was discarded.

4. **Optimiser definition**:

   (a) **SGD**: We initially leveraged stochastic gradient descent with momentum = 0.9 (as in the original paper) for both models. However, our analysis denoted that Adam was superior to it, as it was able to minimise at a faster rate (especially in the beginning, in the first iterations), and, consequently, to approach better weights in the same number of iterations.

In additions to these approaches, a generic tuning of training parameters were adopted. Hyper-parameters such as:

- batch sizes,
- number of epochs,
- learning rates,
- weights decay terms,
- custom class weights,
- focal loss multipliers,

were tuned by picking the best values resulting from a validation phase.

3.4. Validation

A **10-fold-cross-validation** was performed in order to validate models that faced different training settings and select the most performing and robust one. We wished to employ Leave-One-Out Cross-Validation (LOOCV). However, constrained by computational limitations, we could not rotate all 30 images. Consequently, we opted to rotate through a subset of 3 images to ensure computational feasibility, balancing time complexity and true loss estimation quality.

The validation, differently from the original paper, was conducted by computing **full images predictions** (through the probability map), in order to estimate as precisely as possible the actual model's performance.

The purpose of the validation set was for **hyper-parameters tuning, adjustments in the model and training methods**, through the evaluation of prediction computed starting from a set of input images separated from training and testing sets, hence independent from the model itself (for an accurate and unbiased assessment of the model).

All cross-validated models went through the **computation of the metrics** of the 10-fold predictions, which then got averaged by the same subset of images. These 3-averaged metrics were then recorded, averaged all together and finally recorded, along with their standard deviation, for comparison.

The collected metrics are displayed in the following table.

- **Accuracy**: the amount of correct predictions over all predictions.
- **Precision***: the amount of correct positive predictions over all positive predictions.
- **Specificity***: the amount of correct negative predictions over all negative true labels.
- **Sensitivity** (recall): the amount of correct positive predictions over all positive true labels.
- **Intersection over Union** (IoU): a measure of the overlap between the positive true labels set and the positive prediction set.
- **F1 score**: harmonic mean between precision and recall.

These validated metrics, collected for each cross-validated model, have been leveraged for comparison to determine the best models.

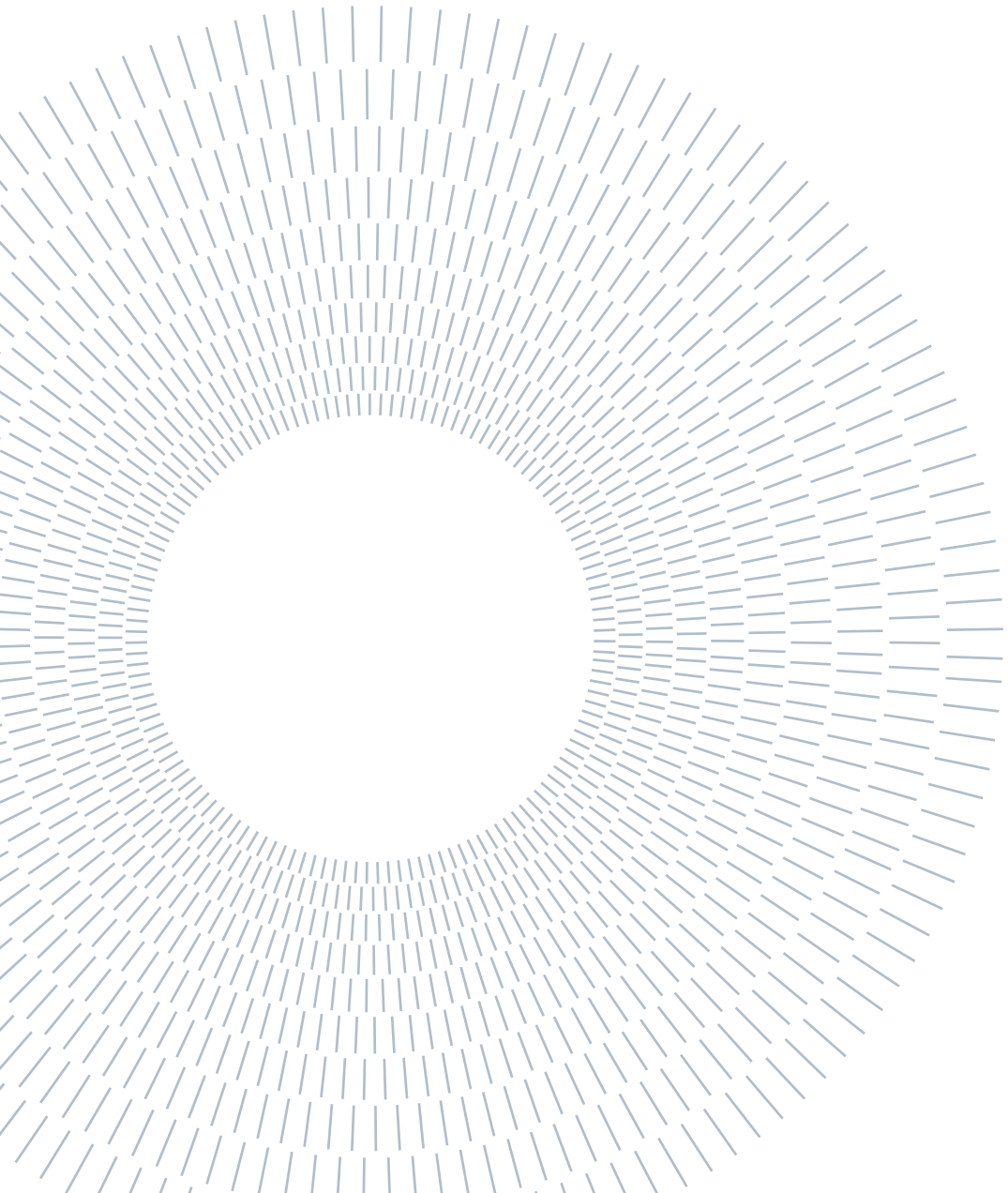
Remarks

- For full-tumor images, the metrics indicated with * were discarded because either they could not be computed or they were uninformative. The rest of their measures have been collected anyway.
- We notice some differences in performance of the different models in the context of full-tumor image predictions. For this reason, these images have also been subject of a separate evaluation for better assessment of the models.

3.5. Testing

We evaluated our models by performing the very procedure described in section 2.5.1 (analogously to the validation phase) on a testing image, separated from validation and training set. However, we could not replicate the same consensus methods used in the original paper because the procedure was quite different: after the neural network inference, two pathologist were tasked to edit the boundaries of the prediction. The authors then used two different consensus methods to compute the final results.

We didn't have access to professional pathologists and the authors shared only a single mask. We therefore evaluated our model using the mask provided. It's important to recognize that our evaluation methods differed from those outlined in the original paper, precluding direct comparison.



4 | Experimental Results

4.1. Evaluation Metrics

As already described in Section 3.4, we have collected several metric indicator to pick the best models among the cross-validated ones.

The two following tables display the averaged metrics resulted from cross-validation, complete with standard deviation, separating results from **overall assessment** and **full-tumor assessments**.

	A1	**A2**	**I1**	**I2**
Acc.	0.9125 ± 0.0044	0.9172 ± 0.0031	0.8966 ± 0.0058	0.9041 ± 0.0048
Prec.	0.9100 ± 0.0071	0.8948 ± 0.0064	0.9138 ± 0.0066	0.9122 ± 0.0069
Spec.	0.6084 ± 0.0163	0.5458 ± 0.0102	0.6549 ± 0.0156	0.6590 ± 0.0106
Sens.	0.9552 ± 0.0069	0.9822 ± 0.0015	0.9338 ± 0.0068	0.9449 ± 0.0042
IoU	0.8929 ± 0.0063	0.9019 ± 0.0043	0.8762 ± 0.0075	0.8854 ± 0.0062
F1	0.9414 ± 0.0037	0.9471 ± 0.0025	0.9320 ± 0.0044	0.9377 ± 0.0036

Table 4.1: Overall validation metrics

	A1	**A2**	**I1**	**I2**
Acc.	0.9852	0.9689	0.7975	0.8581
Sens.	0.9852	0.9689	0.7975	0.8581
IoU	0.9852	0.9689	0.7975	0.8581
F1	0.9925	0.9839	0.8864	0.9232

Table 4.2: Full-tumor validation metrics

4.1.1. Overall Performances

From a general point of view, all models show some unbalanced results: while overall having high scores, they show unimpressive results regarding specificity, and, consequently, also F1 score. We can infer that the models has a **high chance of scoring false positive**; in other words, the models are likely to classify as tumor some areas that are not, disregarding some of the non-tumor regions. The other indicators show fairly high results because of the imbalances of the

dataset: there's a great majority of tumor cells, so whenever the models mistakes a non-tumor nest for a tumor one, the overall accuracy is not significantly affected as non-tumor areas are in minority, but the indicator assessing the prediction quality over false classification is vastly worsened.

Despite the clear issue, it can be considered a minor problem in practice: our task is to develop techniques to highlight risks of tumor in pancreatic tissues, in order to show to experts who has authority on the subject to assess the actual health status of the patient and make informed decisions on how to operate. It is not our objective to develop a model able to pinpoint exactly, with absolute precision, the area of coverage of tumor and non-tumor cells. In other words, our models serves to ring a bell regarding the level of exposure to oncologic risk. Therefore, **it can be actually beneficial to have a system that is especially sensitive to tumors**, and is more susceptible to false positive, as it would be more prone to detect anomalies.

Besides that, we can also notice the sensibility and the F1 score are especially high, so **the models succeed in classifying positive true labels**, which is the indicator that carries more value in an a pathology detection problem such as ours.

4.1.2. Differences Between Architectures

Besides the different training settings, **AlexNet tends to have better overall metrics** while actually having worse scores regarding the balance of class prediction quality, likely witnessing an higher imbalance than its adversary.

At a first glance, indeed, Inception V3 scores higher for what concerns specificity, and, as a consequence, an F1 score as well. From this, we can understand that **Inceptions V3 addresses more effectively class imbalances**, underlining a greater robustness, while still achieving respectable results regarding other metrics indicator.

Unfortunately, from what we can see in the second table, **AlexNet handles better full-tumor predictions**, being it more biased towards tumors. From this, we can understand that Inception V3 still has not effectively grasped the characteristics that separate one class from the other, despite performing better than AlexNet from this point of view.

To sum up, **AlexNet achieves better overall results**, but **Inception V3 show more balanced indicators**, while still delivering **similar performances**.

4.1.3. Differences Between Training Settings

Training parameters affect the final results in a significant way:

- **AlexNet**: **regularisation does not seem to help** with the prediction quality, as it slightly worsen the model's robustness, improving what was good about it but deepening its weaknesses. From this, we are persuaded to believe that **AlexNet does not over-fit**, as it has been stated in the original paper.

- **Inception V3**: leveraged to selectively address the trickier-to-handle non-tumor predictions, **focal loss actually harms the model's performance** from many perspectives. The standard cross-entropy training is to be preferred.

4.2. Comparison with the Article

Our study of the problem and the solution addressed to solve it go in different directions with respect to the paper:

- AlexNet performs better than in the original article, achieving more robustness, especially less false positives, and does not over-fit.
- Inception performs much worse, as it does score the excellent metrics shown in the paper. It is also much sensitive to class prediction imbalances, but, in analogy, it does not over-fit.

4.3. Model Selection and Testing

In light of what has been said so far, we decided to select A1 and I2 out of the four validated, being them the ones delivering the most overall satisfying performances.

At this point, we performed testing on them, to conclude the model assessment phase on another unbiased sample. The testing sample is the 29^{th} image, the last of the dataset, never used for training nor validation purposes, so completely new and unseen to the model as well as to us, to prevent any type of bias of ours.

The following section illustrates the testing results and provides a comparison regarding the performance of the chosen models. The results of the models are always shown one next to each other, in order to ease visual comparisons.

4.3.1. Testing Metrics

The following table collects the evaluation metrics computed on the testing image n. 29 for both selected models.

	A1	**I2**
Acc.	0.9254	0.9292
Prec.	0.9215	0.9412
Spec.	0.5927	0.7072
Sens.	0.9947	0.9754
IoU	0.9170	0.9194
F1	0.9567	0.9580

Table 4.3: Selected models testing metrics

We also provide the confusion matrix in two versions:

- The first one shows the metrics w.r.t the number of classifications.
- The second one shows the metrics w.r.t the percentages of classifications.

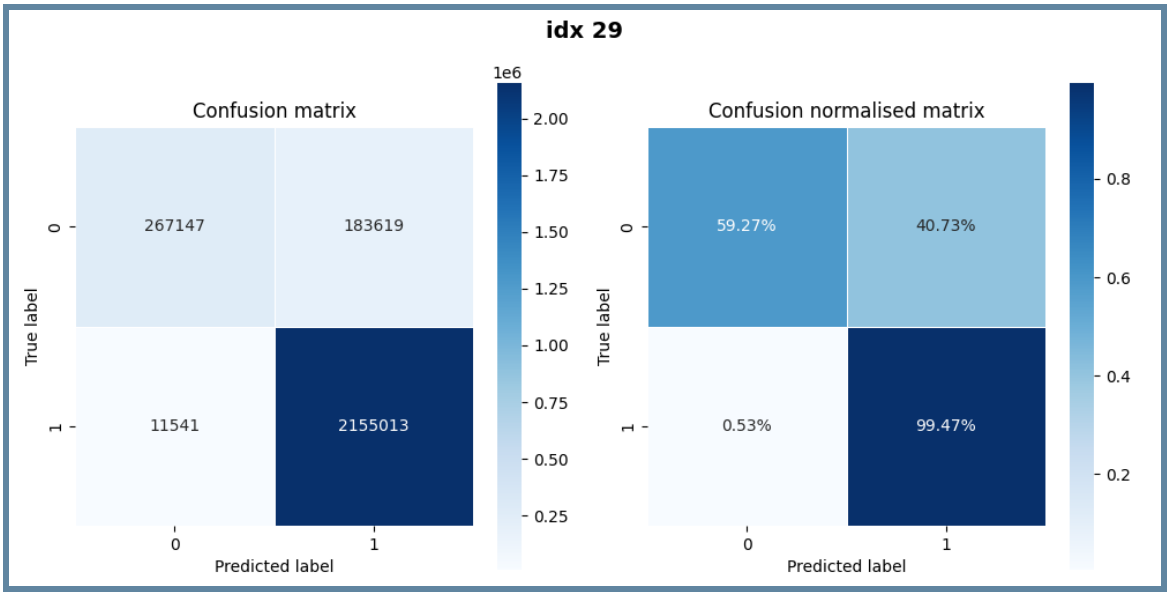


Figure 4.1: Testing confusion matrix of A1 predictions on image n. 29

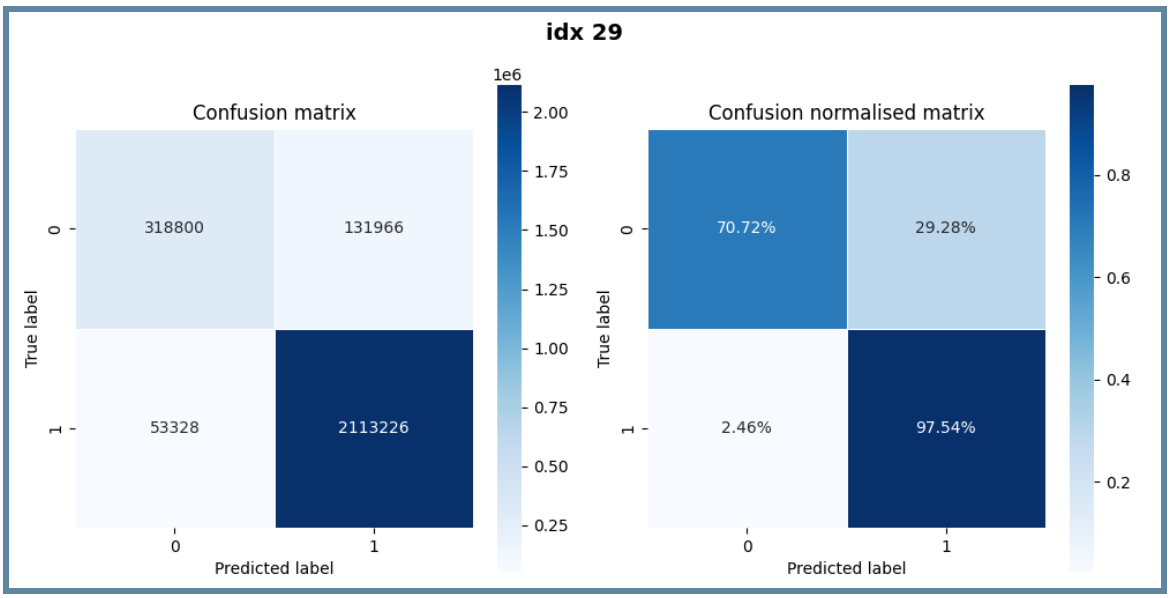


Figure 4.2: Testing confusion matrix of I2 prediction on image n. 29

As expected, in the context of this testing results, we can notice that the **two models behaved quite similarly** and that **I2 had a certain advantage in terms of prediction balance**. However, I2 performed slightly better, achieving marginally better scores in all indicators ex-

cept sensitivity. Overall, the two models proved to deliver respectable and sufficiently reliable performances.

4.3.2. Testing Visualisations

In this section, to conclude the testing procedure, we illustrate some of the visual artifact that allow us to judge by eye the behaviour of the two models.

The prediction procedure on the testing image performed by the models produced the following binary probability mask:

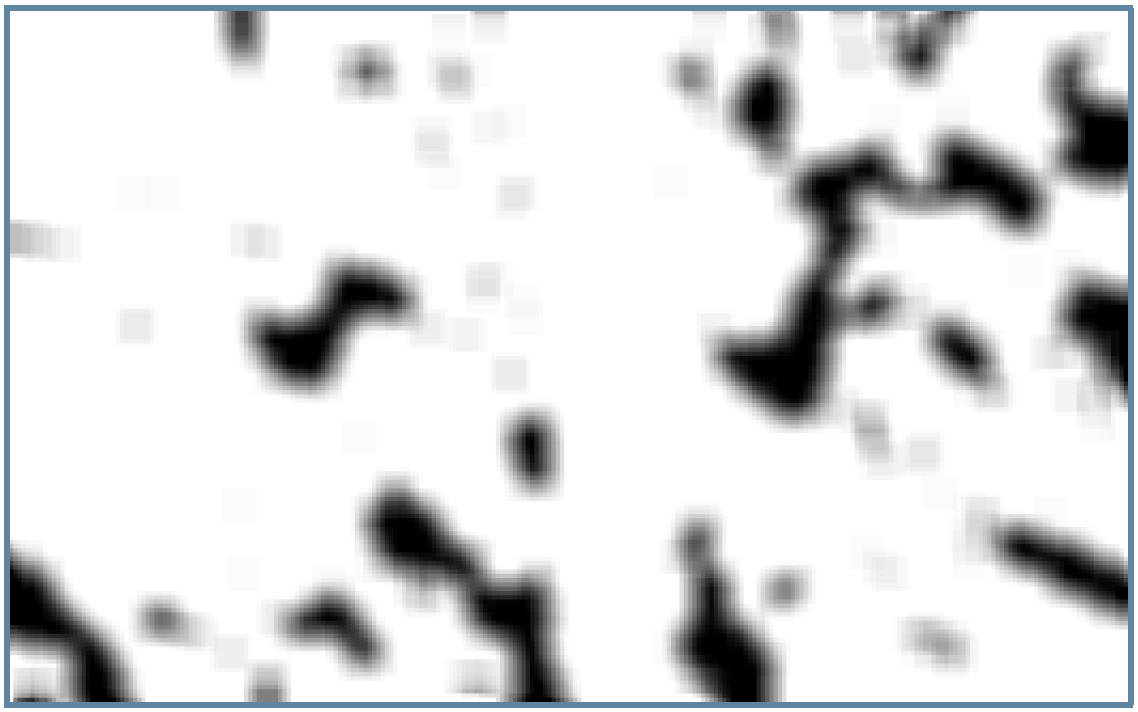


Figure 4.3: Testing prediction performed by A1 on image n. 29

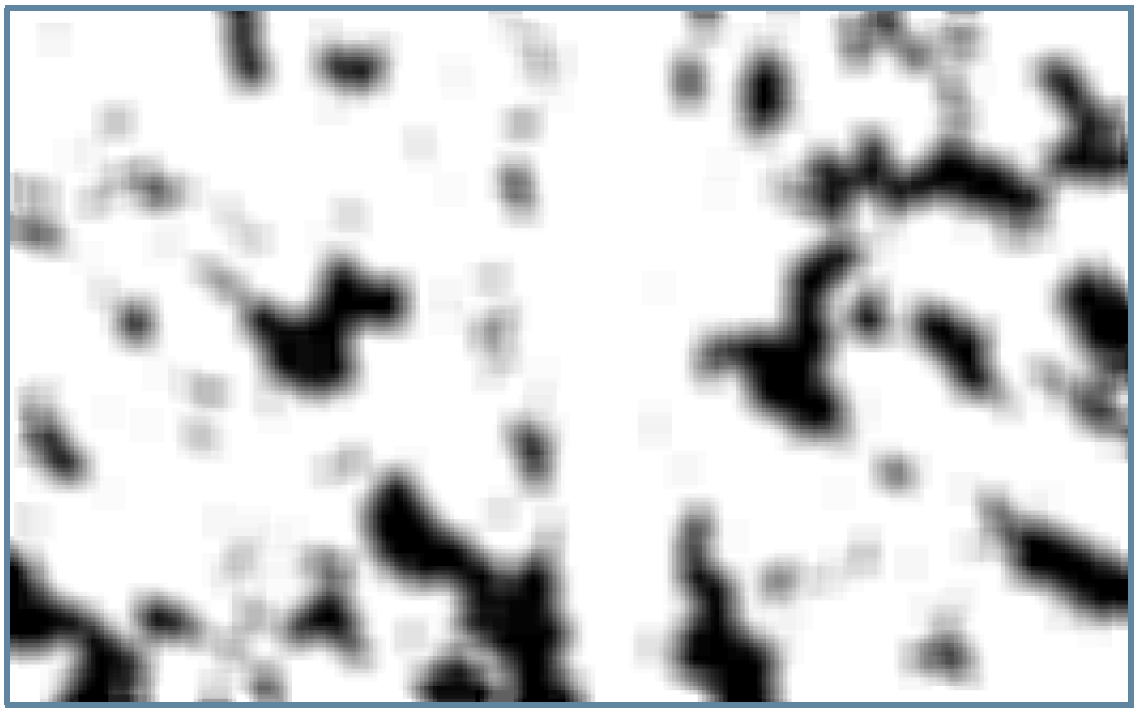


Figure 4.4: Testing prediction performed by I2 on image n. 29

The mask, filtered to a threshold of 0.5, delivers our segmentation prediction:

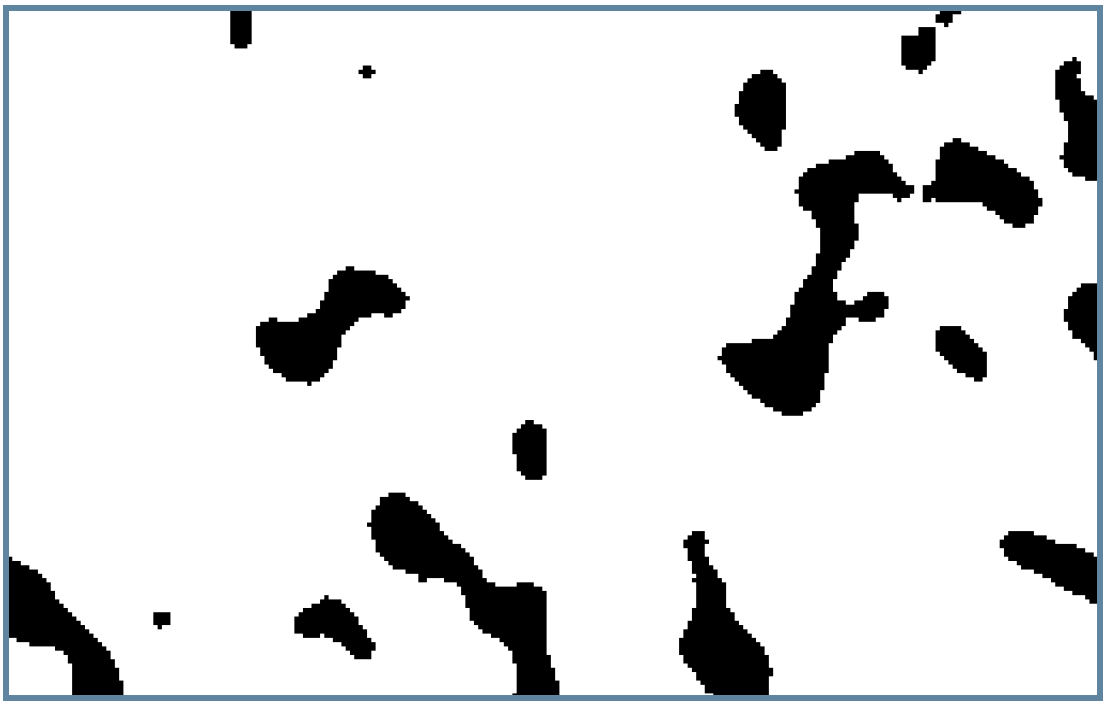


Figure 4.5: Testing thresholded prediction performed by A1 on image n. 29

Figure 4.6: Testing thresholded prediction performed by I2 on image n. 29

The following image shows a visual comparison between the predictions and the true segmentation mask available in the dataset, where:

- White pixels are correct positive classifications.
- Black pixels are correct negative classifications.
- Green pixels are false positive classifications (Type 1 error).
- Red pixels are false positive classifications (Type 2 error).

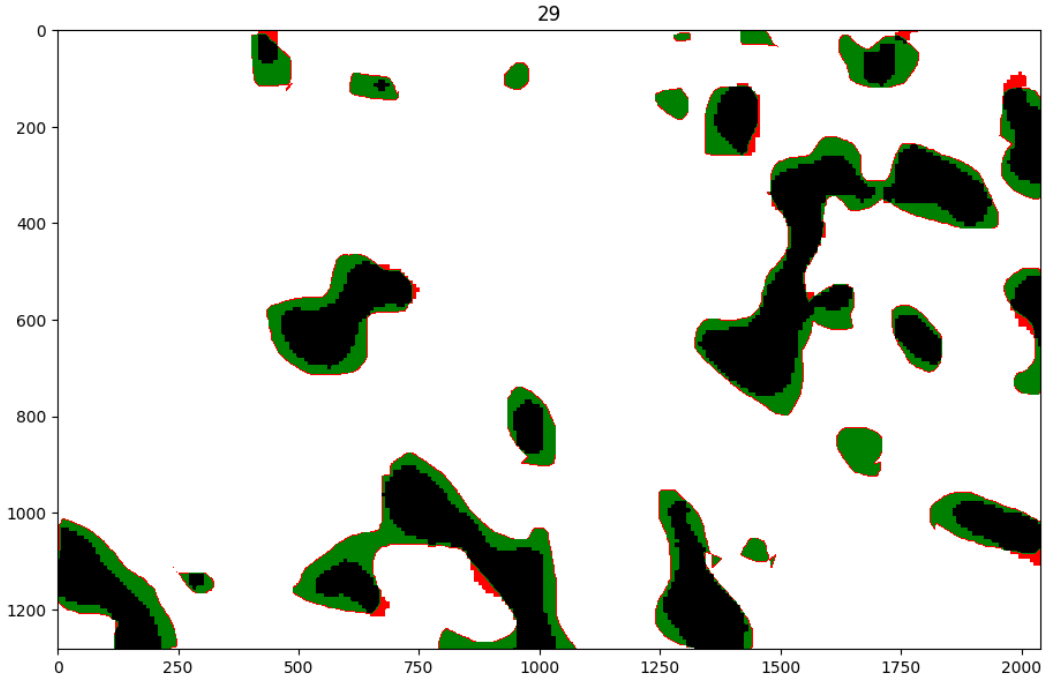


Figure 4.7: Visual comparison of prediction performed by A1 on image n. 29 with true mask

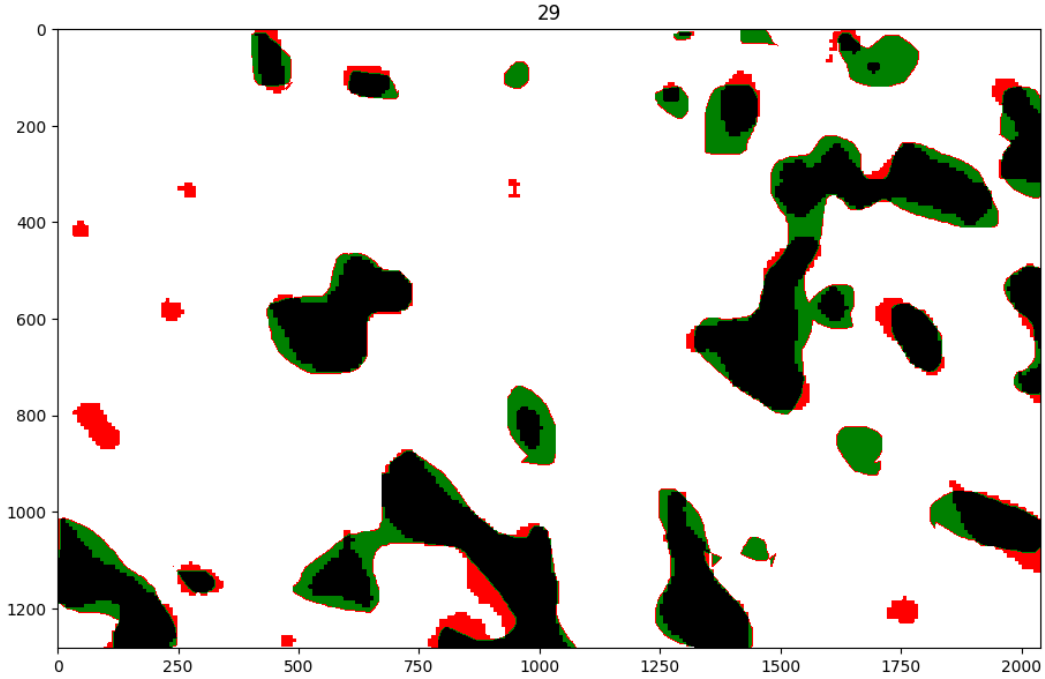
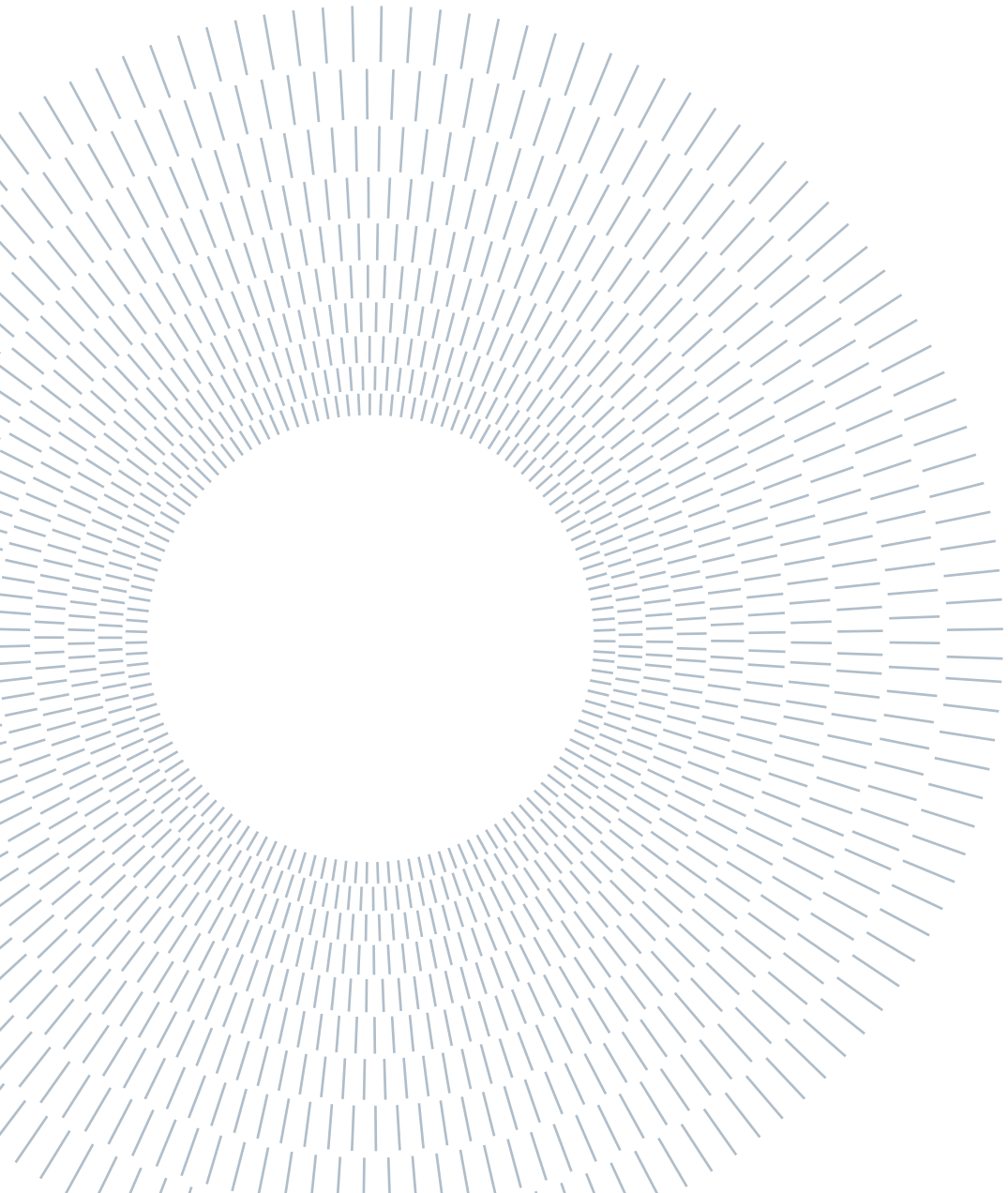


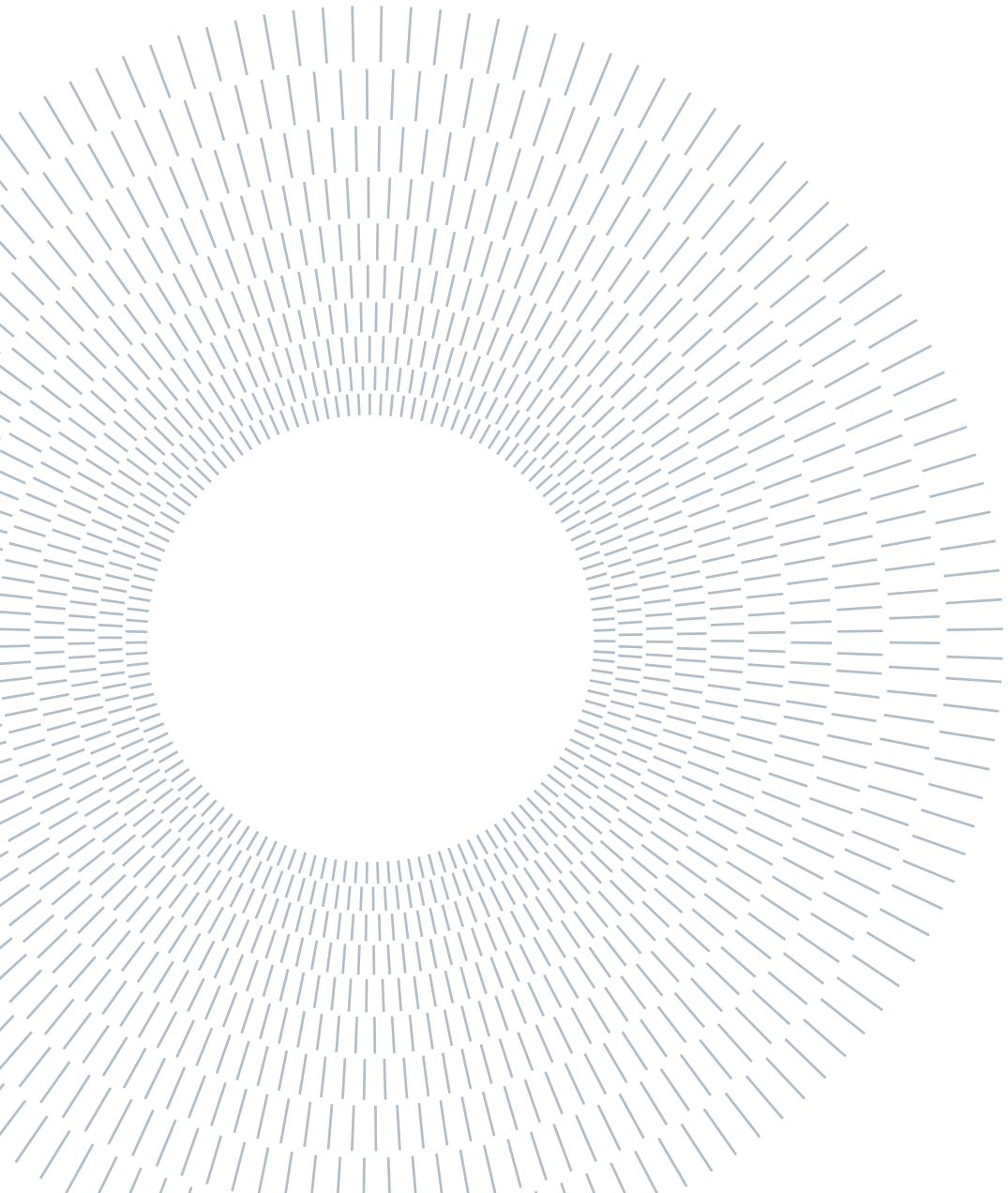
Figure 4.8: Visual comparison of prediction performed by I2 on image n. 29 with true mask

5 | Conclusions

In conclusion, our replication has been successful, despite the limits in the availability of the original data and the computational resources constrains. This shows that identification of tumor cells using machine learning with good accuracy is possible. Looking ahead, continued improvements in machine learning and data collection show great potential for enhancing accuracy in tumor classification.



List of Figures



List of Tables