Finding a Suitable Class Distribution for Building Histological Images Data Sets Used in Deep Model Training - the Case of Cancer Detection

Ismat Ara Reshma, Camille Franchet, Margot Gaspard, Radu Tudor Ionescu, Josiane Mothe, Sylvain Cussat-Blanc, Hervé Luga, Pierre Brousset

Journal of Digital Imaging manuscript No.
(will be inserted by the editor)

Finding a Suitable Class Distribution for Building Histological Images Data Sets Used in Deep Model Training - the Case of Cancer Detection

Ismat Ara Reshma[1,*] · **Camille Franchet**[2] · **Margot Gaspard**[2] · **Radu Tudor Ionescu**[3] · **Josiane Mothe**[1] · **Sylvain Cussat-Blanc**[1,4] · **Hervé Luga**[1] · **Pierre Brousset**[2,5,6]

Received: date

Abstract The class distribution of a training data set is an important factor which influences the performance of a deep learning-based system. Understanding the optimal class distribution is therefore crucial when building a new training set which may be costly to annotate. This is the case for histological images used in cancer diagnosis where image annotation requires domain experts. In this paper we tackle the problem of finding the optimal class distribution of a training set to be able to train an optimal model that detects cancer in histological images. We formulate several hypotheses which are then tested in scores of experiments with hundreds of trials. The experiments have been designed to account for both segmentation and classification frameworks with various class distributions in the training set, such as natural, balanced, over-represented cancer, and over-represented non-cancer. In the case of cancer detection, the experiments show several important results: (a) the natural class distribution produces more accurate results than the artificially generated balanced distribution, (b) the over-representation of non-cancer/negative classes (healthy tissue and/or background classes) compared to cancer/positive classes, reduces the number of samples which are falsely predicted as cancer (false positive), (c) the least expensive to annotate non-ROI (non-Region-of-Interest) data can be useful in compensating for the performance loss in the system due to a shortage of expensive to annotate ROI data, (d) the multi-label examples are more useful than the single-label ones to train a segmentation model, and (e) when the classification model is tuned with a balanced validation set, it is less affected than the segmentation model by the class distribution of the training set.

Keywords Computer-aided diagnosis · Medical information retrieval · Image segmentation and classification · Deep learning · Class-biased training · Class distribution analysis · Histological image

[1] IRIT, UMR5505 CNRS, Univ. de Toulouse, 118 Route de Narbonne, TOULOUSE F-31062 CEDEX 09, France
E-mail: Ismat-Ara.Reshma@irit.fr, Josiane.Mothe@irit.fr, Sylvain.Cussat-Blanc@irit.fr, Herve.Luga@irit.fr

[2] Department of Pathology, Univ. Cancer Institute of Toulouse-Oncopole, 1 avenue Irène Joliot-Curie, TOULOUSE F-31059, France E-mail: Franchet.Camille@iuct-oncopole.fr, Gaspard.m@chu-toulouse.fr, Brousset.p@chu-toulouse.fr

[3] Univ. of Bucharest, 14 Academiei, BUCHAREST 010014, Romania E-mail: raducu.ionescu@gmail.com

[4] Artificial and Natural Intelligence Toulouse Institute, Toulouse, France

[5] INSERM UMR 1037 Cancer Research Centre of Toulouse (CRCT), Univ. Toulouse III Paul-Sabatier, National Centre for Scientific Research (CNRS ERL 5294), Toulouse, France

[6] Laboratoire d'Excellence TOUCAN, Toulouse, France

[*] Corresponding author

1 Introduction

The huge success of deep learning models, such as convolutional neural networks (CNNs), in visual recognition [34, 37, 52] has encouraged researchers to explore their use in various domains, including cancer detection from histological images [5, 43, 62].

Histological images (also referred to as whole slide images or WSIs) are digitalized histological slides[1]. When working on a patient's case, the manual analysis of WSIs demands a high level of concentration and is time-consuming for pathologists. In this context, an automatic system can help by filtering out the healthy parts of the images and indicating possible (otherwise potentially overlooked) cancer regions. During the last few decades, many automatic systems based on machine learning techniques including deep learning [5,21, 26, 33, 42] have been put forward for cancer detection. However, these methods are mainly focused on end-to-end pipeline developments for cancer detection, while the success of such systems depends on several hyper-parameters [25]. We hypothesized that one of the most important hyper-parameters is the class distribution of the training set, as the training set provides the supervision for all learning-based systems [13,14]. Moreover, several studies have focused on class distribution analysis, in particular, the comparison between balanced and imbalanced distributions for different tasks, which illustrates its importance [4, 7, 51, 65]. Nonetheless, the results of different studies are inconsistent, depending on the tasks, data sets and the machine learning techniques employed [51]. This inconsistency casts doubt on their generalization for WSI data in the context of cancer detection, which is our topic of interest.

Generally, in machine learning an imbalanced data distribution has been shown to lead to inferior models compared to a balanced distribution [9, 30], and hence, a lot of effort has been put in to develop methods which overcome data imbalance. Buda et al. [7] have reviewed the popular methods, such as oversampling, undersampling, thresholding, cost sensitive learning, one-class classification and various hybrids. These studies may indeed lay the path for balanced distribution to become the default choice as a deep learning state-of-the-art method [5, 22, 43], although Prati et al. [51], for example, have shown that it is not optimal in all cases. Unfortunately, very few analytical studies on the performance impact of different distributions exist in the literature [51, 64, 65]. Moreover, the available studies have been mainly conducted on toy data sets, even though real data sets may be very different and more complex. Thus, there is no evidence from the conclusions of these studies that they would be appropriate for cancer WSIs.

Furthermore, the outcomes of available studies are contradictory: some support the natural distribution[2] [70], some support an imbalanced distribution [64, 65] while others support a balanced distribution [51]. For this reason, it is not straightforward to decide on a specific class distribution for all types of tasks. We believe that elaborate domain specific analysis is required for each specialized task, especially for a comparatively new and sensitive case, such as cancer detection from WSIs, which has not as yet been studied in any depth.

State-of-the-art methods for cancer detection [5, 22, 43] utilize different types of distribution (usually balanced) and obtain very convincing results for a large-scale breast cancer data set using different models. In particular, the existing systems achieve very high sensitivity[3] [43, 62], while false positives[4] remain an ongoing issue [50]. However, to our knowledge, no analysis indicates if the commonly-adopted balanced distribution is the most appropriate distribution for cancer detection in WSIs nor which class should be over-represented. It is also worth knowing which distribution produces fewer false positives with high sensitivity and why. It would help in choosing training examples and their ratios for building robust training data sets.

In this paper we present a data-driven analysis which determines the performance impact of different class distributions on training data. We derive several hypotheses with regard to WSIs used for cancer detection, and test them with two commonly used target applications: image segmentation and classification. The goal of image segmentation is to segment predefined objects from the input image [17, 48] for the purposes of localization. This is performed by labeling each pixel of the corresponding object with a predefined color that corresponds to a class label. Image classification, on the other hand, is the process of assigning a class label(s) to the entire input image. We test the hypotheses with deep learning state-of-the-art technology for both segmentation and classification tasks. We choose the default choice fully convolutional neural network (FCNN) and CNN architectures for the segmentation and classification task, respectively; finding the optimal architecture is out of the scope of this research. We employ two data sets for training: one is multi-class (annotated with cancer, non-cancer, and other histological structures), and is applicable to the segmentation setting, whereas the other is binary-class (annotated with cancer and non-cancer classes), and is applicable to the classification setting. With both data sets we conduct a series of experiments and analyze the results in detail to be able to provide comprehensive conclusions.

[1] A histological slide is a microscopic examination of tissue used by physicians to study the manifestations of disease.

[2] Natural distribution is the distribution a data originally has, which can be either balanced or biased to a certain class.

[3] Sensitivity is the proportion of actual positive cases that are predicted as positive.

[4] Negative example wrongly predicted as positive class

While the main focus of this paper concerns training data distribution, we also discuss the case of the test data distribution to see if the proposed hypotheses hold for the different test distributions as well. To this end, we use the test set for an additional data set along with the other data sets. However, the test data distribution is a large topic itself and will require a separate elaborate analysis

The rest of the paper is structured as follows. We first review the literature in Section 2. In Section 3 we present the materials and methodology for the relevant hypotheses we formulate on the impact of class distributions. In Section 4 we describe our experimental settings, while the results are presented and discussed in Section 5. In Section 6 we draw our conclusions and discuss future directions.

2 Literature Review

The natural imbalanced distribution of data is generally considered as a problem/obstacle in machine learning. Researchers put a lot of effort into balancing data. In contrast, comparative analytical studies of different distributions are few in number. In this section we investigate several studies related to class imbalance problems and class distribution analysis. We also present the state-of-the-art cancer detection in WSIs.

2.1 Class Imbalance Problems

One of the common problems in machine learning is dealing with class-biased or imbalanced data. In the real world, the availability of some classes makes them an over-represented majority, while the scarcity of other classes makes them an under-represented minority. This imbalance usually makes the classification task challenging for a classifier. There are many studies [1, 7, 9, 10, 27, 30, 35, 39, 45, 57, 63, 67, 68] showing that imbalanced training data leads to a loss in performance, so various methods have to be adopted to make the training data balanced. Prati et al. [51] list some of the most popular methods. Most studies are conducted on classical machine learning methods, and yet, only a few discuss the deep learning perspective [7,28]. Among the studies referring to this perspective, some suggest data-level modifications [1, 23, 27, 39], some prefer tweaking different hyper-parameters of the network or making algorithm-level modifications [3,30,63], e.g. incorporating a new cost function, while others suggest combining data and algorithm-level modifications [24, 53]. Buda et al. [7] performed a comparative study of different methods to address the class-biased problem. According to the authors, imbalanced data have an adversarial effect on the classification accuracy of the CNNs, similarly to classical machine learning techniques. The most recommended solution is to oversample the minority class [7, 45]. Johnson and Khoshgoftaar [28] conducted another detailed survey on recent techniques used to deal with the imbalanced data problem in deep learning. They concluded that not enough evidence exists to suggest that a particular technique is superior in dealing with class imbalance through the use of deep learning.

2.2 Class Distribution Analysis

All the articles from the previous subsection describe the adversarial effects of imbalanced data, and suggest methods for making the class distribution of the training data balanced. However, these methods do not provide the answer to an important question, namely, whether the balanced distribution is optimal for all types of learning techniques and data sets, and if we are to answer this question, comparative studies on different training class distributions will be required. Unfortunately, such studies are few in number, while hundreds are available on solving data imbalance, as mentioned in Section 2.1. Among the available comparative studies, Reshma et al. [54] present a short study on the impact of non-ROIs and natural distribution where the conclusion are in favor to non-ROIs biased data set. Weiss and Provost [64, 65] showed that neither a naturally occurring class distribution nor a balanced distribution is best for learning, and often a substantially better performance can be obtained by using a different class distribution. For their analysis they employed 26 data sets from the UCI repository [16]. According to the analysis by Prati et al. [51] on 20 data sets from UCI [16] and a small number of private data sets, the best distribution for seven different learning algorithms, including neural networks, is a balanced distribution. They conclude that only Support Vector Machines are less affected by class imbalance. However, this study was carried out on toy data sets, and hence there is no evidence that the commonly prescribed balanced distribution is a generalizable solution for more complex real data sets. Consequently, a separate analysis is required for each special kind of data. This conclusion is clearly connected to the No Free Lunch Theorem [66], which states that there is no single model that works best for every task. For example, Zhu et al. [70] conducted an elaborate domain specific study to optimize the training data distribution for land cover data in a bid to detect change. According to the authors, a class distribution proportional to the naturally occurring distribution is

superior to a balanced distribution for land cover data when using random forests as the learning algorithm. To our knowledge, there is no such comparative study for cancer detection in WSIs.

2.3 Studies on Cancer Detection in WSIs

In recent years, the use of deep CNNs has shown incredible performance levels with regard to cancer detection in WSIs [44]. However, the recommended systems have not been adopted at a clinical level, as the performance threshold needed to gain the trust of pathologists has not yet been met [6]. Therefore, developing efficient methods for cancer detection remains an active area of research, as there are still some hyper-parameters to tune in order to achieve a low-cost, high-performing system that could reduce false predictions and cognitive load [18]. Several deep learning systems have been proposed during the last few years, specifically from 2015 onward. Bejnordi et al. [5] have organised a worldwide challenge known as CAMELYON to gather together different methods for cancer detection in WSIs. Before this annual event, the use of WSIs in computational tasks was limited to patches which had been pre-extracted by the pathologist [40]. Most of the proposed methods in the CAMELYON16 challenge are based on deep learning: the variation in the participants' results is induced by hyper-parameter settings and data preprocessing. The winning team [62] trained two 22-layer GoogleNets (V1), one with randomly sampled training patches–probably biased towards negative examples– and another with additional hard negative examples. The final decision was based on the combination of predictions from two models. Additionally, they trained a random forest classifier with 28 handcrafted features extracted from the output heatmaps of the CNNs. They used color normalization to cope with color variation in the WSIs while also applying rotation, and adding extra color noise to training patches for data augmentation during training. Liu et al. [43] (from GoogleAI) used the GoogleNet (V3) [58]. They applied a random patch sampling technique to obtain a balanced training set. Moreover, to deal with the scarcity of tumor patches, they applied several data augmentation techniques, including rotation, mirroring and extensive color disturbance. However, in their extended work [44], they selected a training distribution biased towards the negative class by a factor of four. Lin et al. [40] proposed a framework for fast and dense scanning during prediction. In their framework they converted the modified VGG16 to a fully convolutional neural network (FCNN), which was followed by a patch reconstruction method. In their proposed system they used false positive generated patches in training. Veeling et al. [59] proposed a rotation equivariance framework by adopting G-CNN architecture [12]. To test the effectiveness of their method, they proposed a new data set, known as PatchCamelyon (PCam), which gives a balanced distribution. In [46, 47], authors proposed a new data set for different types of breast cancer, and an end-to-end deep learning framework for multilabel tissue segmentation utilizing their data set while network parameters were determined with a deep analysis. Other cancer detection studies have been covered in [6, 69], and a number of end-to-end pipelines were developed in these studies, but training class distribution was not considered. Furthermore, filtering out non-ROIs and creating a training data set with artificially balanced or a slightly skewed distribution towards the negative class has been a common practice. In other words, the usual natural distribution and non-ROIs have yet to be explored. In this study we investigate the usability of non-ROIs and natural distribution for the WSI data. Moreover, we compare three other commonly adopted distributions, namely balanced, cancer-biased and non-cancer-biased.

3 Materials and Methods

According to Prati et al. [51], the impact of different class distributions on various data sets is not always the same. An accurate study is therefore required to find the optimal class distribution for any crucial task, such as cancer detection from WSIs.

The commonly preferred balanced distribution of the training set [5, 22, 43, 60] is not necessarily the best distribution for cancer detection without carrying out a proper comparative study. Based on this observation, we conducted several preliminary experiments for cancer detection with balanced and imbalanced (both naturally and artificially generated) training sets. From our preliminary experiments we found that the artificially generated balanced set does not provide the best performance. The results encouraged us to draw certain hypotheses and investigate the results at length. These hypotheses led to the design of a new generic framework that can be applied to any new task to provide an optimal choice of class distribution in the training set for a machine learning model. Figure 1 illustrates the methodological framework for this analytical study. However, before describing the proposed hypotheses and corresponding framework, we discuss the main elements of this domain-specific study, i.e. the WSIs and their corresponding details, to be able to familiarize the reader with the general terms and notations.

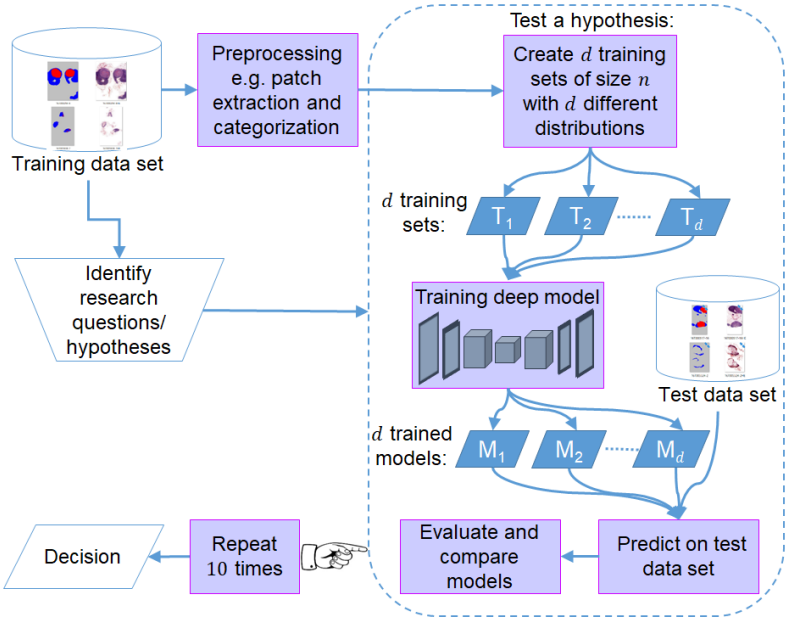


Fig. 1: Our methodological framework. To answer the general question of what the optimal class distribution of the training set should be, we created d training sets with n patches and d different distributions. We then trained d models corresponding to each distribution that we tested on the same unseen test set. The training data distribution of the best model for test data was considered as the best distribution. We repeated each experiment 10 times and calculated the mean to be able to make the final decision.

3.1 Whole Slide Images (WSIs)

We use whole slide images (WSIs) as the image format for this study. WSIs are the digital conversion of conventional histological glass slides containing tissue sample [19,36]. Unlike a normal image, a WSI is stored in a multi-resolution pyramid structured file. The file contains multiple versions of downsampled images, and usually comprises 8 to 10 levels of resolution, including the original level. This allows the extraction of the WSI image at any resolution, and takes low to high magnification levels into account. Level 0 is considered to be the highest magnification level, and at this level a WSI could be several gigapixels in size.

The examination of lymph node sections, either on glass slides or on WSIs, is common practice for the evaluation of tumors and their spread. However, this evaluation has some well-known limitations, such as being time-consuming, laborious and dependent upon the pathologist's expertise and level of fatigue. In this regard, researchers have endeavored to find computer-aided diagnosis (CAD) systems which can act as a tool to assist pathologists. Machine learning techniques, in particular, are used. While applying machine learning, two regions in a WSI are generally recognized:

- Regions of interest (ROIs) are the regions that alert pathologists to check for abnormalities. In the case of lymph node WSIs, the regions containing lymph nodes are ROIs, since the health of lymph node tissue is what pathologists observe. The ROIs can in turn be divided into two classes: positive and negative. Metastasis is considered as a positive class, *cancer* (denoted by $\mathbb{C}$ in the rest of the paper), while any remaining ROIs are considered as a negative class, *non-cancer* (denoted by $\neg\mathbb{C}$).
- Other regions (non-ROIs) are mainly background and histological structures other than lymph node tissue. The non-ROIs, that is to say, non-lymph nodes, are considered as belonging to the negative class, *other* (denoted as $\mathbb{O}$).

In other words, three classes are usually considered in a WSI: the positive ROI class, *cancer* ($\mathbb{C}$), the negative ROI class, *non-cancer* ($\neg\mathbb{C}$), and the negative non-ROI class, *other* ($\mathbb{O}$). Since the cancer class ($\mathbb{C}$) is of the utmost importance to the binary classification task, both $\mathbb{O}$ and $\neg\mathbb{C}$ are merged into one class and simply considered as being the negative class $\neg\mathbb{C}$.

Figure 2 shows an example of a metastatic lymph node WSI with its corresponding regions. In a WSI data set, the non-ROI class is usually over-represented, while the two ROI classes could be balanced or imbalanced, depending on which WSIs have been included in the data set (more details are provided in Section 4.1). On average, a WSI contains 70 to 80% of non-ROI pixels

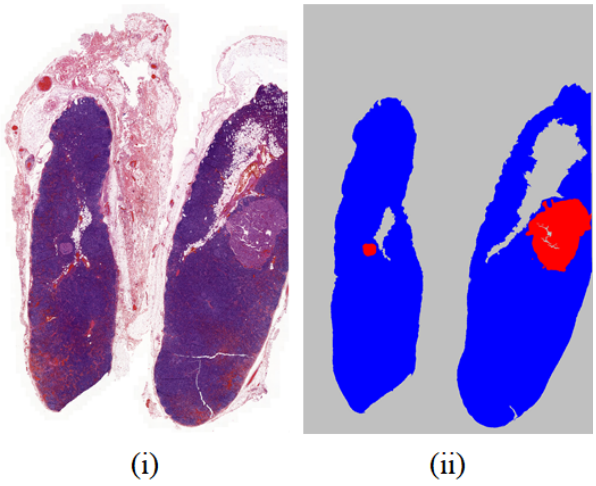


Fig. 2: (i) Metastatic lymph node WSI and (ii) its annotation. In (ii), the colour red represents the cancer class ($\mathbb{C}$), blue represents the non-cancer class ($\neg\mathbb{C}$), and gray represents *other* class, mainly the background ($\mathbb{O}$). Both $\mathbb{C}$ and $\neg\mathbb{C}$ are ROIs, while $\mathbb{O}$ is not.

[41, 62]. We consider this usual bias towards non-ROIs as the *natural* distribution.

3.2 Patch Categories in WSI Data

Due to typical memory limitations, rather than use the whole images as training examples, it is common practice to use patches extracted from the WSIs. We hence consider these patches for training and testing purposes. Furthermore, the patches help create a different distribution in the training set for our analysis. However, in our research the patches need to be placed in different categories. Thus, we consider different categories of patches based on the annotation of the WSIs in the data set.

For a *multi-class WSI data set* annotated using three different classes as shown in Figure 2, we consider four categories of patches as defined in Table 1. Patches that contain more than 99.999% class $\mathbb{O}$ pixels are labeled as the *other* ($\mathbb{O}$) category. The remaining patches belong to the other categories: *cancer* ($\mathbb{C}$), *non-cancer* ($\neg\mathbb{C}$), or *mixed* ($\mathbb{C}\&\neg\mathbb{C}$) with an optional presence of class $\mathbb{O}$ pixels ($< 99.999\%$). For example, in a patch with 100% $\mathbb{C}$ pixels, there are no $\mathbb{O}$ pixels. Among the four patch categories, we consider the category $\mathbb{C}\&\neg\mathbb{C}$ as being multi-label (since it contains two ROI classes at the same time), and the others as single-label. Note that we use the same notations for pixel classes.

For the *binary class WSI data set*, i.e. the data set where only class $\mathbb{C}$ is marked in the ground truth, we extract the patches differently. First of all, we separate the regions without histological structures, i.e. background, using Otsu threshholding [49]; then we extract $\mathbb{C}$ and $\neg\mathbb{C}$ category patches from the corresponding regions. Fortunately, the extraction and categorization of patches has been performed by Veeling et al. [60], giving us the benchmark data set known as PatchCamelyon (PCam) (details are available in Section 4.1.2). We can therefore use the PCam data set as it is. However, we follow the same patch extraction and categorization approach for the other binary class data set, namely CAMELYON16 [2].

We use the categories described above to design several experiments with the aim of testing various hypotheses concerning the class distribution of the training set, which are described in Section 3.3.

3.3 Hypotheses and Relevant Class Distributions

We make several hypotheses and design several experiments with the relevant class distributions to be able to test the proposed hypotheses.

The first hypothesis (H1) revisits the superiority of balanced distribution when training a model, while the other three hypotheses consider the impact of one balanced and two other class-biased distributions (H2 to H4) in different situations.

While testing one hypothesis, the total number of patches in the training set of each experiment is kept the same to ensure fair comparison. The experiments within a group of experiments differ from one another according to the patch categories that make up the training set, and according to their ratio, i.e. the distribution of the classes in the training set. The hypotheses are presented below.

H1: Natural distribution is not an optimal method for training a model.

As mentioned in Section 3.1, the WSI data are naturally biased towards the non-ROI class, $\mathbb{O}$. In other words, the natural distribution of the WSIs is the over-representation of class $\mathbb{O}$. However, class $\mathbb{O}$ is not a region of interest for pathologists. In this case it is a common practice to filter out the excessive examples of class $\mathbb{O}$ and make a balanced training set out of the remaining classes.

We hypothesize that the trained model with over-represented class $\mathbb{O}$ (i.e. natural distribution) will be effective at detecting the regions that are not of interest (non-ROIs) to pathologists, since it will be trained using a large number of different non-ROI cases. However, we also hypothesize that this distribution will be less effective for the less frequently occurring ROI cases, although these cases are much more interesting to detect for pathologists.

Table 1: **Definition** of Patch Categories

Data set type	Patch *category*	Definition	Label type	Comment
Multi-class	*Other* ($\mathbb{O}$)	Containing > 99.999% $\mathbb{O}$ pixels	Single-label	Non-ROI/non-lymph node as the negative non-ROI class
	Cancer ($\mathbb{C}$)	Containing > 0.001% $\mathbb{C}$ pixels and no $\neg\mathbb{C}$ pixels.	Single-label	Metastasis as the positive ROI class
	Non-cancer ($\neg\mathbb{C}$)	Containing > 0.001% $\neg\mathbb{C}$ pixels and no $\mathbb{C}$ pixels.	Single-label	Lymph node without metastasis as the negative ROI class
	Mixed ($\mathbb{C}\&\neg\mathbb{C}$)	Containing > 0.001% $\mathbb{C}$ pixels and > 0.001% $\neg\mathbb{C}$ pixels.	Multi-label	Belongs to both $\mathbb{C}$ and $\neg\mathbb{C}$ class at the same time
Binary class	*Cancer* ($\mathbb{C}$)	Containing $\mathbb{C}$ pixels at a particular center region of the patch.	Single-label	Metastasis as the positive class
	Non-cancer ($\neg\mathbb{C}$)	Containing $\neg\mathbb{C}$ or $\mathbb{O}$ pixels and no $\mathbb{C}$ pixels at the central region.	Single-label	Lymph node or non-lymph node (unlike multi-class data set case) without metastasis as the negative class

Table 2: **Experiment settings E1**: E1 settings are designed to test H1 (natural distribution is not optimal) with a total of 9 units ($\mathbb{U}$) of patches in the training set of each experiment.

Experiment ID	Distribution	Patch ratio ($\mathbb{O} : \mathbb{C} : \neg\mathbb{C}$)
E1.a	Balanced	3 : 3 : 3
E1.b	Over-represented $\mathbb{O}$ (natural)	7 : 1 : 1

To test H1, we designed two experiments: E1.a and E1.b, i.e. in Figure 1, $d = 2$. In E1.a we consider the same number of patches in each of the three classes, whereas in E1.b the training examples are highly biased (7 times) towards class $\mathbb{O}$ (similar to the natural distribution) as presented in Table 2. We hereby introduce $\mathbb{U}$ to denote a unit (fixed amount) of patches. For a particular data set, the size of $\mathbb{U}$ is determined by the size of the smallest patch category (the category with the least number of patches) and its largest presence in the expected patch ratios. The size of $\mathbb{U}$ is determined so that we can consider both the under and over-representation of that smallest category for the given data set. In this regard, we consider only single-label patch categories. For example, let us suppose that $\mathbb{C}$ is the smallest category with size N (total number of extracted patches), and in the expected patch ratios given in Table 2, its largest presence is 3, then $\mathbb{U} = N/3$. To test H1, a total of $9\mathbb{U}$ of patches is used to create both the natural and balanced distributions, i.e. in Figure 1, $n = 9\mathbb{U}$.

This hypothesis can be tested on the multi-class data set only. Since class $\mathbb{O}$ is not annotated separately in the binary class data sets, it is not possible to accurately separate class $\mathbb{O}$ from class $\neg\mathbb{C}$, especially if histological structures are available in $\mathbb{O}$ regions.

H2: Over-representing the $\neg\mathbb{C}$ class in the training set reduces false positives during cancer detection.

We observed that in WSIs the regions for class $\neg\mathbb{C}$ are more heterogeneous than the regions for class $\mathbb{C}$. Indeed, class $\neg\mathbb{C}$ regions may contain germinal centers, macrophage, blood vessels or artifacts (e.g. blur areas), etc. (see Figure 3). Here, some of these are visually similar to class $\mathbb{C}$ regions. Diversity of the $\neg\mathbb{C}$ regions (both in training and test sets) would almost certainly increase in more complex anatomic structures than lymph nodes. To ensure the coverage of different variations and to reduce any confusion with class $\mathbb{C}$ regions, the over-representation of class $\neg\mathbb{C}$ could be useful. Thus, over-representing class $\neg\mathbb{C}$ compared to class $\mathbb{C}$ should reduce the false positive rate for class $\mathbb{C}$.

Table 3: **Experiment settings E2**: E2 settings are designed to test H2 (i.e. $\neg\mathbb{C}$-biased training produces less false positives) with single-label patches. There is a total of $4\mathbb{U}$ patches in the training set of each experiment.

Experiment ID	Distribution	Patch ratio ($\mathbb{C} : \neg\mathbb{C}$)
E2.a	Balanced	2 : 2
E2.b	Over-represented $\neg\mathbb{C}$	1 : 3
E2.c	Over-represented $\mathbb{C}$	3 : 1

To test this hypothesis, we designed three experiments (i.e. in Figure 1, $d = 3$), as presented in Table 3. In E2.a we consider the balanced case between $\mathbb{C}$ and $\neg\mathbb{C}$, while E2.b over-represents $\neg\mathbb{C}$ and E2.c over-represents $\mathbb{C}$ respectively. In the literature, the ratio of under-representation to over-representation is around 1:3 [5, 44], so we follow the same ratio in our experi-

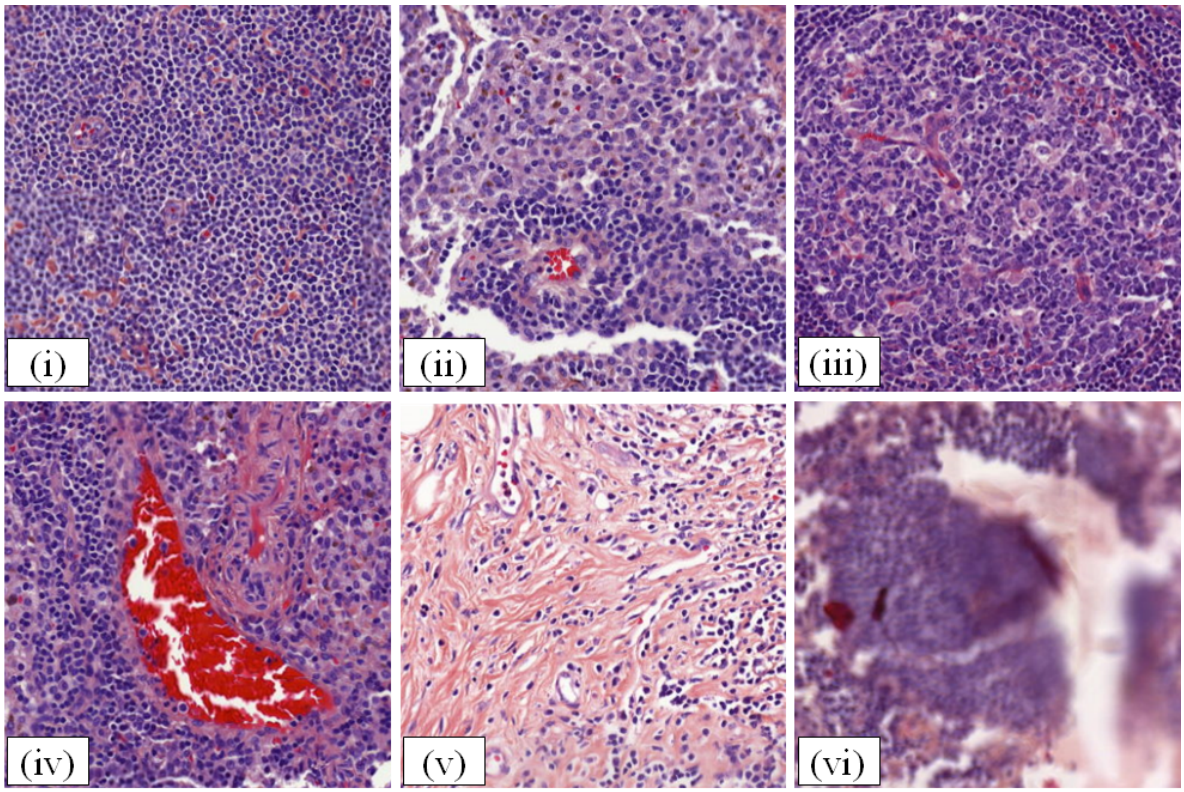


Fig. 3: Samples of heterogeneous patches of class $\neg\mathbb{C}$ (lymph node): (i) usual lymphocytes in lymph nodes, (ii) sinusal macrophages in lymph nodes, (iii) germinal center, (iv) blood vessel, (v) fibrosis with lymphocyte, and (vi) artifact (blur area). Here, (ii) and (iii) have some common visual features with class $\mathbb{C}$; on the other hand, (iv), (v) and (vi) are non-specific structures which might also appear in the other two classes, $\mathbb{C}$ and $\mathbb{O}$.

mental design. The total number of patches in the experiments used to test this hypothesis (H2) is lower than that of H1, mainly because of the expected ratio (1:3) and the total number of extracted patches for the less frequently occurring ROI classes ($\mathbb{C}$ and $\neg\mathbb{C}$). Thus, the results of experiments E1 and E2 are not fully comparable.

This hypothesis can be tested on both multi-class and binary-class data sets. For the multi-class data set, we do not consider patches from category $\mathbb{O}$ – it is important to recall that there are many $\mathbb{O}$ pixels in ROI patches – to be able to focus on the ROI cases, similarly to the common trends in the literature of filtering out non-ROIs. Although we do not use the patch categories $\mathbb{C}\&\neg\mathbb{C}$ and $\mathbb{O}$ in the experiment setting given in Table 3, we use them for H3 and H4 and for re-testing H2. Further details are provided in the following paragraphs.

H3: Multi-label examples are more useful than single-label examples as training data.

This hypothesis states that the patches containing both ROI classes $\mathbb{C}$ and $\neg\mathbb{C}$ (i.e. belonging to category $\mathbb{C}\&\neg\mathbb{C}$) add more valuable information during training than the patches containing a single ROI class (i.e. categories $\mathbb{C}$ or $\neg\mathbb{C}$). Indeed, by having both classes of pixels at the same time, multi-label patches include boundary information for the two classes, and hence there is more contextual information that could be helpful during training, especially for segmentation models, which can localize multiple classes within the same patch. In other words, we consider having multiple ROI classes in the same patch as advantageous.

Table 4: **Experiment settings E3**: E3 settings are designed to re-test H2 when multi-label patches are used and to test H3 (i.e. multi-label patches are more useful than single-label patches).

Experiment ID	Distribution	Patch ratio ($\mathbb{C} : \neg\mathbb{C} : \mathbb{C}\&\neg\mathbb{C}$)
E3.a	Balanced	1.5 : 1.5 : 1
E3.b	Over-represented $\neg\mathbb{C}$	0 : 3 : 1
E3.c	Over-represented $\mathbb{C}$	3 : 0 : 1

To test this hypothesis, we had to design a trio of experiments analogous to E2 (which are designed with single-label patches). To construct class-biased distributions, we replaced the under-represented class examples (category $\mathbb{C}$ or $\neg\mathbb{C}$) in E2.b and E2.c with examples of multi-label patches (category $\mathbb{C}\&\neg\mathbb{C}$). At the same time, to re-evaluate H2 in the current case, we designed the corresponding E2.a experiment as well. The detailed design of the experiments is given in Table 4. First, in E3.a, we considered a balanced case between $\mathbb{C}$ and $\neg\mathbb{C}$. Then, similarly to E2, in E3.b and E3.c we considered over-represented $\neg\mathbb{C}$ and over-represented $\mathbb{C}$ cases respectively. The total number of patches in E3.* is the same as in E2.*, namely $4\mathbb{U}$. Hence, the result of an E2.* experiment is comparable with the result of the corresponding E3.* experiment, where * can be replaced by a, b or c.

In short, we re-tested H2 with the current E3 experiment setting and tested H3 by comparing the corresponding pairs (E3.*, E2.*).

H4: Non-ROI data are useful for training.

Training CNNs requires very large data sets; sometimes data augmentation is applied to make a data set artificially larger [61]. Although data augmentation is useful in increasing the generalization power of CNN models, it can cause overfitting in the case of small data sets [56]. Obtaining a non-artificially augmented large data set of WSIs is generally expensive, since annotating ROI data requires domain experts. However, in our particular case, annotating non-ROIs is less expensive. Furthermore, inside the ROIs there are some common non-ROI regions. Non-ROI examples are required to learn how to classify these common non-ROIs inside the ROIs. By considering all these reasons, we hypothesize that non-ROI data could be useful for training a CNN, especially when the annotated ROI data set is small.

To test this hypothesis, we designed a trio of experiments analogous to E3. Specifically, we replaced some examples of class $\mathbb{C}$ and $\neg\mathbb{C}$ from E3.* with class $\mathbb{O}$ examples (see Table 5) to simulate the shortage of ROI data. However, we still kept the balanced, $\neg\mathbb{C}$-biased and $\mathbb{C}$-biased distribution in the corresponding experiment for ROI classes. Hence, this represents another setting for testing H2 in which, along with class $\neg\mathbb{C}$, we consider another negative class (non-ROI class, $\mathbb{O}$). We are thus revisiting H2 as *over-representing negative classes (both $\neg\mathbb{C}$ and $\mathbb{O}$) compared to the positive class ($\mathbb{C}$) reduces the false positives.*

We designed three experiments denoted as E4.*, as shown in Table 5, analogous to the E3.* experiments. Similarly to the E3 experiment settings, E4 serves two purposes. The first purpose is to test H4 by comparing E4 with E3; the second is to re-test H2 with the current E4 settings. Here, we design E4.a by replacing the same number of patches for categories $\mathbb{C}$ and $\neg\mathbb{C}$ with patches for category $\mathbb{O}$ from E3.a, thereby adding some extra non-ROI information while keeping the balance between ROI classes. We designed the following two experiments, E4.b and E4.c, as the corresponding pair for E3.b and E3.c by replacing $1\mathbb{U}$ patches from the over-represented ROI class with patches of category $\mathbb{O}$. As we kept the total number of training examples $4\mathbb{U}$, the E4.* experiments are comparable with the previous E2.* and E3.* experiments.

Table 5: **Experiment settings E4**: E4 settings are designed to re-test H2 where non-ROI patches are used, and to test H4 (i.e. non-ROI data are useful).

Experiment ID	Distribution	Patch ratio ($\mathbb{O} : \mathbb{C} : \neg\mathbb{C} : \mathbb{C}\&\neg\mathbb{C}$)
E4.a	Balanced	1 : 1 : 1 : 1
E4.b	Over-represented $\neg\mathbb{C}$	1 : 0 : 2 : 1
E4.c	Over-represented $\mathbb{C}$	1 : 2 : 0 : 1

In summary, the E1 experiment settings test H1 on the impact of a natural, balanced distribution, while the E2, E3 and E4 experiment settings test H2 on the impact of balanced, class-biased distributions in three different cases. Moreover, the comparison between E2 and E3 tests the H3 on the impact of multi-label patches. On the other hand, the comparison between E3 and E4 tests H4 on the usability of non-ROI patches.

Note that H2 is the only hypothesis that can be tested on both multi-class and binary-class data sets: the other three are based on multi-class data sets only.

3.4 Processing Pipeline

After presenting the case study on data and formulating the hypotheses, other major steps are preprocessing, training, inference and evaluation. All these steps are applicable for both segmentation at the pixel level and classification at the patch level settings: we mention some minor differences for each step.

At the preprocessing step, we extract overlapping patches at a particular stride and corresponding ground-truth maps/labels from the annotated training data. Instead of the often used random sampling [5, 43, 62] of patches, we extract overlapping patches by maintaining a particular stride, as empirically they perform better than with random extraction. Furthermore, since there is a lack of sufficient contextual information from the border pixels of a patch [29], it is preferable to learn

from overlapping patches uniformly sampled from all regions of a WSI.

For the patch-based data sets, the described preprocessing step is not required.

At the training step, we generate different class distributions in the training set, as described in Section 3.3. The patches from a particular category are selected randomly from all available patches for that category. The random patch selection is followed by shuffling the whole training set to prevent having all the patches in a mini-batch from the same category. This makes the convergence faster during training and provides greater accuracy [32]. The generated training set is used to train a fully convolutional neural network (FCNN) model for the segmentation task (i.e. pixel classification) or a CNN model for the classification task (i.e. patch classification).

Let $T_d = \{(I_k, g_k) : 1 \leq k \leq n\}$ be a training set of size n for a particular distribution, where $I_k \in \mathbb{R}^{l \times l \times 3}$ is a patch of dimension $l \times l$ with its corresponding ground-truth g_k. Here, $g_k \in \mathbb{R}^{l \times l \times C}$ for the segmentation task and $g_k \in \mathbb{R}^C$ for the classification task, and C is the number of classes in the ground-truth annotation (not to be confused with the cancer class $\mathbb{C}$). The training is a process of finding a classifier function $M_d : \mathbb{R}^{l \times l \times 3} \rightarrow \mathbb{R}^{l \times l \times C}$ (for segmentation) or $M_d : \mathbb{R}^{l \times l \times 3} \rightarrow \mathbb{R}^C$ (for classification) by minimizing a loss function $\mathcal{L}(g_k, M_d(\Theta, I_k))$, where Θ is the set of parameters of the classifier.

During inference, the trained model is employed to predict either pixels (in the segmentation task) or patches (in the classification task) for unseen test WSIs. For this purpose we extract same-sized overlapping patches from the WSIs if the patches have not already been extracted in the test set. When the non-ROIs are annotated in the data set (i.e. for the multi-class data set), we predict the patches from WSIs. Otherwise (i.e. for the binary-class data set), the patches from tissue regions of WSIs are considered for prediction. From the extracted patches, we focus on a particular part in the middle, the central region, and the remaining part as a border for each patch. Similar to the PCam data set, the ground-truth of the central region is considered as the ground-truth of the whole patch. Although the patches overlap, we must emphasize that the central regions do not overlap.

Let W_i be the i^{th} WSI from the test set, and I_{ij} the j^{th} patch in W_i. The trained model (M_d) predicts the probability ($\hat{g}_{ij} \in \mathbb{R}^{l \times l \times C}$ or $\mathbb{R}^C$) of each class for each pixel in I_{ij} for segmentation, or for the whole patch for classification. However, in the case of segmentation, the predicted probability of the central region of a patch is taken into account during evaluation. In other words, the predicted probability $s_{ij} \in \mathbb{R}^{r \times r \times C}$ is considered, where $r \times r$ is the dimension of the central region and s_{ij} is the centre crop of $\hat{g_{ij}}$.

During the evaluation, we consider class $\mathbb{C}$ as the positive class and the other class(es) as the negative class for both the binary and multi-class data sets. Moreover, only the evaluation result for class $\mathbb{C}$ is presented, hence focusing on cancer detection. We also consider the patch-based evaluation for both the segmentation and classification tasks. It is worth mentioning here that we aim to detect the regions in a WSI where there are unhealthy parts, and filter out the WSIs that do not show any signs of cancer for the pathologists. In this context, it is not harmful if some pixels around a detected region are miss-classified due to the existing gaps between the actual ground-truth and human annotations. Notably, having annotation gaps in the ground-truth is acceptable for images whose size can be measured in gigapixels, since it is impractical to consider every pixel during manual annotation by a human. Consequently, it becomes less relevant to consider the usual pixel-based evaluation for the segmentation task. Hence, we consider the patch-based evaluation rather than the pixel-based evaluation for both segmentation and classification.

To this end, we need to convert the pixel-based predicted probabilities of the segmentation task to a patch-based probability. Indeed, we require the predicted probability of class $\mathbb{C}$ ($s^{\mathbb{C}}_{ij} \in \mathbb{R}$) for each patch rather than for each pixel, which we obtain by taking the class-wise max over $s_{ij} \in \mathbb{R}^{r \times r \times C}$. The predicted probabilities ($s^{\mathbb{C}}_{ij}$) are computed for all the patches in the test set.

Since one random fold is not enough to validate a hypothesis, we perform 10 trials (runs) for each experiment. The final result for an experiment is given by the mean computed over the 10 runs. We also compute the standard deviation to reflect how accurately the mean represents the 10 runs [38]. For our 11 experiments presented in Tables 2, 3, 4 and 5, we therefore have a total of $11 \times 10 = 110$ runs to test all the hypotheses for a particular hyper-parameter setting in relation to a particular data set.

4 Experimental Details

4.1 The Data Sets and Preprocessing

In our study we used three data sets: one multi-class and two binary-class. The multi-class data set is for the segmentation task, and the binary class data sets are for classification.

4.1.1 Multi-class Data Set

We used the Metastatic Lymph Node data set from the University Cancer Institute of Toulouse-Oncopole, which is abbreviated as MLNTO. This is a private data set composed of metastatic lymph nodes from various primary tumors, such as melanoma, adenocarcinoma and squamous cell carcinoma relating to various anatomical sites. The data set contains 61 WSIs (34 for training and 27 for testing) of lymph nodes. With one exception, all the WSIs included in the MLNTO data set contain metastasis. However, the metastatic WSIs contain enough healthy regions to collect $\neg\mathbb{C}$ examples for training and testing.

In the preparation process of the WSIs, the glass slides were stained with hematoxylin and eosin (H&E), and digitized with a 3DHISTECH Pannoramic 250 digital slide scanner at 0.243 μm per pixel resolution. Two expert pathologists provided the ground-truth segmentation masks for all the WSIs. The ground-truth annotations make MLNTO suitable for the segmentation task, although we can easily perform classification as well.

Three classes were considered during the annotation (see Figure 2): *metastasis/cancer* ($\mathbb{C}$), *lymph node/non-cancer* ($\neg\mathbb{C}$) and so-called *other* ($\mathbb{O}$), the latter being either background or histological structures, such as adipose or fibrous tissue. The ground-truth masks are provided for WSIs which have been downsampled 8 times with respect to the highest resolution, i.e. level 3. The average size of a training WSI is $9,488 \times 14,648$ pixels (after downsampling by a factor of 8 from the highest resolution).

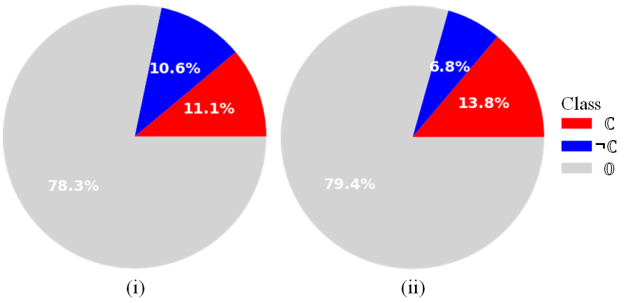


Fig. 4: Class distributions of the pixels in the training (i) and test (ii) sets from the MLNTO data set.

Figure 4 shows the class distributions of the pixels in the training and test sets respectively. The statistics indicate that the natural data distribution is imbalanced, with an over-representation of class $\mathbb{O}$ in comparison to classes $\mathbb{C}$ and $\neg\mathbb{C}$. In the training set, the percentage of classes $\mathbb{O}$, $\mathbb{C}$ and $\neg\mathbb{C}$ are on average 78.3%, 11.1% and 10.6% respectively, while in the test set they are 79.4%, 13.8% and 6.8%. The imbalanced nature of the data convinced us to better understand the impact of a balanced versus an imbalanced distribution of the classes for the learned models as well as the difficulty of choosing which WSIs to annotate for the purpose of training the models.

In our preprocessing step, a total of 127,898 overlapping patches of size $(l \times l =)$ 384×384 pixels were extracted from the training set with a stride of $l/2$. The number of patches that fall into each category are given in Table 6. Considering the number of training patches, the value of $\mathbb{U}$ in the different experiment settings presented in Table 2 to 5 for MLNTO is 5,000. By applying the same extraction process to the test set, we obtain 101,262 patches, of which 17,351 belong to class $\mathbb{C}$.

Table 6: Number of patches belonging to each category in the MLNTO training set, when WSIs are downsampled by a factor of 8 and the stride is $l/2$ (i.e. 192).

patch category	#patches
Other ($\mathbb{O}$)	90,374
Cancer ($\mathbb{C}$)	15,328
Lymph node/non-cancer ($\neg\mathbb{C}$)	17,274
Mixed ($\mathbb{C}\&\neg\mathbb{C}$)	4,922

4.1.2 Binary Class Data Set

We used two well-known binary class benchmarks, namely CAMELYON16[5] [2] and Patch Camelyon[6] (PCam) [60] to further test hypothesis H2 in the binary classification setting.

The CAMELYON16 data set is a breast cancer data set consisting of 399 WSIs of sentinel lymph nodes, with 270 WSIs in the training set and 129 WSIs in the test set. There are 111 training WSIs and 49 test WSIs containing tumors (i.e. $\mathbb{C}$). All the WSIs are in a multi-resolution pyramid structure, generally with 10 levels of resolution.

The PCam data set is a new benchmark generated from CAMELYON16. Unlike the MLNTO and CAMELYON16 data sets, PCam is a ready-to-go data set, as the patches have already been separated. PCam contains 327,680 patches of 96×96 pixels at level 2 ($10\times$ magnification). In PCam, the ratio of training, validation and test data is 6:1:1. The data set is balanced,

[5] `https://drive.google.com/drive/folders/0BzsdkU4jWx9Bb19WNndQTlUwb2M`

[6] https://github.com/basveeling/pcam

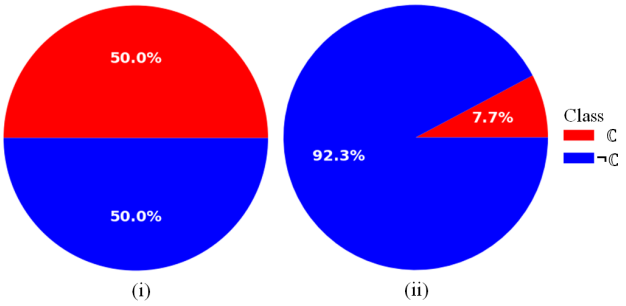


Fig. 5: Class distributions of the PCam data set (i) and the extracted patches from CAMELYON16 test set (ii).

i.e. the ratio of $\mathbb{C}$ and $\neg\mathbb{C}$ patches is 1:1 (Figure 5(i)). The data set considers 32 pixels at the edge as border pixels, while the central region is 32×32 pixels. The patch-level labels are established as follows: if there are any class $\mathbb{C}$ pixels in the central region, the whole patch is labeled as category $\mathbb{C}$; otherwise, the patch-level label is $\neg\mathbb{C}$. Since the ground-truth annotations are class labels instead of class masks, we can only use PCam for the classification task. Besides, since PCam is a binary-class data set, we are only able to test H2 on this data set. To create an imbalanced distribution of a 1:3 ratio in the training set, we set $\mathbb{U} = 43,690$. Note that, we created an artificially imbalanced distribution for the training set only: the validation and test sets are still balanced.

For the CAMELYON16 benchmark we focus on the test set that we used. From the CAMELYON16 test set, the number of extracted test patches of size 96×96 pixels is 10,487,709. The patches were extracted after filtering out the background of the WSIs at level 2, the stride being 32 and the central region being 32×32 pixels. Figure 5(ii) shows the class distribution of the extracted patches, where the ratio of $\mathbb{C}$ to $\neg\mathbb{C}$ is 1:11.9.

4.2 Neural Architecture and Hyper-parameters

In this section we describe the hyper-parameter settings for the segmentation and classification tasks along with the selected network architectures.

4.2.1 Segmentation Task

We selected U-net [55] as our CNN architecture (see Figure 6) for the segmentation task because it has been proven to be effective, even when using a limited number of training images, and only requires a moderate amount of time to train. We implemented the U-net architecture using Keras [11] on the TensorFlow backend.

We used MLNTO to test our hypotheses in the segmentation setting. In all the experiments, we randomly selected 80% of the training data to train the model and the remaining 20% for validation.

Each model was trained from scratch, i.e. without using transfer learning. After a preliminary empirical evaluation, we set the number of epochs (i.e. the number of times the algorithm passes through the entire training data) to 35 and the mini-batch size (i.e. the number of training examples that are provided as input in one iteration) to 5. We opted for the categorical cross-entropy as the loss function, similarly to the original U-net. To optimize the objective function, we used the Adam optimizer [31]. The initial weights were drawn randomly from the zero-mean Gaussian distribution as recommended in [34]. We used a standard deviation of 0.05, which is the default setting in Keras.

4.2.2 Classification Task

The CNNs are inherently translation equivariant, i.e. they can learn the same features from any particular location in an image. In WSIs, besides translation, the histological structures can be in any orientation. In other words, the rotation and reflection are common features in WSIs [20]. Thus, we selected the G-CNN architecture [12], a rotation and reflection equivariant architecture which was tuned for PCam data set by Veeling et al. [60]. The equivariant nature of G-CNNs is effective, and leads to state-of-the-art performance levels for WSI data [60]. For the detailed description of the G-CNN architecture, we refer to the article by Veeling et al. [60].

We used the PCam training and validation sets for training and validation. The models were tested with the test set of both PCam and CAMELYON16 test sets.

For the hyper-parameter setting, we kept the same[7] proposed by Veeling et al. [60] to reproduce the results (when we use the full training set) achieved by them, then used the same setting for our defined training sets of different distributions. We changed however the mini-batch size to match our hardware configuration; the maximum possible mini-batch size while training a G-CNN was 8. Like the original setting, we utilized the cross-entropy as the loss function and Adam as the optimizer. The initial learning rate was $1e^{-3}$, and it was reduced by a factor of $\sqrt{0.1}$ after every 10 epochs if there was no improvement in the validation loss. The training was stopped early if there was no significant improvement in validation accuracy within 12 epochs.

[7] https://github.com/basveeling/keras-gcnn

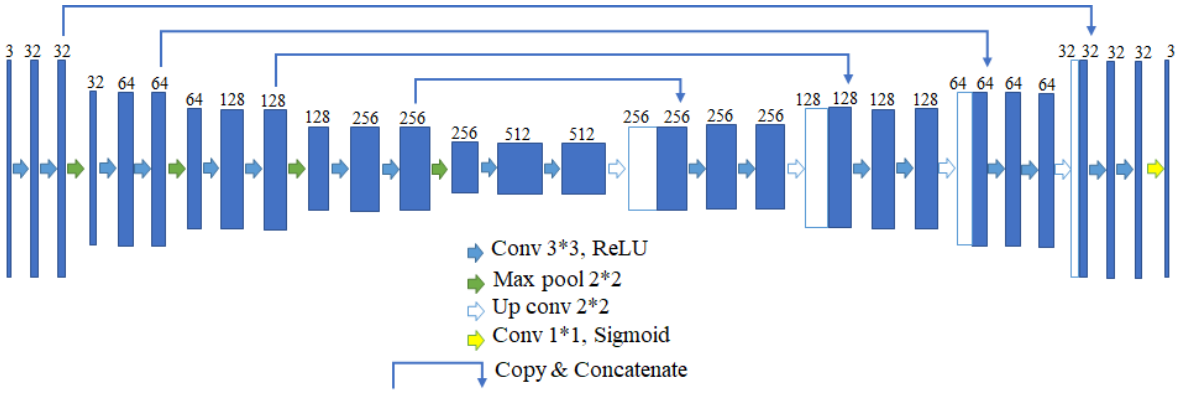


Fig. 6: The U-net architecture: the first and last layers represent the input and the output of the model respectively, while the layers that come after an arrow indicate the output of the operation denoted by the arrow. The number on top of each layer indicates the number of feature channels.

4.3 Evaluation Measures

We considered several metrics in our evaluation because no single metric can serve all our purposes [8]. Table 7 presents the definitions of the evaluation metrics under consideration.

In our hypotheses, we emphasize false positives (FPs), and consider the precision and the false positive rate (FPR) as representative metrics. Along with the FPs, measuring the false negatives (FNs) is also important. We consider recall (also known as sensitivity or true positive rate, TPR) as a metric which is responsive to FNs.

Rather than measuring these values separately, we preferred to resort to curve-based metrics that show the trade-off between two measures. The Receiver Operating Characteristic (ROC) curve plots the TPR at every FPR point. The Precision-Recall (PR) curve plots the precision at every recall point. Note that, in these cases the predicted probability threshold is automatically set to obtain a certain value of, let us say, recall. This means that in a ROC or PR space, a particular point is computed for different models at different thresholds of predicted probability. Thus, a fair comparison has to be prevented to enable hypothesis testing. To resolve this challenge, we computed precision and FPR for a constant list of predefined predicted probability thresholds for all the models, and presented them as precision and FPR curves. The constant list of thresholds includes 101 points from 0.0 to 1.0 by maintaining a regular stride. Moreover, we used the Area Under Curve (AUC) to compare curves overall.

To calculate the mean curve of the 10 runs, we normalized the predicted probabilities for each run by applying min-max normalization, then computed the curve-based metrics for each run. To calculate the mean ROC and PR curves, we selected 1,001 points from both the ROC and PR spaces for each run by using linear interpolation. We then calculated the mean over 10 runs. We considered 1,001 points, as we have a huge number of data points (i.e. test patches) ranging from 32k to 10.5M, depending on the data set.

For FPR and precision curves, no interpolation is required to compute mean curves, since the predicted probability threshold is constant for all runs.

5 Results and Discussion

In this section we present the results which allow us to decide whether a hypothesis is true or false. First, we experiment on the multi-class data set (i.e. MLNTO) while taking the segmentation task into account. Later, we discuss the results for the binary class data sets while taking the classification task into account. In both cases, we make a decision about a hypothesis based on the numeric results before discussing the reasons for the validity or invalidity of a hypothesis.

5.1 Results from Multi-class Data Set MLNTO

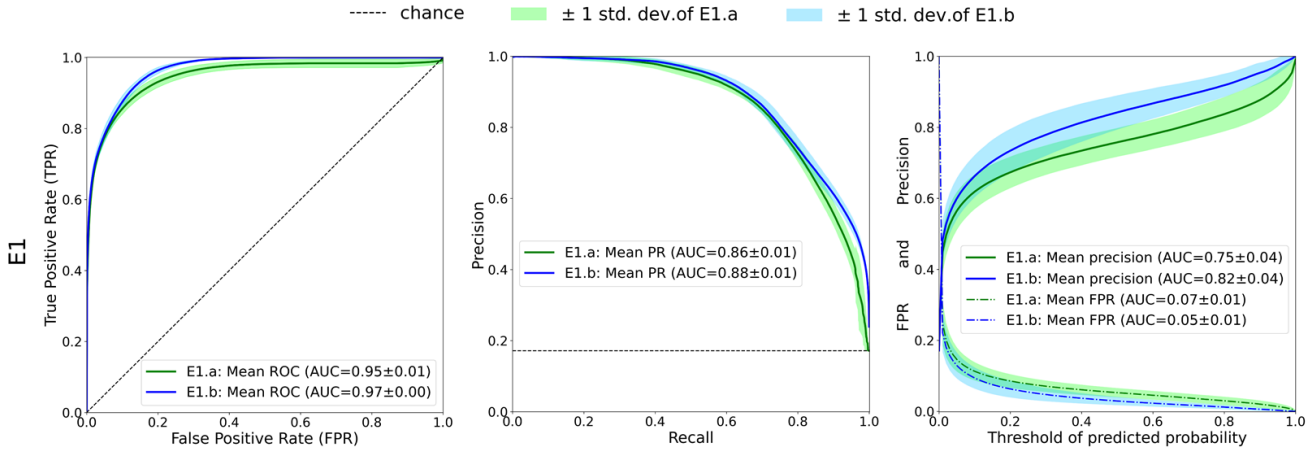
5.1.1 Natural Imbalanced Distribution is better

If we consider the experiment settings E1 as summarized in Table 2, we can observe that the results do not confirm hypothesis H1 and we can conclude that *the naturally imbalanced distribution for class $\mathbb{O}$ (E1.b) produces more accurate result than the artificial balanced distribution (E1.a)* for WSIs in the segmentation setting.

Figure 7 shows the main results for H1. While considering the trade-off between two metrics, in Figure 7(i),

Table 7: Definition of Evaluation metrics. Notations: TP is True Positive, TN is True Negative, FN is False Negative, FP is False Positive, PPV is Positive Predictive Value, TPR is True Positive Rate, FPR is False Positive Rate, ROC is Receiver Operating Characteristic, PR is Precision-Recall, AUC is Area Under Curve.

Metric	Definition
Precision, PPV	TP / (TP + FP)
Recall, Sensitivity, TPR	TP / (TP + FN)
FPR	FP / (TN + FP)
ROC curve	Plot of TPR at every FPR
PR curve	Plot of precision at every recall
FPR curve	Plot of FPR at different threshold of predicted probabilities
Precision curve	Plot of precision at different threshold of predicted probabilities
AUC	Area between the curve and x-axis



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

Fig. 7: **Natural distribution (E1.b) is better than balanced distribution (E1.a) (¬H1).** Mean curves for (i) the ROC curve, (ii) the PR curve, and (iii) the precision and FPR curves. Sub-figures are obtained for 10 runs for balanced distribution (E1.a, lines in green) and for natural over-representation of $\mathbb{O}$ distribution (E1.b, lines in blue). The standard deviation is represented by green/blue colored shading. MLNTO data have been used.

which represents the ROC curve, we can observe that E1.b has a higher true positive rate (TPR) than E1.a at almost every false positive rate (FPR) point. In other words, models trained under the natural distribution give higher recall while also producing fewer false positives (FPs) than models trained under the balanced distribution.

According to the precision-recall (PR) curve presented in Figure 7(ii), the precision of E1.b is also higher at almost every recall point than that of E1.a. Even at the recall point 1.0, the precision of E1.b is better than that of the random chance baseline, while the precision of E1.a is as low as the random chance.

With regard to the precision and FPR curves in Figure 7(iii), the conclusion in favor of E1.b also holds: E1.b has higher precision and a lower FPR than E1.a for all predefined thresholds.

We also applied statistical tests. When precision is calculated using the default argmax decision of the model, precision is about 7% higher in E1.b than E1.a, and the difference is statistically significant according to the t-test (p-value < 0.002).

We further investigated why the distribution helps the models, if it is imbalanced towards $\mathbb{O}$, by manually looking at the predicted masks of the models. Pathologists confirmed that the miss-classification by E1.a (the balanced distribution) is certainly due to inter-class similar regions: the regions containing common histological structures, e.g. blood, fibrous tissue, etc. These histological structures are common in both ROI ($\mathbb{C}$ and $\neg\mathbb{C}$) and non-ROI ($\mathbb{O}$) regions (see Figure 8). When an ROI contains such a histological structure in a small area, the human annotator overlooks the small area, annotating it as a corresponding ROI ($\mathbb{C}$ or $\neg\mathbb{C}$), although the annotation is actually wrong. During training, we therefore need enough examples of these inter-class similar regions (i.e. the over-representation of class $\mathbb{O}$) with their actual annotation to compensate for the unavoid-

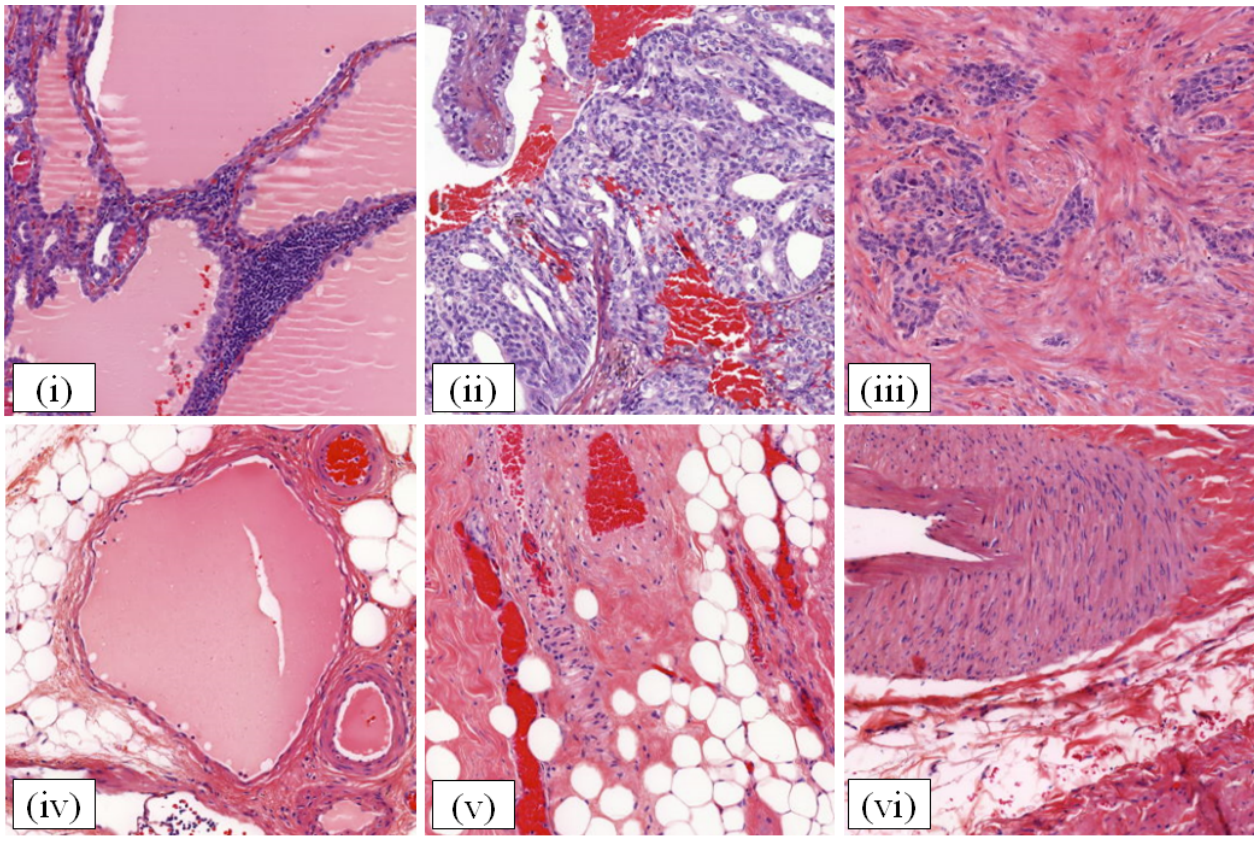


Fig. 8: **Some examples of confusion between other ($\mathbb{O}$) and cancer ($\mathbb{C}$)** classes showing inter-class similar regions with common histological structures. On the first row, (i), (ii) and (iii) are patches of category $\mathbb{C}$ with fluid, blood and fibrosis with lymphocytes respectively, which are taken from the training set. On the second row, (iv), (v) and (vi) are patches of category $\mathbb{O}$ from the test set with the same histological structures respectively. During training, if there are not enough examples of class $\mathbb{O}$ containing the aforementioned structures compared to $\mathbb{C}$, the common histological structures in the test patches are predicted as $\mathbb{C}$ (false positive).

able pitfalls arising from the annotation, or else false positives may be caused during prediction.

5.1.2 Over-representing $\neg\mathbb{C}$ Class Reduces False Positives

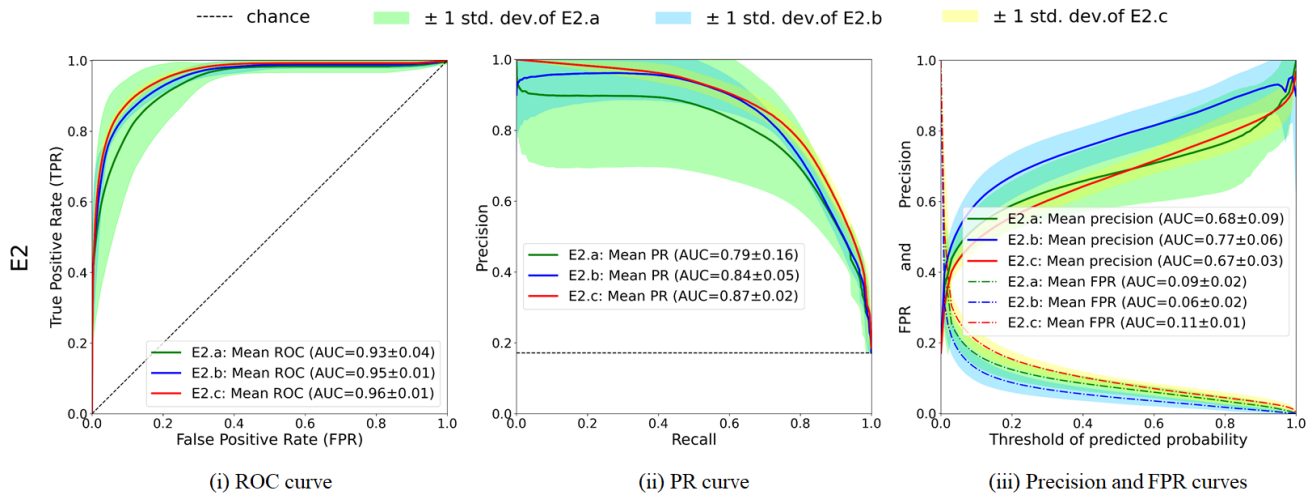
For testing hypothesis H2, *"over-representing $\neg\mathbb{C}$ class reduces false positives"*, we first used the E2 settings, where every example belongs to a single-label category (see Table 3).

Figure 9 presents the main comparative results among the experiments, E2.a (for balanced distribution), E2.b (for over-represented $\neg\mathbb{C}$) and E2.c (for over-represented $\mathbb{C}$) for H2. In Figure 9 we can first observe that the balanced distribution (E2.a) is not as robust as the other distributions (high standard deviation in green). We suspect that the random selection of the under-represented class $\mathbb{C}$ in the training set could not cover all the cancer types from various anatomical sites, thereby inducing more uncertainty in the results and causing this deviation (we further discuss this in Section 5.1.3).

Nonetheless, if we consider the mean results (curves) over 10 runs, we can see that the balanced distribution (green curve) is at its lowest for the ROC curve (see Figure 9(i)). The results are closer for the over-representation of $\mathbb{C}$ (red curve) and $\neg\mathbb{C}$ (blue curve).

The same holds for precision-recall (PR) curves (Figure 9(ii)): while the balanced distribution is the lowest on average and has a high standard deviation, the two other settings are close to each other and more robust, with a very small superiority in favor of the over-representation of $\mathbb{C}$.

While the previous two curves (ROC and PR) are computed for automatic non-fixed thresholds and present the trade-off between two metrics, the precision and FPR curves are computed for predefined constant thresholds for all models and present threshold-wise fair comparison (Figure 9(iii)). When looking at the curves on



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

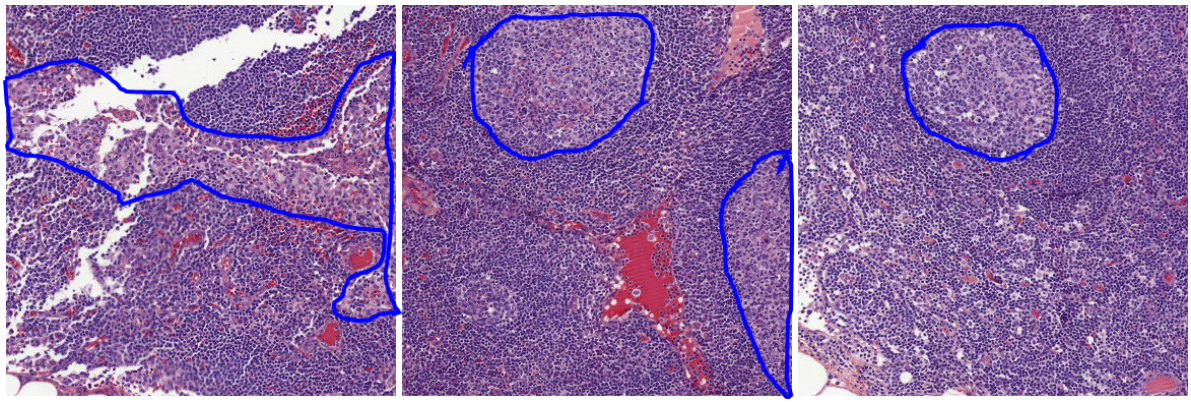
Fig. 9: **Over-representation of negative classes $\neg\mathbb{C}$ reduces false positives (hypothesis H2).** Mean curves with standard deviation for experiment settings E2.a (for balanced distribution in the training set, green), E2.b (for over-represented $\neg\mathbb{C}$, blue) and E2.c (for over-represented $\mathbb{C}$, red). Experiments test H2 in the case of E2, and every example belongs to a single-label category. Same notations as in Figure 7.

Figure 9(iii), we can see that the precision is the highest and the false positive rate (FPR) the lowest in the case of the over-representation of $\neg\mathbb{C}$ (blue curve). We can also examine the area between the curve and the x-axis (i.e., AUC) to compare the curves. The AUC for the over-representation of $\neg\mathbb{C}$ (E2.b) in terms of precision is 0.77 compared to 0.67 and 0.68 for the over-representation of $\mathbb{C}$ (E2.c) and balanced (E2.a) distribution respectively. Moreover, according to the t-test on the precision calculated from the default argmax decision of the models, we found that the precision of E2.b is statistically significantly higher than E2.a and E2.c with p-values < 0.0274 and 0.0013 respectively. These results show that our hypothesis is true: the over-representation of the negative class $\neg\mathbb{C}$ reduces the false positives, and the balanced distribution may be less robust than the others we tested.

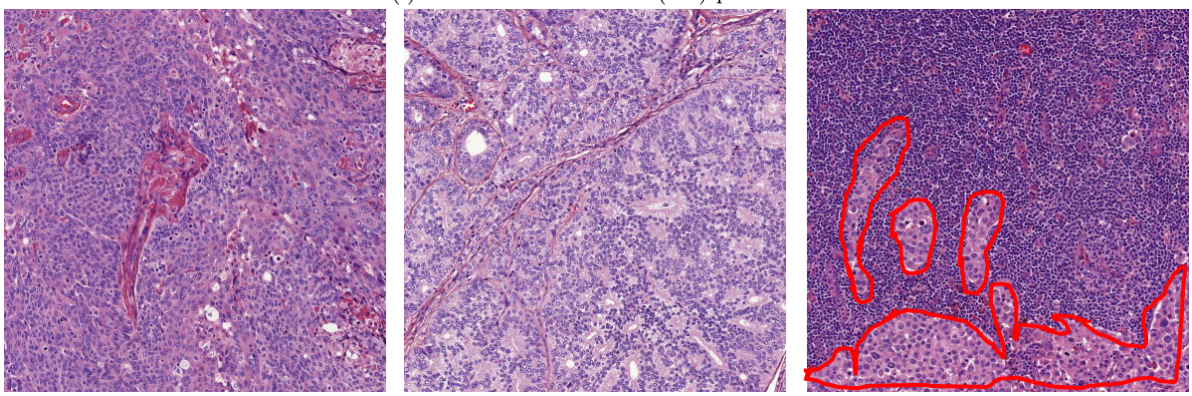
We took a closer look at the data in order to gain a better insight into the numeric results. While examining several heatmap images for the balanced distribution (E2.a) and for the over-represented $\mathbb{C}$ distribution (E2.c), and comparing them to the ground-truth image, we observed that the errors are caused by the inter-class similar regions between classes $\mathbb{C}$ and $\neg\mathbb{C}$ along with the case described in Section 5.1.1 (i.e. in relation to common histological structures).

The inter-class similar regions are the regions which share common visual characteristics between class $\mathbb{C}$ and class $\neg\mathbb{C}$. Specifically, there are some $\neg\mathbb{C}$ regions containing a germinal center, macrophages and a blur area (a kind of artifact) (see Figures 3 and 10) that have some visual characteristics in common with class $\mathbb{C}$ regions.

In Figure 10, we illustrate an example of inter-class similarity and intra-class difference. In Figure 10 (i), the regions inside the blue contours are visually different from the usual $\neg\mathbb{C}$ region (i.e. the region outside of the blue contour); hence, they are the intra-class difference regions of $\neg\mathbb{C}$. These $\neg\mathbb{C}$ regions (inside blue contours) are visually more similar to the $\mathbb{C}$ regions in Figure 10 (ii) than the usual $\neg\mathbb{C}$ regions (outside the blue contours) and hence they are examples of inter-class similar regions between $\mathbb{C}$ and $\neg\mathbb{C}$ classes. For instance, both are lighter in colour than the other parts of the lymph node, the nuclei in both regions are farther from each other than the usual parts of the lymph node, and the nuclei in both regions are larger than the usual size of a nucleus. Hence, during training, if there are not enough examples of $\neg\mathbb{C}$ containing these *specific* inter-class similar regions compared to class $\mathbb{C}$ examples, class $\mathbb{C}$ will dominate. As a result, it is more likely that these regions will be predicted as $\mathbb{C}$ in any of the test WSIs, although many of them will be false positives. Indeed, in practice, the number of these *specific* regions in the lymph node are few compared to the whole lymph node. Thus, it is better to choose a class-biased distribution toward $\neg\mathbb{C}$ so that enough examples containing the aforementioned *specific* regions are used during training. The result of experiments E2 in Figure 9 indicates this conclusion empirically. We further test the H2 in two other settings, E3 and E4, in the following subsections, along with two other hypotheses H3 and H4.



(i) Three full non-cancer (¬ℂ) patches.



(ii) First two patches are cancer (ℂ) patches with 100% class ℂ pixels, while the last one is a mixed (ℂ&¬ℂ) patch (cancer inside the red contour and non-cancer outside).

Fig. 10: **Examples of intra-class difference in ¬ℂ and inter-class similar regions between cancer ℂ and non-cancer ¬ℂ classes**. (i) The regions inside the blue contours are examples of the intra-class different regions for the non-cancer ¬ℂ class, which are visually different from the usual non-cancer ¬ℂ region (i.e. the region outside the blue contour); these regions are visually more similar to the cancer ℂ regions in (ii) than the non-cancer ¬ℂ regions, and are thus examples of inter-class similar regions between cancer ℂ and non-cancer ¬ℂ classes.

Figure 11 shows some examples of heatmaps of class ℂ generated by the experiments E3.a, E3.b and E3.c for the WSI given in Figure 2. From Figure 11, we can see that the prediction by E3.b is better than the two other predictions regarding the number of false positives in the prediction.

5.1.3 Multi-label Examples give Extra Advantages over Single-label Examples

Unlike in the E2 settings, where every example belongs to a single-label category, in the E3 settings presented in Table 4, some of the examples belonging to a single-label category are replaced by multi-label examples (i.e. some pixels are ℂ and some are ¬ℂ in the same patch example). E3 setting serves two purposes; one is to test H3 (multi-label examples are more useful than single-label examples) by comparing with the experiment setting E2, and another is to revisit hypothesis H2 in a different setting. Figure 12 presents the mean curves and the standard deviation of 10 runs for E3.

By comparing single-label (Figure 9) with multi-label (Figure 12), we can immediately see that the results in Figure 12 have improved by considering both the curves and AUC values. The ROC curves in Figure 12(i) are more stable (i.e. have less deviation) and have higher AUC than those in Figure 9(i) except for E3.c (ℂ-biased). Same improvements have been observed in the precision-recall (PR) curves, precision curves, and FPR curves. Multi-label examples significantly improve AUC in all metrics.

According to the t-test on the precision calculated from argmax decision, we found that the precision of E3.a and E3.b are significantly higher (p-value < 0.002 and 0.0002 resp.) than those of E2.a and E2.b respectively, while the precision of E3.c is significantly lower

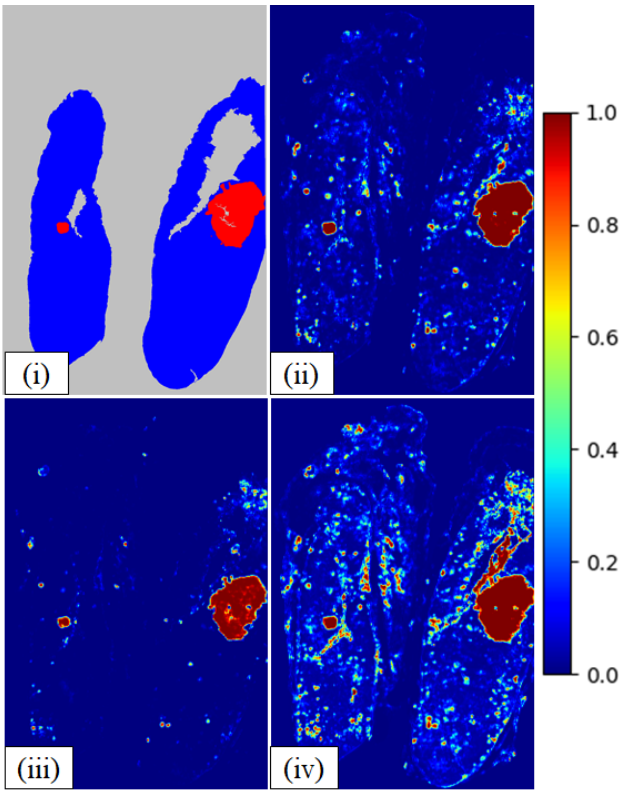


Fig. 11: **Illustrative example E3.b is the best**. Heatmaps of class $\mathbb{C}$ for the WSI given in Figure 2 (one of the WSIs with the most false positives) from the test set of MLNTO. Here, (i) is the ground-truth, (ii) is the heatmap generated by the E3.a, (iii) by E3.b, and (iv) by E3.c.

(p-value < 0.04) than that of E2.c. In other word, the findings are consistent in all evaluation metrics. Hence, the H3 is true for E3.a and E3.b while not for E3.c.

We investigate the reason for having lower result for E3.c than for E2.c. According to our observation, the reason is the ratio of positive and negative classes in the training set of E3.c. Including the patch category $\mathbb{C}\&\neg\mathbb{C}$ in E3.c changes the ratio of classes (negative:positive) from 1:3 (see Table 3) to 1:4 (see Table 4), i.e., it decreases the presence of negative class since $\mathbb{C}\&\neg\mathbb{C}$ category contains class $\mathbb{C}$ (positive class) pixels along with class $\neg\mathbb{C}$ (negative class) pixels. Consequently, increasing the over-representation of the positive class decreases the result of E3.c. Thus, according to our general hypothesis H2, having a lower result in E3.c than E2.c is expected. At the same time, the better result of E3.a and E3.b than the corresponding E2.* empirically proves that H3 is true.

The multi-label patch category $\mathbb{C}\&\neg\mathbb{C}$ gives two advantages: 1) it gives an opportunity to the FCNN models to learn about two classes from the same patches thus gives a chance to learn boundary line between those classes, 2) In the case of micro-metastasis, it is not always possible to extract single-label patches for category $\mathbb{C}$ thus there is a chance of missing examples from a cancer type or WSI, however, it is always possible to extract patches of category $\mathbb{C}\&\neg\mathbb{C}$; thus it almost reduces the chance of missing any cancer type and example from a WSI which reduces the uncertainty and produces a less deviated result.

We re-tested the H2 in the current setting E3 by comparing the results of E3.a, E3.b, and E3.c as described in the following paragraphs.

While considering the trade-off between FPR and TPR in ROC curves (see Figure 12 (i)), the results are comparable for all experiments (all ROC AUCs are 96), i.e., the ROC curve is insensitive to the different distribution in training set for this case.

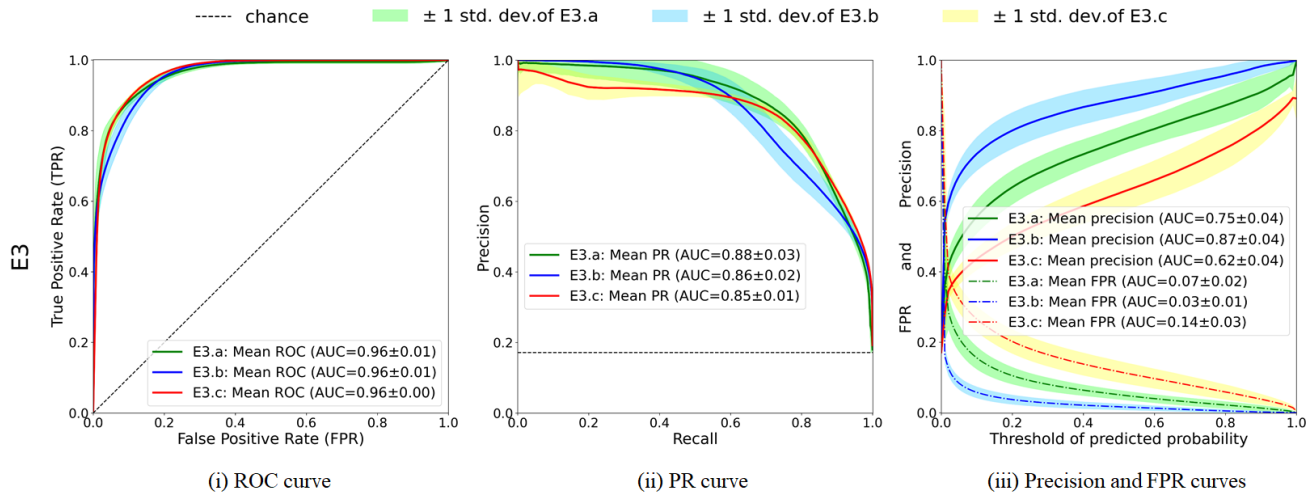
According to the PR curves, and depending on the trade-off under consideration, the settings have a different impact. For high precision, the over-represented $\neg\mathbb{C}$ (in blue) is both robust (low standard deviation) and has a higher impact than the other settings. From these results, it is also clear that along with the distribution of the training set, the cutoff or threshold selection is also important if we are to obtain a desirable performance from the system.

The differences are much greater when we consider precision and false positive rates (FPR) curves (Figure 12 (iii)). If we take a closer look at the blue curves on the right side of Figure 12, the over-representation of $\neg\mathbb{C}$ (E3.b) has the highest precision (AUC 0.87 $\pm$0.04) and the lowest FPR (AUC 0.03$\pm$0.01), while the balanced distribution reaches an AUC of 0.75 $\pm$ 0.04 for precision and 0.07 $\pm$0.02 for FPR. These results are consistent with the results from Figure 9(iii) in terms of the interest of over-representing $\neg\mathbb{C}$, i.e., hypothesis H2.

Additionally, we used the t-test statistical test for precision when it was calculated using the default argmax function. We found that the precision is statistically significantly higher for E3.b (0.888 ± 0.053) than for both E3.a (0.776 ± 0.039) and E3.c (0.639 ± 0.056). Comparison between E3.b and E3.a provides the p-value $< 2.95e^{-05}$, while it is $< 3.62e^{-09}$ for E3.b and E3.c. In other words, according to the t-test, the H2 is significantly true in setting E3 as well.

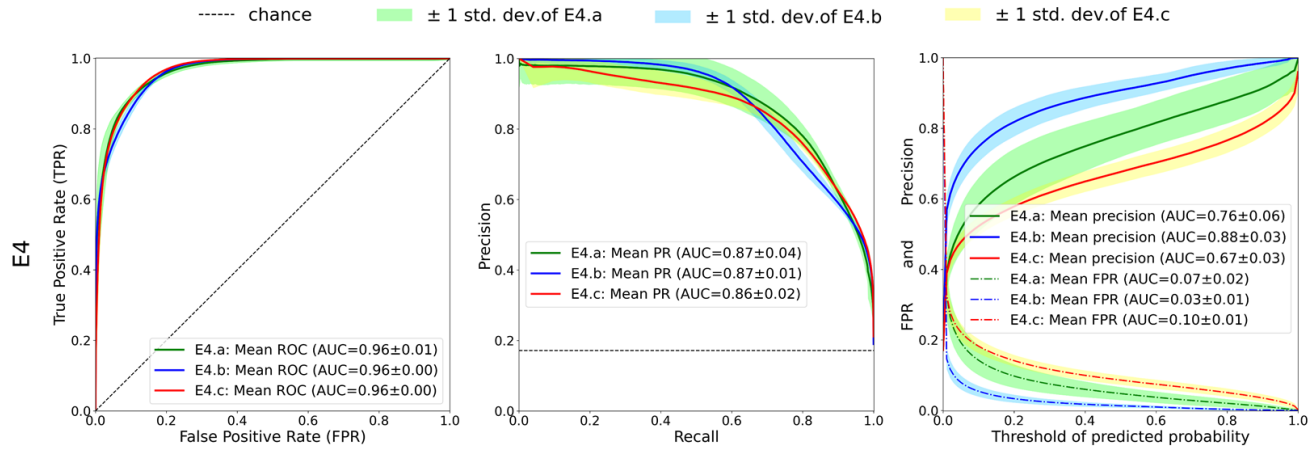
5.1.4 Non-ROI Data are useful

If we compare the result of E4.* (with non-ROI patches) (Figure 13) with the result of E3.* (without non-ROI patches) (Figure 12), we can see that the results for



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

Fig. 12: **Multi-label examples give extra advantages over single-label examples (H3 hypothesis).** Mean curves with standard deviation using E3 settings (Table 4) where there are multi-label examples: E3.a (for balanced distribution in the training set, green), E3.b (for over-represented $\neg\mathbb{C}$, blue) and E3.c (for over-represented $\mathbb{C}$, red) and using the same notations as in Figure 9.



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

Fig. 13: **Non-ROI data are useful (H4 hypothesis).** Mean curves with standard deviation using E4 settings (Table 5) where $\mathbb{O}$ examples replace some $\mathbb{C}$ examples and some $\neg\mathbb{C}$ examples: E4.a (for balanced distribution in the training set, green), E4.b (for over-represented $\neg\mathbb{C}$, blue) and E4.c (for over-represented $\mathbb{C}$, red) and using the same notation as in previous figures.

E4.* are slightly better than, or comparable with, the corresponding result for E3.*.

ROC curves in Figure 13(i) are very similar to each other and also very similar to the curves obtained in E3 (AUC=0.96) (Figure 12(i)). Precision-recall (PR) curves in Figure 13(ii) have similar shapes and are ordered the same way as in E3 (Figure 12(ii)). AUC values are also quite similar to each other when compared to E3.

According to Figure 13(iii), the results for E4 have slightly improved compared to E3 (AUC increases slightly for precision and decreases slightly for the false positive rate (FPR)) in Figure 12(iii). The standard deviation for the 10 runs also slightly decreases, except for the balanced distribution, which still has the highest standard deviation, regardless of the measure being considered. By using the t-test on precision with the default argmax decision on predicted probabilities, we found that the higher precision for E4.b compared to E3.b is not statistically different (p-value 0.22), while it is for the E*.c case (p-value 0.0034).

Nonetheless, the better results for E4.* compared to E3.* indicate that a considerable number of errors come from inter-class similar regions (or common histological structure) between ROIs and non-ROIs, as discussed in Section 5.1.1. However, miss-classification due to common histological structure (e.g. blood) between ROIs and non-ROIs is easy to identify, even with the naked eye, while miss-classification due to inter-class similarity between the ROI classes, $\mathbb{C}$ and $\neg\mathbb{C}$, is more difficult to identify. This explains why the experiment setting for E3.b is more appropriate than the setting for E4.b. Nevertheless, as H4 is a true hypothesis, inexpensive to annotate non-ROI examples will be useful for CNN training. Therefore, if there is a shortage of expensive to annotate ROI examples while training a CNN, this can be compensated by adding relatively easy-to-annotate non-ROI examples. This result is also consistent with our results from hypothesis H1.

As like as E3 setting, we re-test H2 in the E4 setting, and the findings are quite similar to the E3 case. In other words, H2 is true in E4 settings as well.

5.2 Results for the Binary Class Data Sets

In this section, we further discuss the results of hypothesis H2 ($\neg\mathbb{C}$-bias produces fewer false positives) for the binary class data sets PCam and CAMELYON16 for the classification task. Here, all the models were trained on the PCam data set. The trained models were then tested on both PCam (balanced) and CAMELYON16 (highly imbalanced) test sets.

5.2.1 Balanced Test Set: Over-representing the Negative Class reduces False Positives

According to the ROC and PR curves in Figure 14 (i) and (ii), all the distributions produce comparable results (also observed when zooming in). However, we can see that E2.b has a higher level of performance than the other two experiments at low thresholds.

According to the precision and FPR curves in Figure 14 (iii), E2.b has the highest level of performance, although for the threshold of predicted probability greater than 0.8, the results are comparable. According to the evaluation based on the argmax decision on the model, the precision of the E2.b is higher than E2.c (p-value < 0.019) and not statistically different from E2.a.

In other words, for the classification task with balanced test and validation sets, the $\neg\mathbb{C}$-biased distribution produced fewer FPs than the other two distributions, which is consistent with previous results, and confirms the H2 hypothesis. However, this is not always significantly true for this classification task with balanced test and validation sets. We must assume that the balanced distribution of the validation set, which is used for parameter tuning during training, might be the reason.

5.2.2 Imbalanced Test Set: Over-representing the Negative Class reduces False Positives

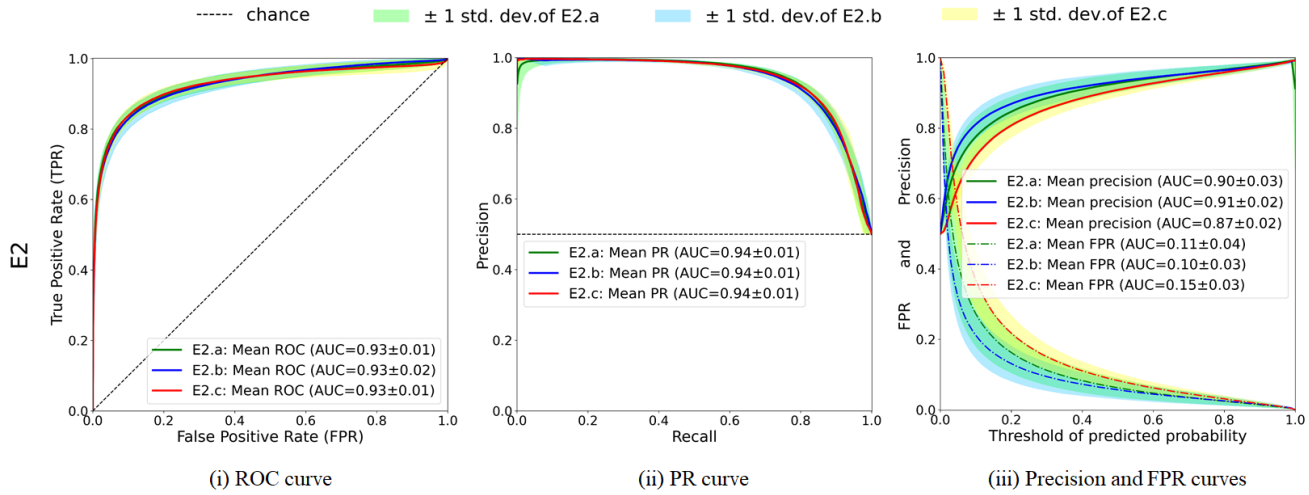
Figure 15 illustrates the test results for H2 on CAMELYON16, whereas the models were trained on PCam. According to this figure, the results on CAMELYON16 are consistent with the results on PCam, i.e. H2 is true, although not always significantly true for both balanced and imbalanced test sets, while the validation set has a balanced distribution.

Figure 15 (i) shows that the ROC curves are insensitive to the different distributions in the training set. It may be because of the highly optimistic nature of ROC curves with regard to a highly imbalanced test set [15].

On the other hand, in Figure 15(ii), it is clear that the PR curves are also insensitive to the different distributions in the training set for a highly imbalanced test set, except for a certain range of recall where we can observe some slight differences in favor of the balanced distribution for the recall range between 0.75 and 0.85, along with a high standard deviation.

On the contrary, in Figure 15(iii), the precision of E2.b is higher than the two others for predicted probability threshold less than 0.7. Moreover, E2.b has the lowest FPR. These results are consistent with the results we obtained for the PCam test set. According to the t-test on the precision calculated from the default argmax decision of the models, the finding is consistent with the PCam case as well. Specifically, we found that the precision of E2.b is not significantly higher (p-value ≈ 0.37) than that of E2.a, i.e., both are comparable, while it is significantly higher (p-value < 0.04) than that of E2.c. In short, for the classification task with the mentioned setting, H2 is true, however, not always significantly true.

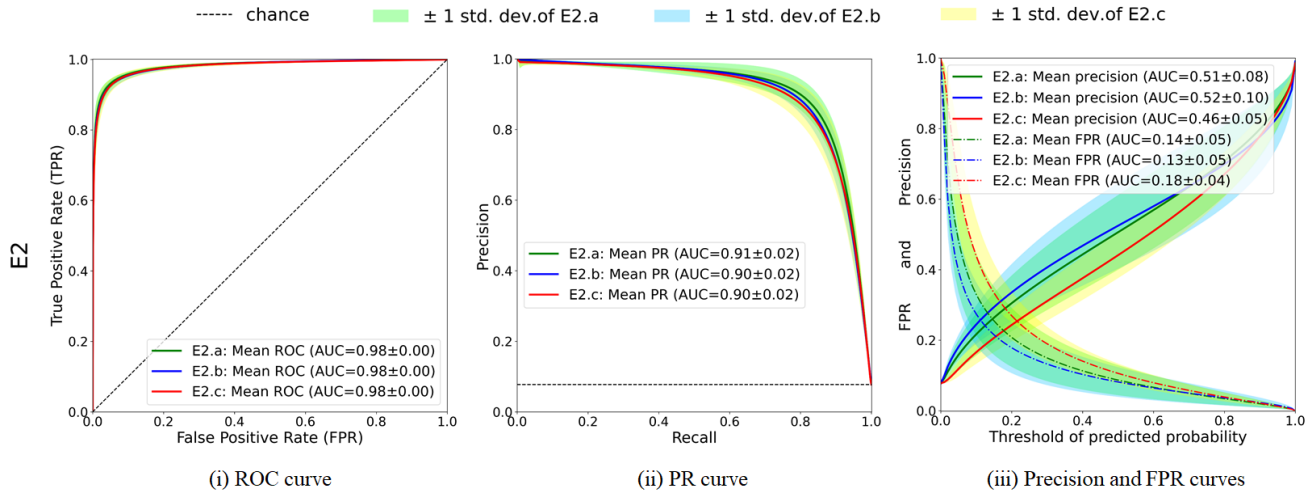
Note that the CAMELYON16 test set is highly imbalanced towards class $\neg\mathbb{C}$ (see Figure 5(ii)). Thus, there is a strong likelihood of producing FPs, which is what the models do. Hence, the AUC of the precision curve in Figure 15(iii) is lower than the other two data set cases presented in Figures 9(iii), 12(iii), 13(iii), and 14(iii). However, the high performance levels of the models in terms of the ROC, PR and FPR curves shown in Figure 15 indicate that they fail to reflect our observation, while the precision curve succeeds. Consequently, the precision curve is more robust



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

Fig. 14: **Over-representing the negative class ¬ℂ reduces false positives - on PCam.** Mean curves with standard deviation (std. dev.) of experiments for balanced distribution (E2.a), for over-represented ¬ℂ (E2.b), and for over-represented ℂ) (E2.c) on PCam.

.



(i) ROC curve (ii) PR curve (iii) Precision and FPR curves

Fig. 15: **Over-representing the negative class ¬ℂ reduces false positives - on CAMELYON16.** Mean curves with standard deviation (std. dev.) of experiments E2.a (for balanced distribution), E2.b (for over-represented ¬ℂ) and E2.c (for over-represented ℂ) using the same keys as in the other figures. In all the experiments, the models were trained on PCam and tested on CAMELYON16.

than the ROC, PR and FPR curves in comparing the performances of the different models.

We also observed the predictions of different experiments for the CAMELYON16 test set, and we found that, like the MLNTO data set, the false positives are the major problem caused by inter-class similarity and intra-class difference. Figure 16 shows some examples of heatmaps predicted by different experiments. According to this figure, the E2.b (for over-represented ¬ℂ) produces fewer false positives than the other two experiments, which confirms that H2 is true.

6 Overall Discussion and Conclusions

In this research we performed a data-level analysis to determine the optimal distribution of the classes in the training set for WSIs when using deep learning. In natural distribution, the WSI data is highly biased towards the non-ROIs, while the distribution of the two ROI classes is variable, depending on the WSIs which have been included in the data set. The biases and variability of the distribution make us interested to investigate different distributions in-depth. To conduct this analy-

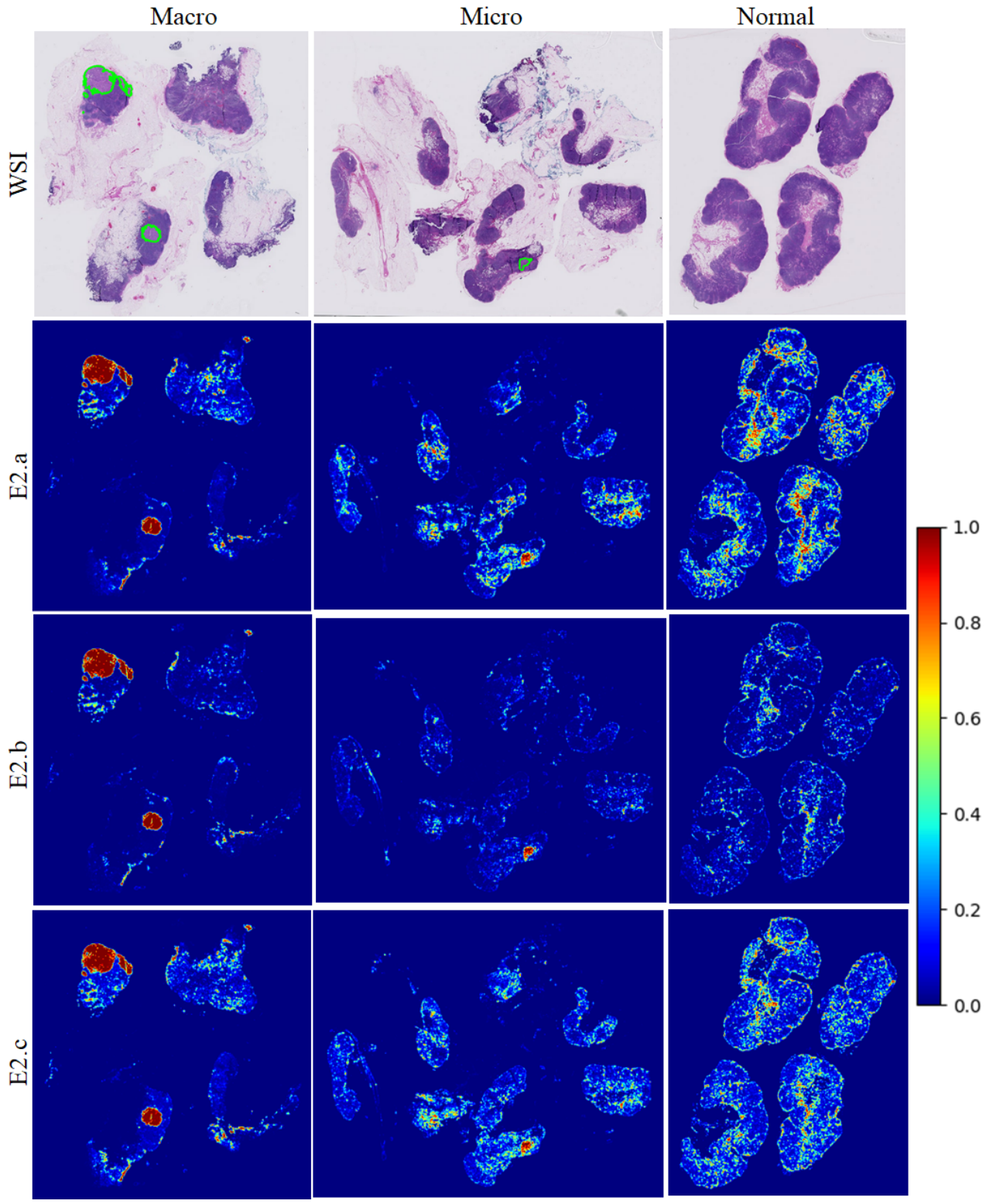


Fig. 16: **Illustrative examples for ¬ℂ-biased training (E2.b) produces the smallest number of FPs (H2).** Example heatmaps of class ℂ generated by the experiments E2.a (for balanced distribution), E2.b (for over-represented ¬ℂ) and E2.c (for over-represented ℂ) for WSIs with macro-metastasis, micro-metastasis and normal tissue from the CAMELYON16 test set. Here, in the WSIs, the regions inside the green contours are ground-truth annotations for class ℂ.

sis, we considered the case of FCNN for segmentation and CNN for patch classification.

To the best of our knowledge, our analysis is pioneering in the case of class distribution analysis of WSI data for deep learning models; previous research has focused on end-to-end pipeline development for cancer detection.

We first compared the natural distribution (biased towards non-ROI pixels/patches) with the more commonly used balanced distribution (H1, experimental setting E1). We found that the natural distribution of the WSI data is superior to the artificially balanced distribution. Since non-ROI examples are usually much easier to annotate, and annotation of ROI is costly in domains where only experts can provide the labels, this result is of huge importance.

We then focused on the distribution of ROI classes. The ROI class distribution can be balanced, as in the MLNTO training set, or highly imbalanced, as in the CAMELYON16 test set. Since the natural distribution of the ROI classes in a WSI data set is variable, choosing an optimal distribution for the ROI class while building a training set is an issue. In H2, and with experimental setting E2, we show that the generally recommended balanced distribution is not the best. Instead, the non-cancer-biased training set produces the best performance, bearing in mind precision and the number of false positives.

In the literature, multi-label patches are considered problematic [22]. According to the test result of hypothesis H3, we found that multi-label patches give extra advantages over single-label patches. Both the AUCs and curves were significantly improved for the multi-label patches case. Thus, when building a training data set, they can be of huge importance. In other words, it is better to have multi-label patches than additional positive examples.

We carried out an in-depth analysis of the results from the first hypothesis that non-ROIs can still be of use. They are indeed useful as a replacement of ROI data in a case where the ROI data are limited or small. Moreover, they are easier to annotate than cancer/non-cancer (i.e., ROI data). This finding is very important because non-ROIs could be the choice for obtaining a large enough data set at a low cost for training a deep model.

In addition to observing the results of deep learning models, we also had a close look at the data to be able to form medically-oriented conclusions. While manually observing the predicted mask, we learned the importance of class heterogeneity and inter-class similarity. When building a new data set of histological images, more examples should be added from a heterogeneous class, which can compensate for the confusion caused by inter-class similarities.

When it comes to the two different tasks, segmentation and classification, we found that classification is less sensitive to the different distributions in the training set.

While we mainly focused on the training set distribution and, to a lesser extent, on the test set distribution for which we obtained consistent results, we did not study the distribution of the validation set. This set can also have an impact, since it is used for tuning the models during training. For example, for the classification task and when anaylzing the impact of the test set distribution, we kept the original (balanced) distribution of the PCam validation set. The balanced distribution of the validation set may have an impact on the relative insensitivity of the test distribution with regards to the classification task. We aim to address this challenge in future work. Moreover, we found that the precision curve is more robust than the popular ROC and PR curves in differentiating models and reflecting the FPs production. A separate comparative study among different evaluation metrics could be another track worth investigating in future work. We tested the hypotheses for the single level of WSIs. It could also be interesting to investigate them for multiple levels of WSIs. We kept this point for future work.

In summary, our study is representative, although not exhaustive. The conclusions could be further tested in other domains. We believe that the outcomes of the analysis will be helpful for researchers who are building a training data set of WSIs or other applications. Such analyses could also help in other real-world problems where data have a complex history, as discussed by Crawford [14] regarding the importance of building a training set with proper distribution.

Acknowledgements

Authors would like to thank Dr. Md Zia Ullah for his fruitful discussions. The research leading to these results has received funding from the NO Grants 2014-2021, under project ELO-Hyp contract no. 24/2020.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Afzal, S., Maqsood, M., Nazir, F., Khan, U., Aadil, F., Awan, K.M., Mehmood, I., Song, O.Y.: A data augmentation-based framework to handle class imbalance problem for alzheimer's stage detection. IEEE Access **7**, 115528–115539 (2019)
2. Alexi, B., Altuna, H., Babak, B.E., Wauters Carla, Geert, L., Jeroen, L.V., Dijk Van Marcory, Maschenka, B., Meyke, H., Nikolas, S., Oscar, G., Paul, D.V., Peter, B., Bult Peter, Manson Quirine, Vogels Rob, Rob, D.L.V.: Supporting data for "1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset" (2018). DOI 10.5524/100439. URL `http://gigadb.org/dataset/100439`
3. Baloch, B.K., Kumar, S., Haresh, S., Rehman, A., Syed, T.: Focused anchors loss: Cost-sensitive learning of discriminative features for imbalanced classification. In: Asian Conference on Machine Learning, pp. 822–835 (2019)
4. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter **6**(1), 20–29 (2004)
5. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
6. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A.: Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology **16**(11), 703–715 (2019)
7. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks **106**, 249–259 (2018)
8. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE transactions on pattern analysis and machine intelligence **41**(3), 740–757 (2018)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
10. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: European conference on principles of data mining and knowledge discovery, pp. 107–119. Springer (2003)
11. Chollet, F., et al.: Keras. `https://keras.io` (2015)
12. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International conference on machine learning, pp. 2990–2999 (2016)
13. Cracknell, M.J., Reading, A.M.: Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences **63**, 22–33 (2014)
14. Crawford, K.: Artificial intelligence's white guy problem. The New York Times **25**(06) (2016)
15. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp. 233–240 (2006)
16. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017). URL `http://archive.ics.uci.edu/ml`
17. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010). DOI 10.1007/s11263-009-0275-4. URL `http://dx.doi.org/10.1007/s11263-009-0275-4`
18. Fan, K., Wen, S., Deng, Z.: Deep learning for detecting breast cancer metastases on wsi. In: Innovation in Medicine and Healthcare Systems, and Multimedia, pp. 137–145. Springer (2019)
19. Farahani, N.: Whole slide imaging in pathology: advantages, limitations, and emerging perspectives (2015)
20. Graham, S., Epstein, D., Rajpoot, N.: Dense steerable filter cnns for exploiting rotational symmetry in histology images. IEEE Transactions on Medical Imaging (2020)
21. Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: A review. IEEE reviews in biomedical engineering **2**, 147–171 (2009)
22. Halicek, M., Shahedi, M., Little, J.V., Chen, A.Y., Myers, L.L., Sumer, B.D., Fei, B.: Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Scientific reports **9**(1), 1–11 (2019)
23. Hamad, R.A., Kimura, M., Lundström, J.: Efficacy of imbalanced data handling methods on deep learning for smart homes environments. SN Computer Science **1**(4), 1–10 (2020)
24. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis **35**, 18–31 (2017)
25. Hinz, T., Navarro-Guerrero, N., Magg, S., Wermter, S.: Speeding up the hyperparameter optimization of deep convolutional neural networks. International Journal of Computational Intelligence and Applications **17**(02), 1850008 (2018)
26. Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q.: Deep learning for image-based cancer detection and diagnosis- a survey. Pattern Recognition **83**, 134–149 (2018)
27. Jaccard, N., Rogers, T.W., Morton, E.J., Griffin, L.D.: Detection of concealed cars in complex cargo x-ray imagery using deep learning. Journal of X-ray Science and Technology **25**(3), 323–339 (2017)
28. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 27 (2019)
29. Kellenberger, B., Marcos, D., Tuia, D.: Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. Remote sensing of environment **216**, 139–153 (2018)
30. Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems (2017)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of ICLR (2015)
32. Koller, O., Ney, H., Bowden, R.: Deep learning of mouth shapes for sign language. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 85–91 (2015)
33. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology Journal **16**, 34–42 (2018)

34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
35. Kubat, M., Matwin, S., et al.: Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol. 97, pp. 179–186. Nashville, USA (1997)
36. Kumar, N., Gupta, R., Gupta, S.: Whole slide imaging (wsi) in pathology: Current perspectives and future directions. Journal of Digital Imaging (2020)
37. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). DOI 10.1038/nature14539. URL `http://dx.doi.org/10.1038/nature14539`
38. Lee, D.K., In, J., Lee, S.: Standard deviation and standard error of the mean. Korean journal of anesthesiology **68**(3), 220 (2015)
39. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–42 (2015)
40. Lin, H., Chen, H., Dou, Q., Wang, L., Qin, J., Heng, P.A.: Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 539–546. IEEE (2018)
41. Lin, H., Chen, H., Dou, Q., Wang, L., Qin, J., Heng, P.A.: Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 539–546. IEEE (2018)
42. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
43. Liu, Y., Gadepalli, K.K., Norouzi, M., Dahl, G., Kohlberger, T., Venugopalan, S., Boyko, A.S., Timofeev, A., Nelson, P.Q., Corrado, G., Hipp, J., Peng, L., Stumpe, M.: Detecting cancer metastases on gigapixel pathology images (2017). URL `https://arxiv.org/abs/1703.02442`. Initial publication on arxiv, then submit to MICCAI.
44. Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., Stumpe, M.C.: Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists. Archives of pathology & lab. medicine (2018)
45. Masko, D., Hensman, P.: The impact of imbalanced training data for convolutional neural networks (2015)
46. Mejbri, S.: Deep learning applied to multivariate medical data. PhD dissertation, Université Toulouse III-Paul Sabatier (2019)
47. Mejbri, S., Franchet, C., Reshma, I.A., Mothe, J., Brousset, P., Faure, E.: Deep analysis of cnn settings for new cancer whole-slide histological images segmentation: the case of small training sets. In: 6th International Conference on Bioimaging (2019)
48. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, pp. 1520–1528. IEEE Computer Society, Washington, DC, USA (2015). DOI 10.1109/ICCV.2015.178. URL `http://dx.doi.org/10.1109/ICCV.2015.178`
49. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
50. Pham, H.H.N., Futakuchi, M., Bychkov, A., Furukawa, T., Kuroda, K., Fukuoka, J.: Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. The American journal of pathology **189**(12), 2428–2439 (2019)
51. Prati, R.C., Batista, G.E., Silva, D.F.: Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowledge and Information Systems **45**(1), 247–270 (2015)
52. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. **29**(9), 2352–2449 (2017). DOI 10.1162/neco-a-00990. URL `https://doi.org/10.1162/neco-a-00990`
53. Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F.J., Granda-Gutiérrez, E.E.: Data sampling methods to deal with the big data multi-class imbalance problem. Applied Sciences **10**(4), 1276 (2020)
54. Reshma, I.A., Cussat-Blanc, S., Ionescu, R.T., Luga, H., Mothe, J.: Natural vs balanced distribution in deep learning on whole slide images for cancer detection. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp. 18–25 (2021)
55. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, vol. 9351, pp. 234–241. Springer (2015). URL `http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a`
56. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of Big Data **6**(1), 60 (2019)
57. Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition **40**(12), 3358–3378 (2007)
58. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
59. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention, pp. 210–218. Springer (2018)
60. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention, pp. 210–218. Springer (2018)
61. Walach, E., Wolf, L.: Learning to count with cnn boosting. In: European Conference on Computer Vision, pp. 660–676. Springer (2016)
62. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. CoRR **abs/1606.05718** (2016). URL `http://dblp.uni-trier.de/db/journals/corr/corr1606.html#WangKGIB16`
63. Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J.: Training deep neural networks on imbalanced data sets. In: Neural Networks (IJCNN), 2016 International Joint Conference on, pp. 4368–4374. IEEE (2016)

64. Weiss, G.M., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Rutgers Univ (2001)
65. Weiss, G.M., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research **19**, 315–354 (2003)
66. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Computation **8**(7), 1341–1390 (1996)
67. Wu, Y., Ding, Y., Feng, J.: Smote-boost-based sparse bayesian model for flood prediction. EURASIP Journal on Wireless Communications and Networking **2020**, 1–12 (2020)
68. Yuan, X., Xie, L., Abouelenien, M.: A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recognition **77**, 160–172 (2018)
69. Zhou, X., Li, C., Rahaman, M.M., Yao, Y., Ai, S., Sun, C., Wang, Q., Zhang, Y., Li, M., Li, X., et al.: A comprehensive review for breast histopathology image analysis using classical and deep neural networks. IEEE Access **8**, 90931–90956 (2020)
70. Zhu, Z., Gallant, A.L., Woodcock, C.E., Pengra, B., Olofsson, P., Loveland, T.R., Jin, S., Dahal, D., Yang, L., Auch, R.F.: Optimizing selection of training and auxiliary data for operational land cover classification for the lcmap initiative. ISPRS Journal of Photogrammetry and Remote Sensing **122**, 206–221 (2016)