Toward reliable false discovery rate control in classification problems under distribution shift

Andrey Borevskiy[1] and Attila Kertesz-Farkas[1,*]

[1]Laboratory on AI for Computational Biology, Faculty of Computer Science, HSE University, 11 Pokrovsky Bvld., Moscow 109028, Russian Federation
[*]Email: kfattila@yandex.com

November 12, 2025

Abstract

Domain shifts or batch effects can significantly degrade the performance of machine learning-based classifiers in software production without any warning. This can particularly be problematic in medical applications which involve human lives. The shift changes the shape of the prediction score distribution of the test samples compared to that of the training samples; hence, decision boundary re-calibration would be required to conform to the expected performance standards. We present a simple, yet robust heuristic method to calculate valid p-values (uniformly distributed) for raw prediction scores of test samples under data shift in general classification problems. Our method is based on an adjustment of the test prediction scores so that the test null distribution (approximated distribution of the prediction scores of negative test samples) aligns to the training null distribution (actual distribution of prediction scores of the negative training samples). We termed it the Test Null Adjustment (TNA) method. Then, accurate empirical p-values of the adjusted test prediction scores can be calculated with respect to the training null distribution to be used in any statistical applications which require test samples accompanied by valid p-values. We demonstrate the usability of our TNA method in FDR control with Benjamini-Hochberg (BH) protocol. The main advantage of the TNA method is that its application does not require any domain specific knowledge. It works in binary and multi-label classification scenarios, it is fully data-driven, and it operates in the 1-dimensional space of the prediction scores ensuring stable numerical calculations. We have performed four experimental tests on the biomedical image analysis domain to demonstrate that the TNA method is able to mitigate the bias in FDR induced by data shift in real-life applications.

Key words: Empirical P-values, False Discovery Rate, Data Shift, Batch Effect

1 Introduction

Modern AI applications not only need to increase their discriminative power to reduce the overall number of incorrect predictions, but also need to simultaneously minimize a quantified risks of false positive (FP) and negative (FN) error by an optimal decision threshold calibration while also being able to handle (or detect) data distribution shift [1, 2]. False positive (type I) and false negative (type II) errors can have significantly different costs or impacts on people and society [3]. For instance, a false positive HIV, COVID-19 or cancer diagnosis can cause emotional trauma and the person may undergo invasive treatments with undesirable complications and side effects [4, 5, 6, 7] until a more elaborated clinical test establishes the correct diagnosis. However, a failure to diagnose an infectious disease may lead to a false sense of safety [8], and may put not only their but other peoples' lives in danger [9]. An undetected cancer may either prolong or hinder treatment [10], and misdiagnosed human papillomavirus (HPV) may progress to cancer precursor lesions [11, 12].

False discovery rate (FDR) control procedures are plausible techniques to keep the FP rate among the total positive classifications at a certain, user-defined level [13], when the risk of FP prediction is significantly higher than the cost of FN predictions. FDR control became popular with the emergence of high-throughput and shotgun experiments which measured large numbers of distinct variables simultaneously per sample in a cost efficient way. For instance, automatic patch clamp systems employ a deep-learning-based image processing module to automatically detect target cells (e.g. nerve cells in living brain). Then it calibrates the movement of a micropipette to approach the cell for future experiments [14]. Cell morphological profiling in high-throughput imaging assays for studying compound libraries and human diseases [15] relies on automatic cell segmentation and detection often in 3D images [16]. The FDR control here could curb the error in selecting wrong cells for subsequent experiments and feature quantification. Microarrays measure the expression levels of thousands of genes simultaneously and the FDR control allows the selection of promising genes for followup studies [17, 18, 19]. Tandem mass spectrometry can be used to quantify and identify thousands of proteins in a complex biological/chemical sample and FDR control makes it possible to choose the most trustworthy protein identifications [20].

The False Selection Rate (FSR) control procedure can restrain the FP and FN errors simultaneously [21]. This approach determines two thresholds: one to make definitive positive predictions, and another one to make definitive negative predictions. Any samples not passing either threshold remain in *indecision*. Therefore, the misclassification error (accuracy) among definitive predictions can be bounded, while a subset of samples without predictions are returned for manual investigation by a human expert or for further data collection [22]. Typically, the FDR and FSR controls, and many other statistical methods heavily rely on valid, i.e. uniformly distributed p-values, while typical deep learning models produce raw discriminative scores or logits for test samples.

In this paper, we are interested in providing valid p-values for a set of discriminative scores $T_s = \{t_1, t_2, \ldots, t_m\}$, where each score t_i is produced by a machine learning classifier for a test instance t_i without knowing its true label. The scores t_is are from a mixture of unknown null and non-null hypotheses; i.e. t_i is generated under one of the two possible distributions: (1) null distribution: $T_i \sim T_0$, and (2) non-null distribution: $T_i \sim T_1$; in a mixture morel: $T_i \sim T = \pi_0 T_0 + (1 - \pi_0)T_1$, where π_0 is the prior probability of that a score comes from the null distribution. The null distribution is associated with the negative class, the non-null distribution is associated with the positive class. The null and the non-null distributions as well as the assignments are unknown due to the lack of information on the true class labels. Implicitly or explicitly, but knowing or estimating the null distribution is essential to obtain accurate, valid p-values with respect to it. In certain practical applications it is possible to label some of the data as negative with high confidence. For instance, decoy peptides [20] in tandem mass spectrometry and internal negative control proteins [23] in proteomic profiling serve this purpose. However, this is not the case with general machine learning applications, and we propose to approximate the test null distribution T_0 with the empirical training null distribution $X_0 = \{x_1, x_2, \ldots, x_n\}$, where all x_is prediction scores belong to the training negative class; that is, $T_0 \approx X_0$.

Therefore, an *empirical p-value* (EPV) of a given discriminative score t could be defined as the fraction of the scores of the negative training data that are greater than or equal to t, formally $p_t = (1 + \sum_{x_i \in X_0} \mathbb{1}\{x_i \geq t\})/(n + 1)$ [eq. 1]. This empirical p-value is also referred to as conformal p-value [25] or as rank among negative controls (RANC) [23]. The EPVs of the negative test samples are super uniform under the null if the data distribution of the test and training samples are identically and independently distributed. This property is called *exchangeability*. In this case, the Benjamini-Hochberg (BH) procedure controls the FDR at $\pi_0\alpha$ level, where α is a user-defined confidence level. The EPVs have been successfully used in recent methods for outlier detection [25, 26] and in classification [22, 27] under the assumption of exchangeability.

Unfortunately, the validity of p-values and EPVs can be compromised by various problems including data distribution shift (a.k.a. covariate shift), data label distribution shift, batch effects [29] caused by some confounding factors, and overfitting. As a consequence multiple testing procedures that rely on valid p-values may fail to control the relevant statistical errors. In natural sciences, this shift can be attributed including, but not limited, to alterations in laboratory or environmental conditions, change in the reagents or atmospheric ozone levels, as well as to personnel preferences and to variation of technical sources [30]. Various high-throughput instruments produced by different manufacturers may also produce slightly differently distributed data about the same experiments. Quantitative morphological features of cells from multi-channel fluorescence microscopy images may depend on lamp intensities, filter patterns, and dyes [31]. Tandem mass spectrometry data are also affected by differences in experimental protocols and data acquisition conditions, reagent batches, or changes in instrumentation [32, 33]. These impacts may lead to incorrect conclusions about the experiments and to improper clinical drug therapies, etc. Similarly in economics and industry, machine learning-based predictions may also become inaccurate after economic shocks, technological advancement, geopolitical turbulence, and pandemic [34]. For instance, energy consumption prediction needs to be recalibrated after energy price change in order to maintain carbon emission reasonably low [35]; credit risk assessment needs to be adapted to monetary policy tightening, increased government expenditure, and income distribution shift caused by inflation [36, 37, 38].

Statistically speaking, the data shift violates the exchangeability assumption resulting in different test and training null hypotheses ($T_0 \not\approx X_0$) that further results in invalid p-values and conservatively or liberally (anti-conservatively) biased FDR control. Recent methods, which deal with data shift, aim to (1) detect data shift [39], (2) identify features in tabular data, which cause data shift [40], (3) detect and correct label shift [41], (4) provide explainability about data shift [42, 43], and (5) aim to adapt to data shifts [44, 45, 46], respectively. Most of these methods operate in the feature space and rely on data clustering or graphical causal models. They are also challenged by high-dimensional space, data sparsity (i.e. curse-of-dimensionality) [47], and correlated features.

In this paper, we propose a method to calculate EPVs for raw test scores under data or class distribution shift in general classification problems. These EPVs can be used in any statistical applications which rely on valid p-values. We demonstrate the usability of our method in FDR control with BH procedure. Our method, first, approximates the unknown test null distribution T_0 relying on the known training null and non-null distributions. Then, our method adjusts the test prediction scores so that the adjusted test null distribution $\hat{T}_0$ aligns well to the training null distribution; that is, $\hat{T}_0 \approx X_0$. Hence, we termed our method Test Null Adjustment (TNA). Figure 1 illustrates this idea. The empirical p-values of the adjusted test scores can be calculated with respect to the training null distribution X_0 with eq [1]. The test null adjusted empirical p-values (TNA EPVs) can then be used by the user in any statistical methods that require valid p-values, including π_0 estimation, or FDR control with any procedure. Valid p-values should be uniformly distributed over the range of $[0, 1]$. We propose to visually monitor the uniformity of the p-values with Q-Q plots, in which the p-values are plotted against their ordered normalized positions. Any deviation from the diagonal line of the Q-Q plot would indicate a data shift.

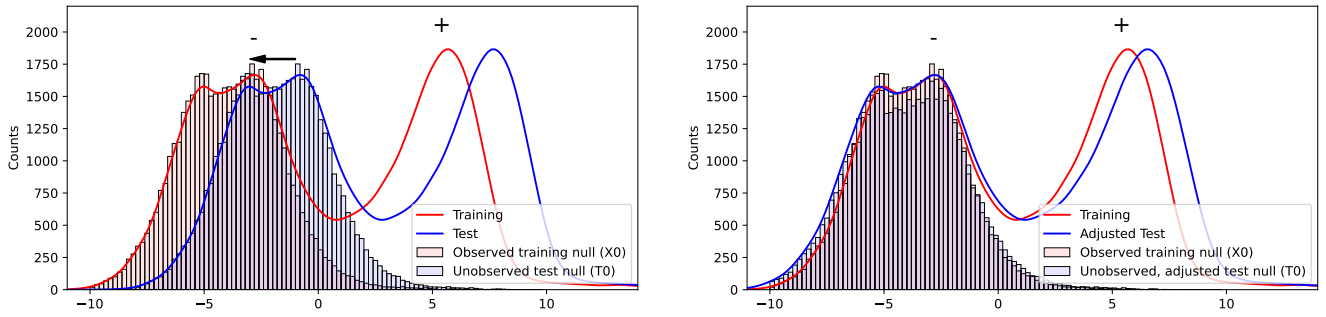

Figure 1: **Sketch of TNA.** (A) The plot illustrates the distribution shift of the discriminative scores of the test (blue) instances from the training (red) instances. The shaded areas illustrate the observed training and unobserved test null distributions, respectively. The black arrow indicates that the test scores should be shifted to leftward. The TNA method aims to approximate the unknown test null distribution (blue shaded area) based on the known true training null distribution (red shaded area). (B) TNA aims to perform a test score adjustment so that the test null aligns to training null distributions better.

The TNA method operates solely at the space of prediction score, which is typically one dimensional for each class. Therefore, TNA is not hindered by the curse-of-dimensionality [47]. It can be employed with any machine learning model, given that the classification is based on discriminative scores. This includes deep neural networks, black box models, support vector machines [48], with any kind of data such as tabular, text, audio, image, graph data. The TNA method is data driven as it does not rely on any class of analytical distributions. The only assumption we have for the TNA method is that the score distributions "somewhat resemble" the ones in Figure 2A or in Figure 10 that are typical.

2 Test null adjustment (TNA) method

2.1 Preliminaries

Let us suppose that we are given domain D such as images, text, speech, graphs, time series, tabular data, or the mix of these; furthermore, let $X \subseteq D$ and $T \subseteq D$ denote the training and the test data samples, respectively. Let $Y_t \in \{0, 1\}$ denote the true class label of sample $t \in D$. Let $f : D \to R$ be an appropriately trained binary discriminator (classifier), such as a deep neural network, to predict positive instances, where R denotes the real numbers. Without loss of generality, we assume that higher score $f(t)$ indicates a stronger membership to the positive class. The training null distribution is constructed as $X_0 = \{f(t) : Y_t = 0,\ t \in X\}$ from the scores of negative instances, the training target distribution is constructed as $X_1 = \{f(t) : Y_t = 1,\ t \in X\}$ from the scores of positive instances, and the training score distribution is $X_s = \{f(t) : t \in X\}$. The true test null T_0, true test target T_1, and the test score T_s distributions are constructed in a similar way using true labels. We say that we trust (or accept) a sample t predicted as positive if and only if $f(t) > s_t$ for a given decision threshold s_t and we make indecision for samples if $f(t) < s_t$. The false discovery proportion (FDP) in a set of samples $K = \{t \in D\}$ at a score threshold s_t is calculated as

$$FDP(s_t, K) = \frac{1 + |\{t \in K : Y_t = 0,\ f(t) \geq s_t\}|}{|\{t \in K : f(t) \geq s_t\}|},$$

where $|.|$ denotes set cardinality. We reserve the FDR for the statistical expectation of FDP. FDR control methods aim to calibrate the threshold s_t so that the error rate among trusted predictions is at most level of α.

The empirical p-values (EPVs) of the test samples are calculated using eq. 1 with respect to X_0. The p-value calculation can be time consuming for large X_0 sets, in this case, a binned, empirical cumulative distribution function of X_0 could be used to calculate the p-values accordingly in an accelerated way. Then, the FDR of a set T can be controlled using the BH protocol and the EPVs at a predetermined α level. For the sake of simplicity, we replace the α with $\alpha/\hat{\pi_0}$ inside BH algorithm, where $\hat{\pi_0}$ is an estimation of π_0 introduced by Storey [49] in the following way: $\hat{\pi_0} = (1 + \sum_{i=1}^{m} \mathbb{1}\{p_i \geq \lambda_S\})/(m(1 - \lambda_S))$ for $\lambda_S \in (0, 1)$. The q-value of a test sample $t \in T$ is defined as $q_t = \min_{z \leq f(t)} FDP(z, T)$; that is the minimum α level when the sample t would become a trusted positive prediction. The smaller the q-value of a test sample t is, the more trusted positive prediction t is. Note that, the q-value of a sample depends not only on its prediction score, but also on the other samples in the set.

Learn-then-test (LTT). We acquired the LTT code from its authors' GitHub repository. We specified the following hyperparameters as following: 100 uniformly distributed queried levels of FDR control are iteratively being processed, starting from the 1e-3, with the most appropriate threshold λ_L being selected among the 300 potential threshold values for

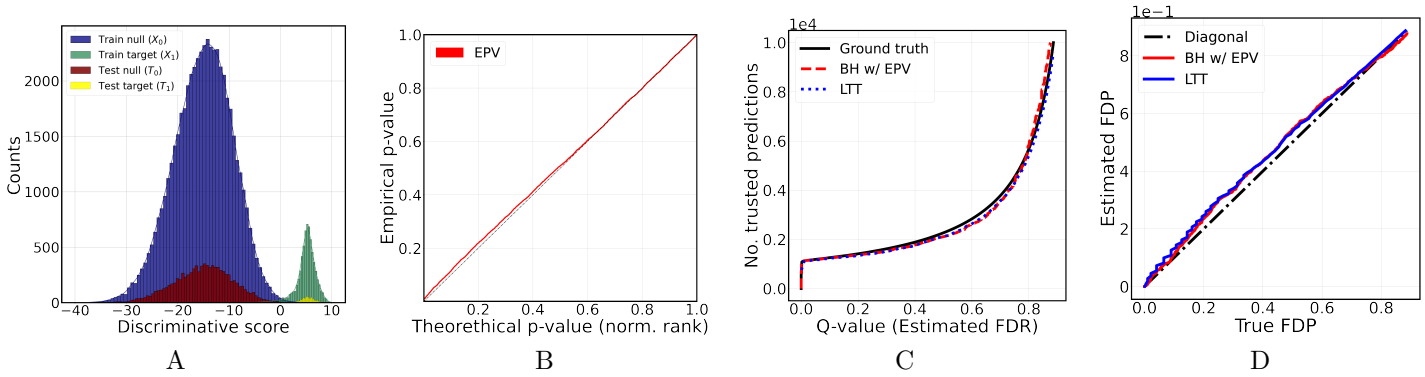

Figure 2: **FDR control in MNIST with EPVs.** (A) The discriminative score distributions. (B) The Q-Q plot of the EPVs of the test samples against the theoretical uniform distribution. (C) The number of the accepted (trusted) positive predictions at various q-values calculated with true labels (black solid line), BH with EPVs (red dashed line), and LTT (blue dotted line). (D) Deviation between the true and the estimated FDR produced by BH with EPV (red line) and LTT (blue dots), respectively.

each iteration. The remainder of the methodology is entirely dependent on the published pipeline, including the comprehensive calculation of the p-values for each threshold during each alpha level test, with a further Bonferroni procedure employed to select the most appropriate border. The final value of the test false positive rate is equal to the number of samples with higher scores than the selected λ_L.

Example 1. We used the well-known MNIST dataset for demonstration. It contains 60K training and 10K test images of 28x28 grey-scaled pixels in 10 classes. There is no data or label shift between the training and test data. We trained a convolutional neural network to classify the digit '1' as target (positive) against all the other digits (negative) (binary classification for the sake of simplicity). Our ConvNet model contained two convolutional layers consisting of 8 kernels with size of 3×3 each, and it was trained with an Adam optimizer with a learning rate of 0.01 for three epochs. Our trained model achieved a 99% accuracy on the test set.

The relevant results are presented in Figure 2. Figure 2A shows the training and test null and target distributions. Figure 2B shows a Q-Q plot of the test EPVs against the theoretical uniform distribution; that is the scatter plot of the EPV against its normalized rank in our case. The dots line up along the diagonal line which shows that the EPVs calculated are uniformly distributed and, consequently, valid. This also indicates that there is no shift between the training and test distributions. Figure 2C shows the number of accepted (trusted) predictions as positives at various q-values (α level of FDR). The ground truth (black solid line) was calculated using the true test labels. Any FDR controlling method (a) yielding more trusted predictions than the ground truth is implied to be liberally biased, and (b) yielding less trusted predictions is conservatively biased. The BH procedure with EPVs (red dashed line) yields closely the same number of trusted predictions at various FDR levels of α (calculated without using the true test labels). This repeatedly indicates that the EPVs are uniform and the BH procedure is an unbiased method. The LTT method also accurately controls the FDR at any α levels (blue dotted line). Finally, Figure 2D shows that the deviation of the estimated FDPs from the true FDPs is small.

TNA protocol. When the training null distribution X_0 is different from the test null distribution T_0, the EPVs will not be uniformly distributed and the BH procedure results in liberal or conservative bias in FDR estimation. In the following steps, we describe two versions of TNA methods that both adjusts the prediction scores so that the approximated test null distribution $\hat{T}_0$ aligns to the actual training null distribution X_0. The first variation, called TNA- (TNA minus), relies on that the shapes of the training null (X_0) and test null (T_0) distributions are similar, but there is no assumption on the form of the non-null distributions. The second variation, called TNA+ (TNA plus), relies on that the shapes of the training non-null (X_1) and test non-null (T_1) distributions are "somewhat similar", but there is no assumption on the form of the null distributions. The proportion of the nulls may freely vary between the training and test data in the case of both TNA methods, there is no assumption on this. The two approaches differ only at the 3rd step of the following algorithm.

1. Get the density histograms of X_0, X_1, and T_s distributions with identical bin borders and denote them as $\overline{X}_0$, $\overline{X}_1$, and $\overline{T}_s$, respectively. The $\overline{X}[i]$ denotes the value of bin i. We used 100 bins in our experiments.
2. Let $\hat{\pi}_{0_T}$ be an estimation of the proportion of the negative test predictions calculated as follows: $\hat{\pi}_{0_T} = N/(N+P)$, where N and P denotes the total number of the negative and positive predictions in the given test data, respectively, as predicted by the classifier f. We note that this $\hat{\pi}_{0_T}$ estimation can be done with raw discriminative scores and it does not require valid p-values.

3a. TNA-: We create a new density histogram $\hat{T}_0$ that will be an approximation of the test null T_0. The value of the histogram bin i is calculated by $\hat{T}_0[i] = \overline{T}_s[i] \cdot c[i]$, where $c[i] = X_0[i]/X_s[i] \cdot \pi_{0_X}/\hat{\pi}_{0_T}$ and π_{0_X} is the actual proportion

of the null samples in the training data calculated with using the true class labels.

3b. TNA+: This approach relies on the following formulation: $X_s = (\pi_0) \cdot X_0 + (1 - \pi_0) \cdot X_1$, where π_0 is the proportion of the negative samples; therefore, the null distribution can be expressed as $X_0 = (X_s - (1 - \pi_0) \cdot X_1)/\pi_0$. We create a new density histogram $\hat{T}_0$ that will be an approximation of the test null T_0. The value of the histogram bin i is calculated by $\hat{T}_0[i] = (\overline{T}_s[i] - (1 - \hat{\pi}_{0_T})\hat{T}_1[i])/\hat{\pi}_{0_T}$. The unknown $\hat{T}_1$ is approximated by T_s scaled with the proportion of the target calculated from the training data; that is:

$$\hat{T}_1[i] = \overline{T_s}[i] \cdot \frac{\overline{X_1}[i]}{\hat{\pi}_{0_T}\overline{X_0}[i] + (1 - \hat{\pi}_{0_T})\overline{X_1}[i]}.$$

Mind that the training null and target distributions (X_0, X_1) are weighted with test class proportions $\hat{\pi}_{0_T}$. The outcome of division by zero is set to zero.

4. Let μ_{X_0}, σ_{X_0}, $\mu_{\hat{T}_0}$, and $\sigma_{\hat{T}_0}$ are the mean and std of the X_0 and the $\hat{T}_0$ distributions, respectively.
5. Adjust the test scores $\hat{t_s} = (t_s - \mu_{\hat{T}_0})/\sigma_{\hat{T}_0} \cdot \sigma_{\hat{X}_0} + \mu_{X_0}$.

Finally, calculate the TNA EPVs for each adjusted $\hat{t_s}$ test score with respect to the training null (X_0) distribution with using Eq. [1]. The TNA EPVs then can be used in any statistical applications that rely on valid p-values. In our experiments in the following sections, we demonstrate the usability of the TNA EPVs in FDR control with BH procedure. With TNA EPVs, we re-estimate $\hat{\pi}_0$ for the test instances with the Storey method with $\lambda_S = 0.8$, and run BH procedure to control the FDR at a desired level.

We note that the actual number of bins or the size of bin-width in step 1 may depend on the range and/or the number of the discriminative scores. Essentially, there should be enough data in each bins to ensure accurate estimations. This might require the the analyst to manually adjust the bins in practice.

We also note that we tested a one-dimensional optimal transportation (OT) to adjust training null to test null. Unfortunately, this approach is very sensitive to the test null approximation especially at the decision boundary. In our experiment, OT severely overfitted, while TNA remained more robust to poor tail approximation of the null distributions. OT did not provide any good results (data not shown) in our benchmark.

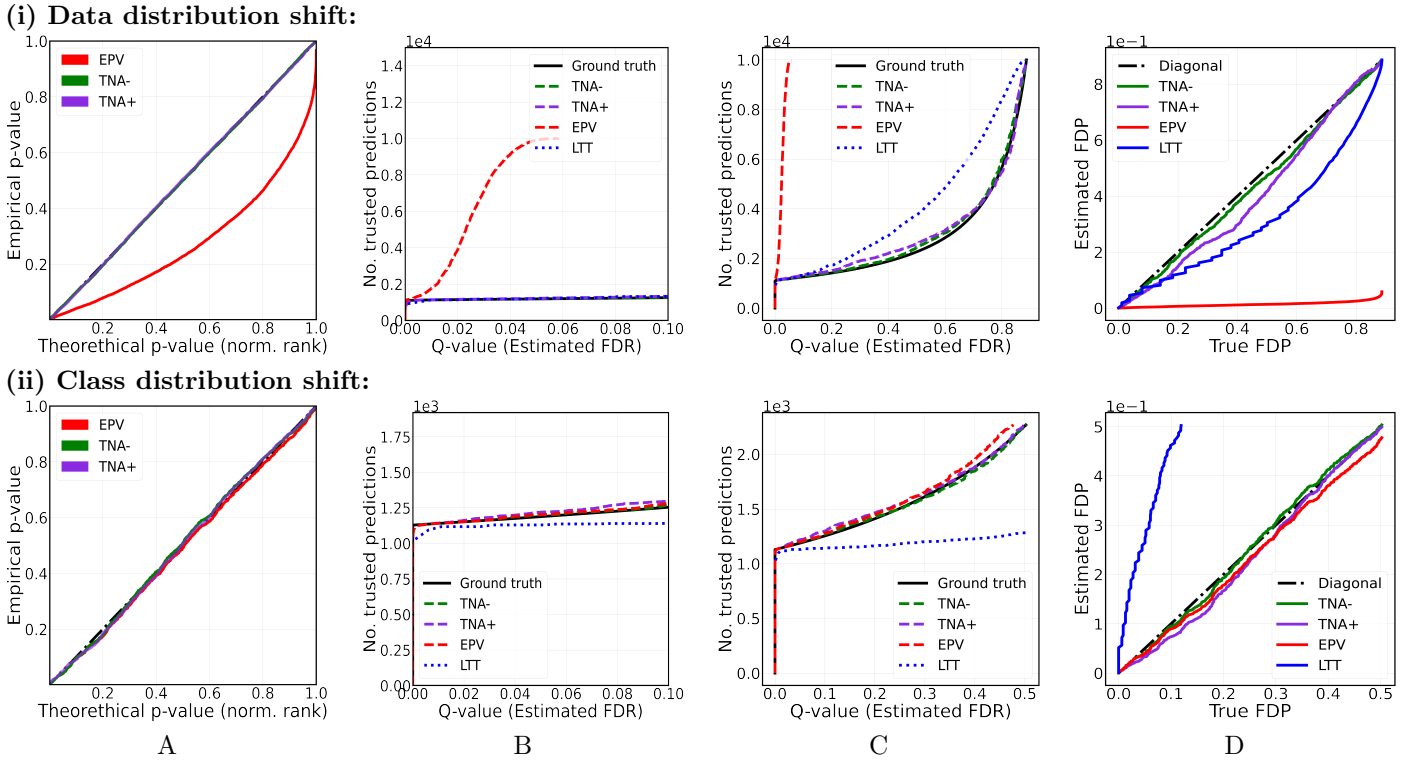

Figure 3: **FDR control under data (top row) and label (bottom row) shift.** (A) Q-Q plot of the EPVs (red dots), the TNA- EPVs (green dots) and the TNA+ EPVs (purple dots) against the theoretical uniform distribution (normalized rank). (B) The number of accepted classifications as a function of the Q-values over a critical range (0-0.1) obtained with (i) ground truth (black line), (ii) BH with EPV (red), (iii) BH with TNA- EPV (green), (iv) BH with TNA- EPV (purple), and (v) LTT (blue). (C) Same as (B) but over the entire q-value range. (D) Deviation of the estimated FDP from the true FDP obtained with (i) BH with EPV (red line), (ii) BH with TNA- EPV (green line), (iii) BH with TNA+ EPV (purple line), and (iv) LTT (blue line).

Example 2. To illustrate the case of data distribution shift, we down-scaled the pixel intensity by 10% of the test images in the MNIST dataset (but training images remained the same). The accuracy of the CNN classifier (the same as trained in Example 2.1) remained at 99 %. The evaluation plots are shown in Figure 3(i). The Q-Q plot (Fig. 3(i)A) shows that the test EPVs (red) became biased due to data distribution shift; however, both TNA methods can correct the EPVs yielding uniformly distributed p-values (green, purple). The number of accepted predictions ((Fig. 3(i)BC) obtained with BH with EPVs (red) differs from the ground truth substantially (black line), it became liberally biased. The LTT yielded similar amount of accepted predictions (blue) compared to the ground truth in the critical range of q-values ((Fig. 3(i)B); however, LTT also becomes liberally biased at lower ranges of q-values ((Fig. 3(i)C). The number of accepted predictions obtained with BH with both TNA EPVs (green, purple) remained close to the ground truth over both the critical and lower ranges of q-values implying accurate FDR control under data distribution shift. The actual deviation of the estimated FDP from the true FDP is shown in Figure 3(i)D for each method. The TNA+ is slightly less accurate than TNA- in this test scenario.

Example 3. The FDR estimation is sensitive to changes in class proportions. To illustrate the case of class distribution shift, we resampled the test data so that the number of true negative and positive instances became equal. This approach was used in other articles, e.g. [50]. The accuracy of the CNN classifier (the same as trained in Example 2.1) remained at 99 %. The evaluation plots are shown in Figure 3(ii). The Q-Q plot (Fig. 3(ii)A) shows that the EPVs (red) remained unbiased and so did TNA EPVs (green, purple). This is in fact expected, because the shift in class distributions should not affect the actual null distributions; Therefore, the results of BH-based methods rely on the $\hat{\pi_0}$ estimation. BH with EPV and TNA EPV remained accurate and produced results close to the ground truth (red, green, and purple lines in Fig 3(ii)BCD) indicating successful $\hat{\pi_0}$ estimation. The BH is robust to changes in class proportions because it can involve the (re-) estimation of the negative data proportion ($\hat{\pi_0}$) that can be done with, e.g. Storey method. The LTT is also sensitive to changes in class proportions because it relies on the class proportions that are calculated with the training (validation) datasets, and it assumes that class proportions remain unchanged in test phase. Here, LTT resulted in conservative FDR control even at the critical range of q-values (blue in Fig. 3(ii)BC). The bias in FDR produced by LTT (blue) is also visible in Figure 3(ii)D.

3 Experimental results

3.1 PCam: tumor classification

The PatchCamelyon (PCam) benchmark dataset is a collection of histopathologic scans of lymph node sections [51] derived from the Camelyon 16 challenge [52]. PCam consists of 327,680 colored images of size 96x96 each. An arbitrarily chosen sample data is shown in Figure 4. The challenge is to identify the presence (positive) or absence (negative) of any histopathology in one image. The dataset creators of PCam underline that all the positive and negative instances in the training, validation, test sets are equally distributed, resulting in a class balance of 1:1. We used the pre-trained resnet34-pcam model, which is a popular, deep residual network from the TIA (Tissue Image Analysis) toolbox [53] with an overall f1-score of 0.889, as stated by the authors. We evaluated the uniformity of the p-values of the test samples and the performance of the FDR controlling methods. The results are shown in the first row of Figure 5. The Q-Q plots (Fig. 5-(i)-A) of the test EPVs (red dots) indicate that the test p-values are slightly biased, possibly indicating a slight overfitting and/or distribution shift between the training and test datasets. However, the Q-Q plots of the test EPVs adjusted with TNA- (green dots) and TNA+ (purple dots) show a more uniformly distributed p-values. The plots on Fig. 5-(i)-BCD indicate that all four FDR control methods are mostly accurate with this dataset especially at critical α levels (0-0.1), despite the small bias in the EPVs. The TNA methods (both, TNA- and TNA+) manage to reduce the bias from EPVs and yield more accurate FDR control with BH (green and purple lines), than the BH method with the original EPVs (red line). However, LTT outperformed both BH-based methods in this benchmark.

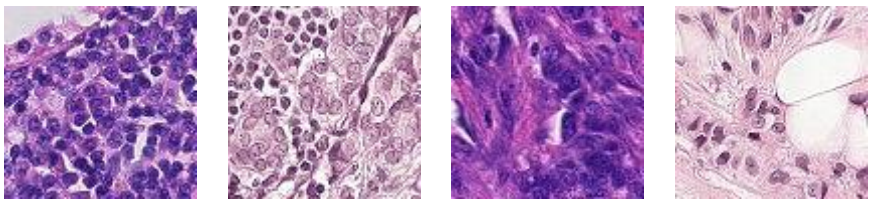

Figure 4: **Instances of PCam dataset.**

In our opinion, the actual class proportions in real-life applications can be significantly different from 1:1, making the FDR control with methods like LTT inaccurate in practical applications. We manually resampled the test data so that the positive and negative class balance became 31%:69%, following the original class balance of the Camelyon 16 dataset. The evaluation plots are shown in the second row of Figure 5. The Q-Q plot (Fig. 5-(ii)-A) reveals that the EPVs (red dots) become slightly biased; however, our TNA methods manage to reduce this bias from the EPVs (green dots). The LTT

method became liberal in FDR control, the BH protocol with standard EPVs resulted in conservative FDR control; whereas, the BH with the TNA EPVs remained accurate, especially at critical range of q-values (0-0.1) (Fig. 5-(ii)-C).

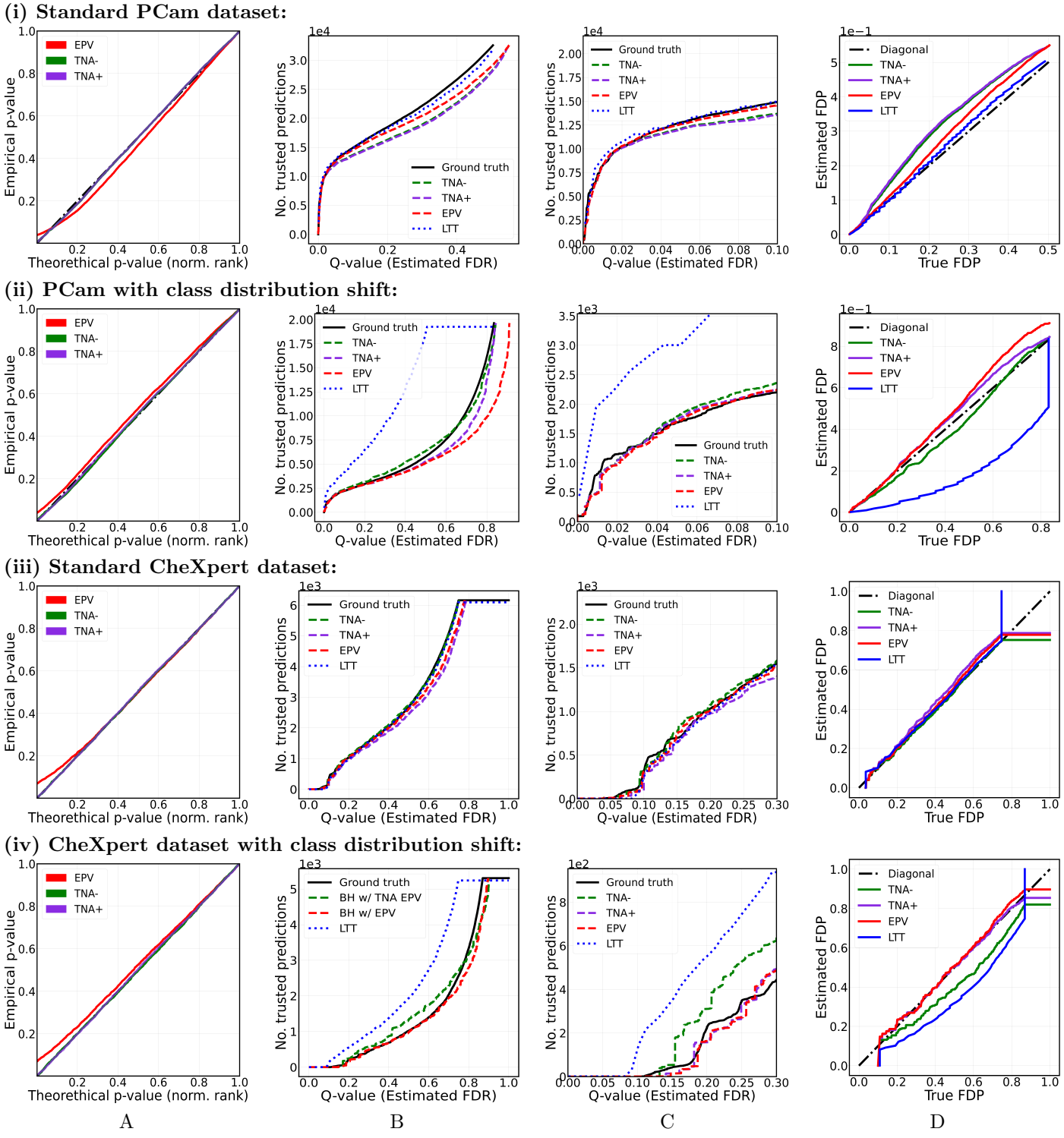

Figure 5: **Performance evaluation of FDR controlling methods.** See the caption of Figure 3 for plot descriptions.

3.2 CheXpert: chest x-ray dataset

The CheXpert dataset contains a total of 224,316 radiographs of both frontal and lateral views of the thorax from ∼65K patients in order to facilitate the automated interpretation of chest X-rays for multi-label classification tasks (i.e. multiple labels can be predicted for each instances) with a total of 14 labels. An arbitrarily chosen sample data is shown in Figure 6. We used the deep neural network model, called DenseNet121, from the `Torchxrayvision` package [54, 55], specifically developed for this CheXpert dataset. For the sake of simplicity, we turned the CheXpert dataset into a binary classification

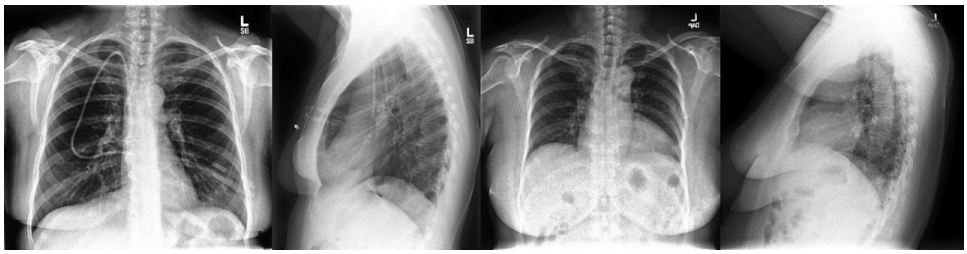

Figure 6: **Instances of CheXpert dataset.**

task by treating all instances of pleural effusion as positive and all the others as negative. We selected this label because this was the most prevalent disease in the dataset. The proportion of the positive class became 27 % and 23 % in the training and test datasets, respectively. The DenseNet121 model achieved accuracies of 89 % and 84 % on the training and test sets, respectively.

We evaluated the four FDR controlling methods and plotted the results on Figure 5-(iii). Interestingly, the most confident prediction of the DenseNet121 model in the test set is wrong. This can be seen by the flat region of the ground truth evaluation (black line) at low q-values around 0-0.05 (Fig, 5-(iii)-C). Fortunately, all four FDR controlling methods under investigations managed to capture this and remain close to the ground truth (Fig, 5-(iii)-B-D). The standard EPVs (red dots) tend to be off at the critical range (p-value < 0.1), but this does not affect the FDR controlling results in this benchmark, perhaps because of the confident incorrect predictions. Overall, we conclude that all four methods provide accurate FDR control. In our opinion, this indicates the absence of any database distribution shits. These plots also show that our TNA- and TNA+ algorithms do not corrupt the EPVs in the absence of data shift.

The actual frequency of pleural effusion may be different among all chest x-ray radiographs performed. It may vary over locations, age groups, in-patients and out-patients [56, 57]. In the United States, 275 million conventional radiology procedures are performed annually [58], but only around 1.5 million of US citizens develop pleural effusion [57]. Pleural effusion develops in about 20-40% of in-patient with pneumonia [59]. Whatever the real abundance of the pleural effusion is among chest x-rays performed, it can be significantly different from the 27 % of the CheXpert dataset. Taking this into account, we re-sampled the test set in the CheXpert dataset to achieve a lower positive class ratio of 13%. The results of the re-evaluation are shown in Figure 5-(iv). The LTT and the TNA- EPVs are both slightly biased. The BH with the standard EVPs and with TNA+ EPVs achieve the most accurate FDR control over the entire range of α levels (0-1) in general. The LTT method is very sensitive to class distribution shift demonstrated by this case, wheree it resulted in a liberally biased FDR control.

3.3 TissueNet: cells segmentation

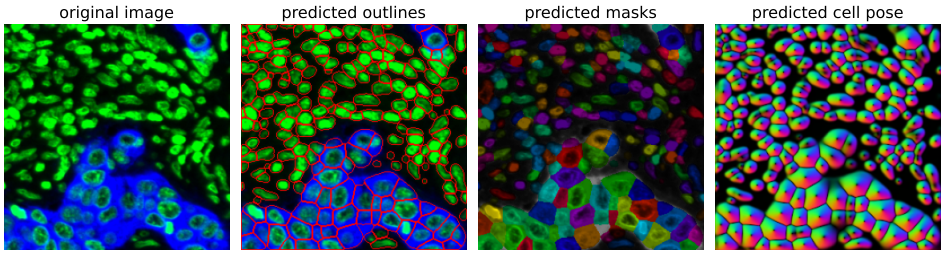

Figure 7: **Instance of TissueNet dataset.** Two-channel original images are being processed by the model, in turn producing masks & flows to correctly segment cells

TissueNet dataset contains 1.3 million microscopic images of cells obtained from six various platforms and nine organs including both histologically healthy and diseased tissues. Each images has size of 512x512 pixels, was manually segmented at pixel level and paired whole-cell and nuclear annotations. An arbitrarily chosen sample data is shown in Figure 7. Our binary classification task was to predict whether a pixel is inside a cell (positive) or not (negative). The overall positive-negative ratio was 59%:41% in the training data and in contrast to 66%:34% in the test data. However, the class ratio varies over the images considerably. We used the Cellpose pre-trained deep neural network model [60] and fine-tuned it for five epochs.

It achieved accuracy of 95% and 94% on training and test data, respectively. Our evaluation is reported in Figure 9-(i). The plots reveal that the BH protocol with standard EPVs becomes liberally biased (Fig. 9-(i)-B-C). The latter possibly originates from the data distribution shift or from overfitting. Both LTT and the BH with TNA EPV remain accurate.

The situation changes when we evaluate the segmentation problem with a single image. We arbitrarily chose one test image (ID: 007-008-009) and run the Cellpose model to classify each pixel. The positive and negative class balance was: 33%:67%. The Q-Q plot (Fig. 9-(ii)-A) of the test EPVs (red dots) indicates that the scores of negative samples from training and test datasets are different. The TNA EPVs on the Q-Q plot shows that the TNA methods have managed to adjust the p-values of the test samples (green and purple dots) closer to the uniform distribution. The plots on Fig. 9-(ii)-B-D indicate that LTT has turned out to be liberally biased, the BH method with standard EPVs strongly conservative. The BH with TNA- EPVs (green) and with TNA+ EPVs (purple) remain accurate in the critical region of q-values (0-0.4). Thus, the TNA methods have succeeded in correcting the p-values at the critical range at least.

3.4 BCSS: Breast cancer semantic segmentation dataset

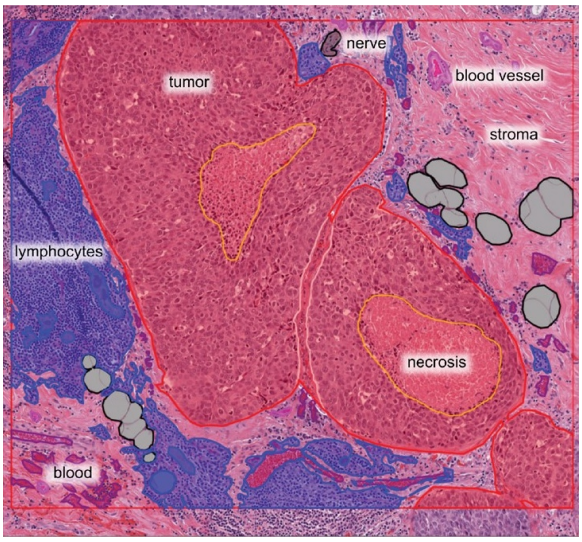

Figure 8: **An instance from BCSS dataset.**

The Breast Cancer Semantic Segmentation (BCSS) dataset [61] contains around 20K image segmentation objects annotated through Digital Slide Archive, with a particular focus on tissue regions derived from breast cancer imagery obtained from The Cancer Genome Atlas (TCGA). The image pixels are labeled by: 'Stroma' (class id: 0), 'Tumor' (class id: 1) and 'Others' (class id: 2). An arbitrarily chosen sample data is shown in Figure 8. The corresponding class ratios in the training and test sets are 56%:38%:6% and 33%:59%:8%, respectively. We used the deep learning model called `'fcn resnet50 unet-bcss'` from the TIA (Tissue Image Analysis) toolbox [53] to classify the pixels to one of the three classes. The model achieved an accuracy of 84 % and 82 % on the training and test datasets, respectively.

Since this dataset involves a multi-class classification problem, we slightly modified the FDR control procedure. We calculated the p-values, the π_0 estimations with Storey method with respect to a single class (deemed as positive); thus, for each pixels, we got three p-values, each corresponding to one of the three classes. Then we concatenated all of these p-values and calculated the q-values. We note that, we did not evaluate LTT on this benchmark, because it is not clear how to adjust it to multi-class classification problems.

The evaluation of the FDR controlling methods is shown in Figure 9-(iii). At first glance, the results suggest that this datasets is challenging and it has a strong shift between the training and test data and/or the deep learning model has been profoundly overfit. The Q-Q plots of the EPVs and the TNA EPVs are far from being uniform (Fig. 9-(iii)-A). Overall, the BH control with standard and TNA EPVs is conservative (Fig. 9-(iii)-B), but it is slightly liberal over the critical range of q-values (0-0.1) (Fig. 9-(iii)-C). BH protocol seems to be more accurate with standard rather than with TNA EPVs. The FDR estimation error is significant (Fig. 9-(iii)-D).

Next, we further analyzed the results by each class separately. We plotted the training and test positive and negative discriminative score distributions class-wise provided by the deep learning model (Fig. 10 first row). The distribution plots show that the test scores are differently distributed from the training scores in all the classes. We also plotted the distribution of the adjusted scores of the negative test samples, adjusted by our TNA methods (green and purple lines in Fig. 10) in order to see how well they are aligned to the training null distribution. In the case of the Stroma class, the mode of adjusted test scores (green and purple lines) move a bit closer to the one of the true training null distribution (X0, filled orange bars). However, their tails on the right hand side remain off. In the other two classes, the adjusted test scores align visibly better to the training null distribution (X0). The corresponding FDR curves are shown in the second row of Figure 10. Both the

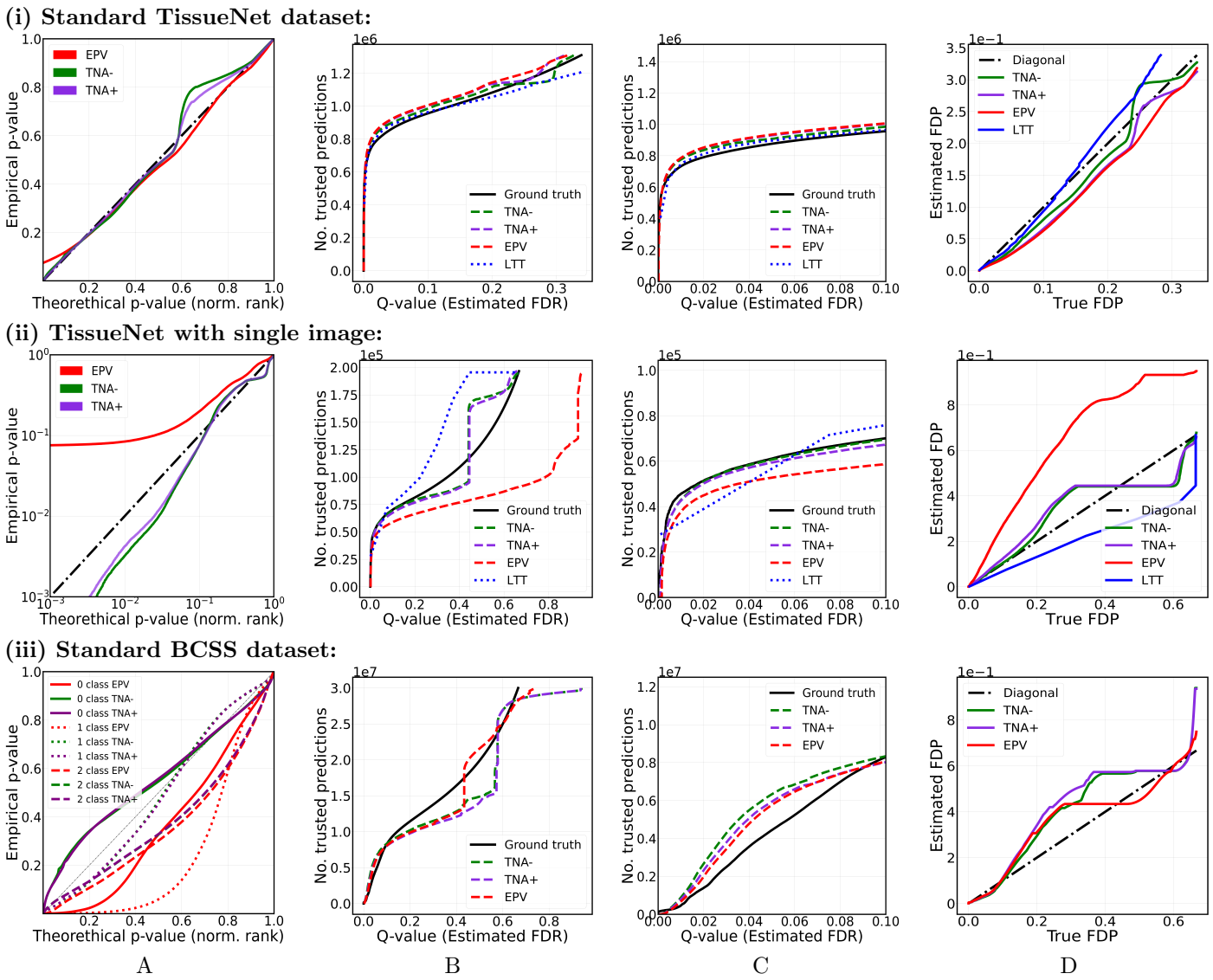

Figure 9: **Performance evaluation of FDR controlling methods.** See the caption of Figure 3 for plot descriptions.

TNA- and the TNA+ methods worsen the results in the case of the Stroma class; however, they make the FDR control more accurate in the other two classes. This is mainly due to the fact that the Storey method provides better $\hat{\pi_0}$ estimation using the TNA EPVs tha using the standard EPVs.

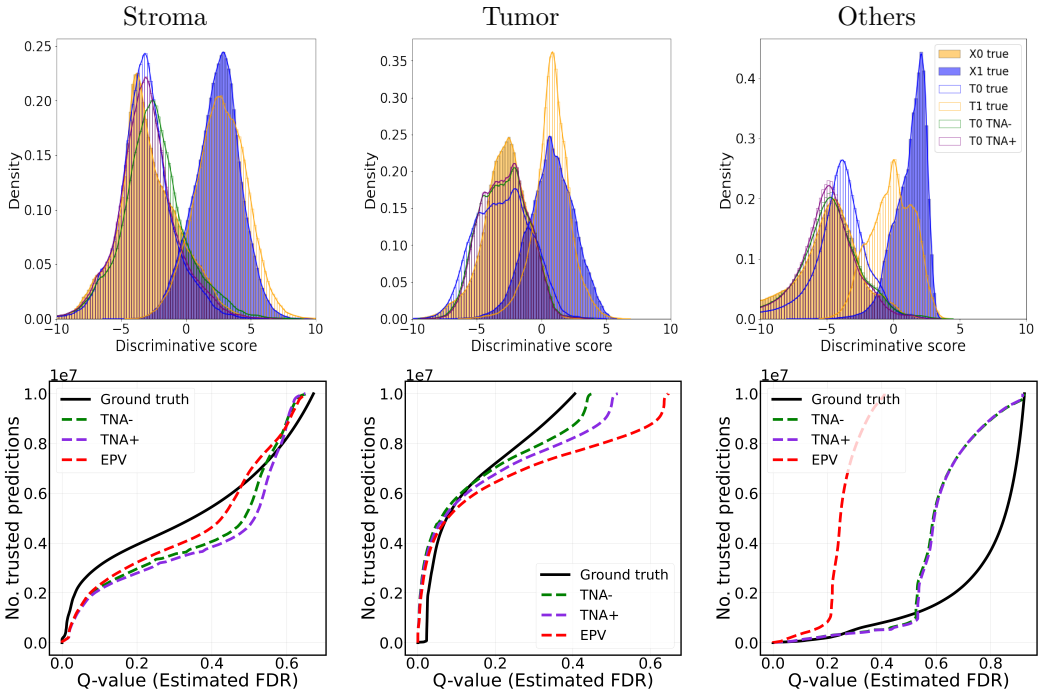

Figure 10: **Classwise score distributions and FDR control in the BCSS dataset.**

4 Conclusions

In this study, we examined the effects of data shift on False Discovery Rate (FDR) control. In our opinion, the BH procedure with standard EPVs provides a very accurate FDR control in the absence of any data shift. Unfortunately, data shift is rather common in practice, and we demonstrated that the standard BH and the LTT methods are quite sensitive even to small data or class distribution shift. In this paper, we presented a simple, but robust method, called Test Null Adjustment (TNA), in order to mitigate the effect of the data shift on the FDR control. Our method operates in the space of prediction scores, thereby it is not affected by the curse-of-dimensionality problem, which often hinders methods that aim to mitigate data shift in the input feature space. The TNA method consists of two main steps. First, it approximates the null distribution among the scores of the test instances via certain statistics from the scores of the training instances. Second, it recalibrates the test scores so that the test and training null distributions would move closer to each other. TNA is fully data-driven, it does not rely on any theoretical assumptions or forms of the data. We have demonstrated in our experimental test with four biomedical datasets that TNA can successfully mitigate the bias caused by data shift; while, leaving the distributions intact in the absence of data shift. We note that, the handling of data shifts in practice is challenging, and unfortunately our method does not provide any theoretical guarantee that the FDR control with TNA EPVs becomes accurate, or conservatively or liberally biased without any additional assumptions; hence TNA is a heuristic in its general form.

With this study, we also aim to emphasize and popularize the importance of accurate error control in any predictions in the biomedical domain. We have shown that the deviations in the simple Q-Q plots of the calculated vs. the theoretical p-values can be used to detect data shifts between the training and test data. Moreover, we also hope that the FDR control and plots with q-values will become a standard metric in performance evaluation for prediction systems in the biomedical domain.

References

[1] Jean Feng, Rachael V Phillips, Ivana Malenica, Andrew Bishara, Alan E Hubbard, Leo A Celi, and Romain Pirracchio. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of ai algorithms in healthcare. *npj Digital Medicine*, 5(1):66, 2022.

[2] Mugahed A Al-Antari. Artificial intelligence for medical diagnostics—existing and future ai technology!, 2023.

[3] Laure Wynants, Maarten Van Smeden, David J McLernon, Dirk Timmerman, Ewout W Steyerberg, Ben Van Calster, Topic Group 'Evaluating diagnostic tests, and prediction models' of the STRATOS initiative. Three myths about risk thresholds for prediction models. *BMC medicine*, 17:1–7, 2019.

[4] David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "big three". *Diagnosis*, 8(1):67–84, 2021.

[5] Talya Salz, Alice R Richman, and Noel T Brewer. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-Oncology*, 19(10):1026–1034, 2010.

[6] Anna N. A. Tosteson, Dennis G. Fryback, Cristina S. Hammond, Lucy G. Hanna, Margaret R. Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D. Pisano. Consequences of False-Positive Screening Mammograms. *JAMA Internal Medicine*, 174(6):954–961, 06 2014.

[7] Ying Xu, Dennis Nguyen, Armin Mohamed, Cheryl Carcel, Qiang Li, Mansur A Kutlubaev, Craig S Anderson, and Maree L Hackett. Frequency of a false positive diagnosis of epilepsy: a systematic review of observational studies. *Seizure*, 41:167–174, 2016.

[8] Dimitra S Mouliou and Konstantinos I Gourgoulianis. False-positive and false-negative covid-19 cases: respiratory prevention and management strategies, vaccination, and further perspectives. *Expert review of respiratory medicine*, 15(8):993–1002, 2021.

[9] Steven Woloshin, Neeraj Patel, and Aaron S Kesselheim. False negative tests for sars-cov-2 infection—challenges and implications. *New England Journal of Medicine*, 383(6):e38, 2020.

[10] Stephen H Bradley, Brian D Nicholson, and Garth Funston. Interpreting negative test results when assessing cancer risk in general practice. *British Journal of General Practice*, 71(708):298–299, 2021.

[11] Anna Macios and Andrzej Nowakowski. False negative results in cervical cancer screening—risks, reasons and implications for clinical practice and public health. *Diagnostics*, 12(6):1508, 2022.

[12] Paul F Pinsky. Principles of cancer screening. *Surgical Clinics*, 95(5):953–966, 2015.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[14] Krisztian Koos, Gáspár Oláh, Tamas Balassa, Norbert Mihut, Márton Rózsa, Attila Ozsvár, Ervin Tasnadi, Pál Barzó, Nóra Faragó, László Puskás, et al. Automatic deep learning-driven label-free image-guided patch clamp system. *Nature communications*, 12(1):1–11, 2021.

[15] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *Nature Communications*, 15(1):1594, 2024.

[16] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

[17] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

[18] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[19] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of statistics*, pages 2055–2085, 2015.

[20] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3):207, 2007.

[21] Guanlan Zhao and Zhonggen Su. Controlling fsr in selective classification. *arXiv preprint arXiv:2311.03811*, 2023.

[22] Bradley Rava, Wenguang Sun, Gareth M James, and Xin Tong. A burden shared is a burden halved: A fairness-adjusted approach to classification. *arXiv preprint arXiv:2110.05720*, 2021.

[23] Zijun Gao and Qingyuan Zhao. Simultaneous hypothesis testing using internal negative controls with an application to proteomics. *arXiv preprint arXiv:2303.01552*, 2023.

[24] Yulia Danilova, Anastasia Voronkova, Pavel Sulimov, and Attila Kertesz-Farkas. Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *Journal of Proteome Research*, 18(5):2354–2358, 2019.

[25] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

[26] Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee, 2023.

[27] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

[28] Chip Huyen. *Designing Machine Learning Systems*. O'Reilly Media, USA, 2022.

[29] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[30] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[31] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

[32] Ser-Xian Phua, Kai-Peng Lim, and Wilson Wen-Bin Goh. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Computational and Structural Biotechnology Journal*, 2022.

[33] Jelena Čuklina, Chloe H Lee, Evan G Williams, Tatjana Sajic, Ben C Collins, María Rodríguez Martínez, Varun S Sharma, Fabian Wendt, Sandra Goetze, Gregory R Keele, et al. Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial. *Molecular systems biology*, 17(8):e10240, 2021.

[34] Valerie A Ramey. Macroeconomic shocks and their propagation. *Handbook of macroeconomics*, 2:71–162, 2016.

[35] Tobias Clement, Hung Truong Thanh Nguyen, Nils Kemmerzell, Mohamed Abdelaal, and Davor Stjelja. Coping with data distribution shifts: Xai-based adaptive learning with shap clustering for energy consumption prediction. In *Australasian Joint Conference on Artificial Intelligence*, pages 147–159. Springer, 2023.

[36] Mark Kritzman, Sebastien Page, and David Turkington. Regime shifts: Implications for dynamic strategies (corrected). *Financial analysts journal*, 68(3):22–39, 2012.

[37] Yue Guo, Chenxi Hu, and Yi Yang. Predict the future from the past? on the temporal data distribution shift in financial sentiment classifications. *arXiv preprint arXiv:2310.12620*, 2023.

[38] Marvin Zhang. *Adaptation Based Approaches to Distribution Shift Problems*. PhD thesis, EECS Department, University of California, Berkeley, Dec 2021.

[39] Tamraparni Dasu, Shankar Krishnan, Dongyu Lin, Suresh Venkatasubramanian, and Kevin Yi. Change (detection) you can believe in: Finding distributional shifts in data streams. In *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, August 31-September 2, 2009. Proceedings 8*, pages 21–34. Springer, 2009.

[40] Sean Kulinski, Saurabh Bagchi, and David I Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in neural information processing systems*, 33:19523–19533, 2020.

[41] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[42] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In *International Conference on Artificial Intelligence and Statistics*, pages 1666–1674. PMLR, 2021.

[43] Sean Kulinski and David I Inouye. Towards explaining distribution shifts. In *International Conference on Machine Learning*, pages 17931–17952. PMLR, 2023.

[44] Yongduo Sui, Qitian Wu, Jiancan Wu, Qing Cui, Longfei Li, Jun Zhou, Xiang Wang, and Xiangnan He. Unleashing the power of graph data augmentation on covariate distribution shift. *Advances in Neural Information Processing Systems*, 36, 2024.

[45] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[46] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022.

[47] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

[48] Nello Cristianini. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[49] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205, 2004.

[50] Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Saliency is a possible red herring when diagnosing poor generalization. In *International Conference on Learning Representations*, 2021.

[51] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

[52] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017.

[53] Johnathan Pocock, Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Srijay Deshpande, Giorgos Hadjigeorghiou, Adam Shephard, Raja Muhammad Saad Bashir, Mohsin Bilal, Wenqi Lu, David Epstein, Fayyaz Minhas, Nasir M Rajpoot, and Shan E Ahmed Raza. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2(1):120, sep 2022.

[54] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020.

[55] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.

[56] Hany A. Zaki, Bilal Albaroudi, Eman E. Shaban, Ahmed Shaban, Mohamed Elgassim, Nood Dhafi Almarri, Kaleem Basharat, and Aftab Mohammad Azad. Advancement in pleura effusion diagnosis: a systematic review and meta-analysis of point-of-care ultrasound versus radiographic thoracic imaging. *The Ultrasound Journal*, 16(1):3, Jan 2024.

[57] Katherine Cashen and Tara L Petersen. Pleural effusions and pneumothoraces. *Pediatrics in Review*, 38:170 – 181, 2017.

[58] Mahadevappa Mahesh, Armin J Ansari, and Fred A Mettler Jr. Patient exposure from radiologic and nuclear medicine procedures in the united states and worldwide: 2009–2018. *Radiology*, 307(1):e221263, 2022.

[59] Eman Shebl and Manju Paul. Parapneumonic pleural effusions and empyema thoracis. 2018.

[60] Carsen Stringer, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. 02 2020.

[61] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A. Atteya, Mai A T. Elsebaie, Lamia S. Abo Elnasr, Rokia Adel Sakr, Hazem S. E. Salem, Ahmed F. Ismail, Anas M. Saad, Joumana Ahmed, Maha A. T. Elsebaie, Mustafijur Rahman, Inas A Ruhban, Nada M. Elgazar, Yahya Alagha, Mohamed Hosny Osman, Ahmed M. Alhusseiny, Mariam M. Khalaf, Abo-Alela F. Younes, Ali Abdulkarim, Duaa M. Younes, Ahmed M. Gadallah, Ahmad M. Elkashash, Salma Y. Fala, Basma Mostafa Zaki, Jonathan D. Beezley, Deepak Roy Chittajallu, David Manthey, David A. Gutman, and Lee A. D. Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35:3461 – 3467, 2019.

Author contributions

AKF concieved the idea and supervised the project, AB worked out the methods and carried out experiments. AKF and AB wrote the manuscript.

Data availability

The data used in this study are secondary data obtained from publicly available sources. All data sources are properly cited within the manuscript. No new primary data were generated or analyzed in this research. For access to the datasets used, please refer to the original sources as indicated in the references section.

Additional information

The authors declare no conflicts of interest.