JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 0, 2021
© Mary Ann Liebert, Inc.
Pp. 1–16
DOI: 10.1089/cmb.2020.0551

ISBRA 2020

Randomness in Cancer Breakpoint Prediction

KSENIIA CHELOSHKINA,[1] ISLAM BZHIKHATLOV,[2] and MARIA POPTSOVA[1]

ABSTRACT

Cancer genomes are susceptible to multiple rearrangements by deleting, inserting, and translocating genomic regions. Recently, the problem of finding determinants of breakpoint formations was approached with machine learning methods; however, unlike cancer point mutations, breakpoint prediction appeared to be a more difficult task, and various machine learning models did not achieve high prediction power often slightly exceeding the threshold of random guessing. This raised the question of whether the breakpoints are random noise in cancer mutagenesis or there exist determinants in structural mutagenesis. In the present study, we investigated randomness in cancer breakpoint genome distributions through the power of machine learning models to predict breakpoint hot spots. We divided all cancer types into three groups by degree of randomness in their breakpoint formation. We tested different density thresholds and explored the bias in hot spot definition. We also compared prediction of hot spots versus individual breakpoints. We found that hot spots are considerably better predicted than individual breakpoints; however, some individual breakpoints can also be predicted with a satisfactory power, and thus, it is not proper to filter them from analyses. We demonstrated that positive–unlabeled learning can provide insights into insufficiency of cancer data sets, which are not always reflected by data set sizes. Overall, the present results support the view that cancer breakpoint landscape can be represented by predictable dense breakpoint regions and scattered individual breakpoints, which are not all random noise, but some are generated by detectable mechanism.

Keywords: cancer breakpoint hot spots, cancer breakpoints, cancer genome rearrangements, machine learning, random forest.

1. INTRODUCTION

The major characteristic property of cancer genomes is genome rearrangements resulting in the formation of deletions, inversions, translocations, and copy number variants. Due to the efforts of the international cancer genome consortiums—The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC)—hundreds of thousands of cancer breakpoints has been documented for different types of cancers (Nakagawa et al., 2015; Harewood et al., 2017; Nakagawa and Fujita, 2018). One of the biggest problems in analyses of cancer mutations is their heterogeneity that makes it difficult to find

[1] Laboratory of Bioinformatics, Faculty of Computer Science, National Research University Higher School of Economics, Moscow, Russia.
[2] Faculty of Control Systems and Robotics, ITMO University, St. Petersburg, Russia.

strong statistical signals in cancer mutation data and delays biomarker discovery. Heterogeneity of cancer mutations has been noticed long ago (Salk et al., 2010), and it finally led to the emergence of the notion of cancer genome landscapes (Vogelstein et al., 2013). Cancer sequencing genome data provided insights into cancer mutation processes at the level of individual genomes, and the overall information was combined into the pan-cancer genome (Cancer Genome Atlas Research et al., 2013; Zhang and Wang, 2015) revealing common and individual properties of cancer genome mutations. Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the ICGC and TCGA reported the integrative analysis of more than 2500 whole-cancer genomes across 38 tumor types (Consortium, 2020). Analysis of the accumulated cancer genome data confirmed that genome rearrangement with creation of structural elements is often an early event in cancer evolution sometimes preceding point mutation accumulation (Javadekar and Raghavan, 2015; Consortium ITP-CAoWG, 2020; Li et al., 2020).

Recently, finding regularities in cancer genome mutagenesis become a task for machine learning algorithms. Initially, machine learning models were applied to predicting densities of somatic mutations based on epigenetic markers and DNA structural organization from localized DNA structures to chromatin packaging (Polak et al., 2015; Georgakopoulos-Soares et al., 2018). Thus, in Polak et al. (2015), the authors could predict density of somatic point mutations with the densities of HDNase, a marker of an open/close chromatin state, and densities of histone marks with a high coefficient of determination of 0.7–0.8. In Georgakopoulos-Soares et al. (2018), the authors studied contribution of non-B DNA structures and epigenetic factors to origination of cancer point mutations. The authors showed that machine learning models can predict density of point mutations with the major contribution of non-B DNA structures; however, taking both groups of factors into account results in even higher prediction power.

Despite success of machine learning approach in finding determinants of somatic point mutations, cancer breakpoint prediction models showed low or moderate power (Georgakopoulos-Soares et al., 2018; Cheloshkina and Poptsova, 2019). From one hand, this can be due to the absence of causal determinants in the models, from the other hand, machine learning algorithms have limitations.

Previously, we showed that the breakpoint density distribution varies across different chromosomes in different cancer types (Cheloshkina and Poptsova, 2019). Also, we showed that determination of hot spot breakpoints depends on a threshold, and the choice of the threshold could vary between different types of cancers and could affect the results of machine learning models. In the current work, we performed a series of experiments with machine learning models to reveal how breakpoint hot spots density thresholds influence prediction power of machine learning models. Along with building machine learning models predicting breakpoint hot spots, we built machine learning models predicting individual breakpoints and machine learning models trained to distinguish breakpoint hot spots from individual breakpoints. With this approach, it is possible to quantify the degree of randomness in generation of individual breakpoints and reveal cancer genomes where breakpoints can be predicted with a relatively high power.

We addressed the cancer breakpoint data insufficiency issue with positive–unlabeled (PU) learning approach. Overall, our study helps to understand the degree of randomness in cancer breakpoint genome distribution across various cancer types and make a step forward toward determining signatures of cancer structural variation.

2. METHODS

2.1. Data

Data on cancer breakpoints were downloaded from the ICGC. The data set comprises more than 652,000 breakpoints of 2803 samples from more than 40 different types of cancers that we grouped into 10 groups of cancer according to tissue types and further refer as cancer types. We cut the genome into nonoverlapping windows of 100 kb of length and excluded regions from centromeres, telomeres, blacklisted regions, and Y chromosome. Then, for each window, we estimated breakpoint density as the ratio of the number of breakpoints in the window to the total number of breakpoints in a given chromosome. We used the density metric to designate hot spots, that is, genomic regions with a relatively high concentration of breakpoints. In the study, we investigate three labeling types of hot spots—99%, 99.5%, and 99.9% percentiles of breakpoint density distribution. Besides, we assigned ''individual breakpoints'' label to windows containing at least one breakpoint. The proportion of the number of these windows from the total number of windows varied from 2.8% to 90% for different cancer types.

For predictors, we collected data sets from different group of cancers including genomic regions, non-B DNA secondary structures, transcription factor binding sites, histone markers, chromatin accessibility, topologically associated domain boundaries, and DNA methylation. The data sets on transcription factor binding sites were downloaded from Encode; non-B DNA structure and repeat markups were taken from non-B database and DNA punctuation project. Data sets on histone marks, DNA methylation, HDNase accessibility, and TAD boundaries were taken from Encode and UCSC Genome Browser. For each individual feature, we calculated feature density for 100 kb regions. The final data set represented a collection of breakpoint density regions with corresponding feature vectors in the form of feature densities for the same regions.

2.2. *Machine learning models*

During this study, we tested three different quantile thresholds for breakpoint hot spots labeling: 99%, 99.5%, and 99.9%. In total, we assembled 30 data sets for 10 cancer types and different hot spot labeling. The train–test splits with stratification by a chromosome were created for each data set in proportion of 70–30, retaining 30% of data for testing. To get a reliable estimate of quality metrics, we performed train–test splits 30 times for each data set because of high-class imbalance (very small ratio of positive examples). For this reason, we also applied the class balancing technique (oversampling) when training a machine learning model.

As a machine learning algorithm, we took random forest and estimated the following hyperparameters: the number of trees, number of features to grow a tree, minimal number of examples in a terminal node, and maximal number of nodes in a tree. We took the average from the model performance metrics for all cancer types.

2.3. *Evaluation metrics*

The following metrics were taken for model evaluation: ROC AUC (area under receiver operating curve), precision, recall, lift of precision, and lift of recall. Because of the class imbalance, we performed 30 random train–test splits and took the mean (or median) ROC AUC on the test set controlling for its standard deviation. This approach helps to reveal how strongly the results depend on the distribution of examples in train and test sets. We calculated recall and precision for different probability percentiles—from 0.5% to 50%.

Additionally, we reported two metrics, the lift of recall and lift of precision, which are metrics estimating, how well a model performs in comparison with a random choice. In contrast to ROC AUC, these metrics shows model's ability to distinguish classes. The lift of recall for a given probability percentile shows how many times the recall of the model (estimated on examples labeled as the positive class according to a probability percentile threshold) is higher than a random choice (it is equal to the re-call of the model divided by the probability percentile). It is supposed that in case of random choice model selection of *n* % of sample holds *n* % of positive examples so that the ratio of probability percentile used as threshold and recall will approach 1. This way it is expected from a good model to have value of lift of recall higher than 1. Similarly, the lift of precision demonstrates how many times precision for a given probability percentile is higher than a random choice (equal to the precision of the model divided by the proportion of positive examples in a data set).

2.4. *PU learning*

To address the cancer breakpoint data insufficiency issue, we used PU learning approach. The approach supposes that in a sample there are examples labeled as positive and all the rest examples are of unknown label (may be positive or negative). The task is to label all unlabeled examples taking into consideration known labels and different feature distributions.

In the article, we adopted the PU learning algorithm (Liu et al., 2002, 2003) and run this algorithm, for all data sets for each cancer type. The general idea of the algorithm is to iteratively update positive and negative sets until convergence by setting new certainty thresholds after building a model on new positive and negative sets. The algorithm has hyperparameter ε (width of certainty interval). The algorithm can be summarized as follows:

```
1. For $\varepsilon$ in {0.01, 0.03, 0.05}:
    1.1. Apply the auxiliary algorithm ‘‘get prediction’’ (see below) and get predictions, $r_{up}$ and $r_{low}$. If
         the algorithm fails move to next $\varepsilon$.
```

1.2. Label RN (reliable negatives) and RP (reliable positives) data points in the training set: if the predicted probability is greater than r_{up}, then a data point is labeled as RP, if predicted probability is less than r_{low}, then a data point is labeled as RN.

1.3. Compose a new training set from RN, RP, and initial positive examples. The remaining data points are included in the unlabeled set.

1.4. If there are labeled RN as well as unlabeled data points, then proceed with the next step, otherwise save results and move to the next ε.

1.5. For each iteration ($k = 5$):

  1.5.1. Apply the auxiliary algorithm ''get prediction'' and get predictions, r_{up} and r_{low}. If it fails move to the next ε.

  1.5.2. Label RN (reliable negatives) and RP (reliable positives) data in the unlabeled set: if the predicted probability is greater than r_{up}, then the data point is labeled as RP, if the predicted probability is less than r_{low}, then the data point as RN.

  1.5.3. Update the training set by the addition of new labeled RN and RP and remove them from the unlabeled set.

  1.5.4. The process stops when:

  - the number of new labeled RN is greater than the number in previous iteration or less than number of initial positives
  - $r_{up} < r_{low}$
  - unlabeled set is empty
  - there is no new labeled examples (neither RN nor RP)

Otherwise proceed with the next iteration
The auxiliary algorithm ''get prediction'':

1. Train machine learning model on training set.
2. Get probability predictions for the train, test (and unlabeled if determined) sets.
3. Calculate r_{up} and r_{low}:

Let q_1 be the 90% percentile of the training set probability distribution and q_2 be the 10% percentile of the training set probability distribution. If $q_1 = q_2$, then exit the function with an error, else if $q_1 - q_2 \leq 2\varepsilon$, then $r_{up} = q_1$ and $r_{low} = q_2$, else $r_{up} = q_1 - \varepsilon$ and $r_{low} = q_2 - \varepsilon$.
Output: predicted probabilities, r_{up} and r_{low}.

The above algorithms were implemented in two modes: RP mode (fully corresponds to the described algorithm) and RN mode (modified version of the algorithm when a set of positive examples remains fixed and only RN examples are updated at each iteration according to the procedure).

3. RESULTS

3.1. Dependence of machine learning model powers on density thresholds

We train random forest models on all 30 data sets for hot spot prediction and 10 data sets for individual breakpoint prediction. The distribution of ROC AUC on test set by cancer type and labeling type is given in Figure 1. It could be observed that for the half of cancer types including blood, brain, breast, pancreatic, and skin cancer, the higher the hot spot labeling threshold the higher the median of test ROC AUC. For bone, liver, and uterus cancer, there is no monotonically increasing median test quality, but for the highest labeling type, this value is higher than for the lowest, while for the rest of cancers (ovary and prostate), there is no significant difference between labeling types.

The median values of the test ROC AUC by cancer type and labeling type are presented in Figure 2 and Table 1. The highest quality in terms of considered metric belongs to the breast cancer for all labeling types while the lowest—to the blood and skin cancer for 99% labeling type. The difference greater than 0.10 between the median test ROC AUC for models of the 99% and 99.9% hot spot labeling types is observed for blood, brain, breast, liver, and pancreatic cancer. For the half of the cancer types, the highest labeling type implies significantly higher quality according to the median test ROC AUC.

However, if one looks at the shapes of distributions for different labeling types (Fig. 1), one can observe that it becomes broader with higher labeling threshold. In Figure 3, we plotted the difference between the

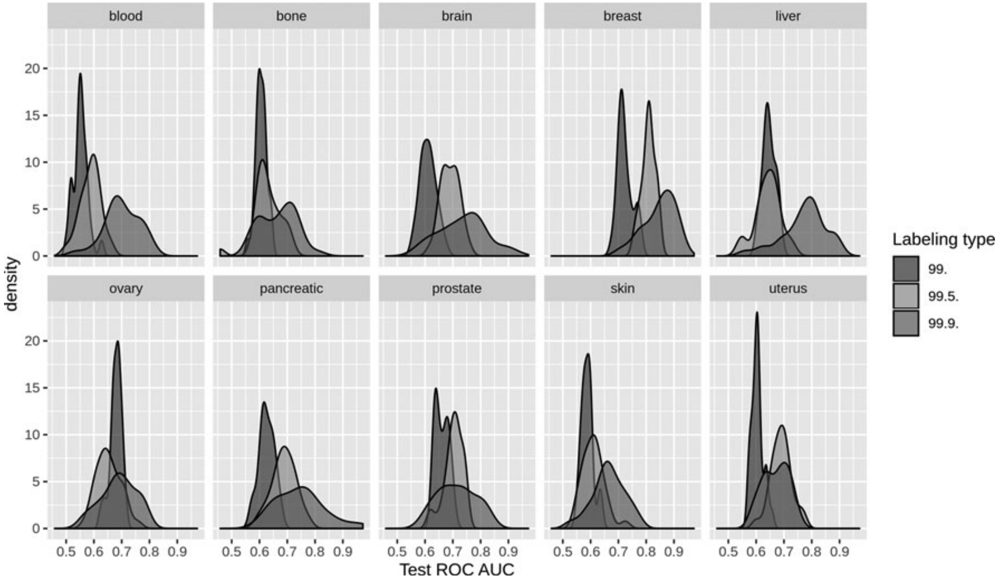

FIG. 1. Distribution of test ROC AUC for all cancer types by hot spots labeling. ROC AUC, area under receiver operating curve.

train and test ROC AUC against the median test ROC AUC. One can observe from Figures 1 and 3 that the more rare the hot spots (higher labeling type), the higher the variance of the test ROC AUC on the test set as well as the difference between the train and test ROC AUC. This could be explained by the fact that for the case of rare hot spots there is a small number of positive examples in a data set, and its random permutation between the train and test set leads to different results. Moreover, for all cancer types except for the breast cancer, difference between the median train and test ROC AUC for the 99.9% labeling type approaches 0.2 ROC AUC and is two to three times larger than for the 99.5% and 99% labeling types. This

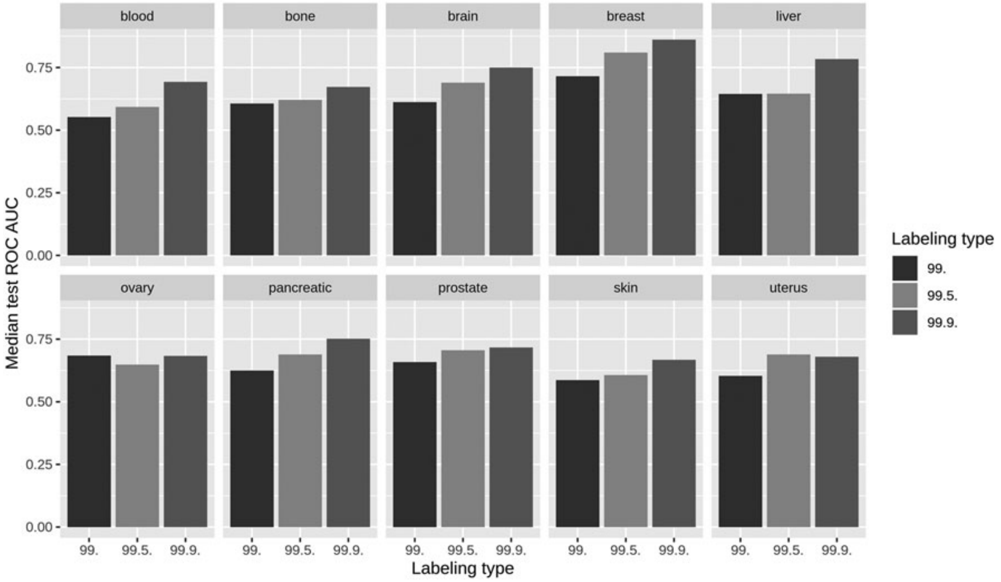

FIG. 2. Median test ROC AUC for each cancer and labeling type.

Table 1. Median Test Area Under Receiver Operating Curve for Each Cancer and Labeling Type

Cancer type	*Median test ROC AUC (99%)*	*Median test ROC AUC (99.5%)*	*Median test ROC AUC (99.9%)*
Blood	0.552	0.593	0.693
Bone	0.606	0.621	0.673
Brain	0.612	0.689	0.75
Breast	0.715	0.81	0.861
Liver	0.645	0.646	0.784
Ovary	0.684	0.648	0.683
Pancreatic	0.625	0.689	0.752
Prostate	0.659	0.706	0.717
Skin	0.587	0.607	0.668
Uterus	0.603	0.689	0.68

AUC ROC, area under receiver operating curve.

means that the highest labeling threshold is not always a good choice, and the second, and even the third, labeling type will be a more reasonable choice.

3.2. *Lift of precision and lift of recall*

The results on the distribution of other quality metrics such as the lift of recall and lift of precision are presented in Figures 4 and 5, respectively. Here, confidence intervals for the mean of these metrics are plotted against different probability quantiles selected as a threshold for model predictions for each cancer and hot spot labeling type. The main conclusion that could be made according to these results is that there is no single labeling type, which guarantees the best classification results for all cancer types. However, three groups of cancer types were distinguished: the best labeling type for blood, brain, liver, and pancreatic cancer is 99.9%, for bone, breast, and uterus cancer—99.5%, for the rest (ovary, prostate, and skin cancer)—99%.

When comparing the best labeling type determined with the median test ROC AUC and with the lift of recall/precision, it is the same only for ovary, bone, and uterus cancer. As we are mainly interested in selection of minimal number of genome regions with the maximal concentration of hot spots, the final choice of the best labeling type will coincide with the decision according to the lift of recall/precision.

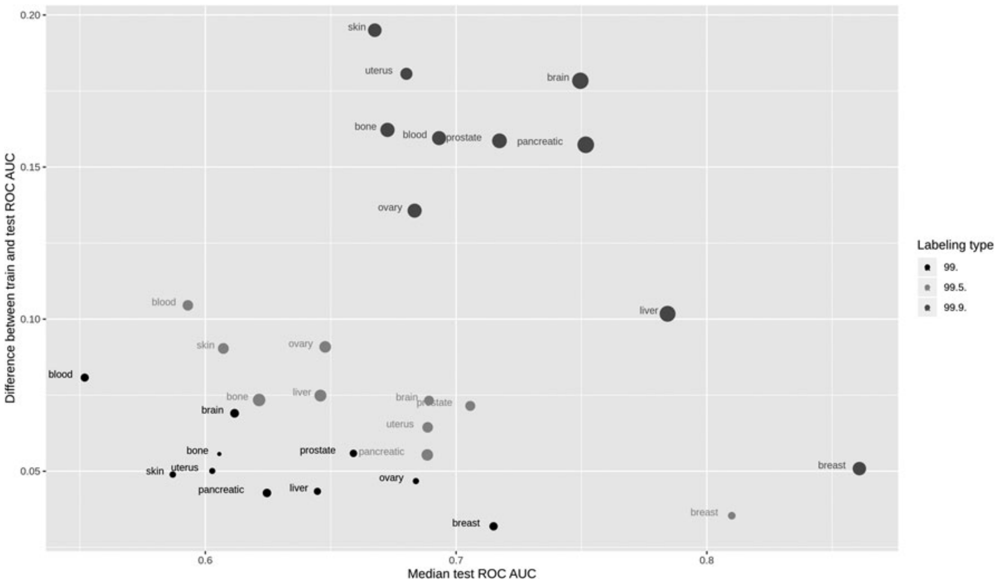

FIG. 3. Dependence of the difference of train and test ROC AUC from test ROC AUC at different labeling types and different standard deviations of test ROC AUC.

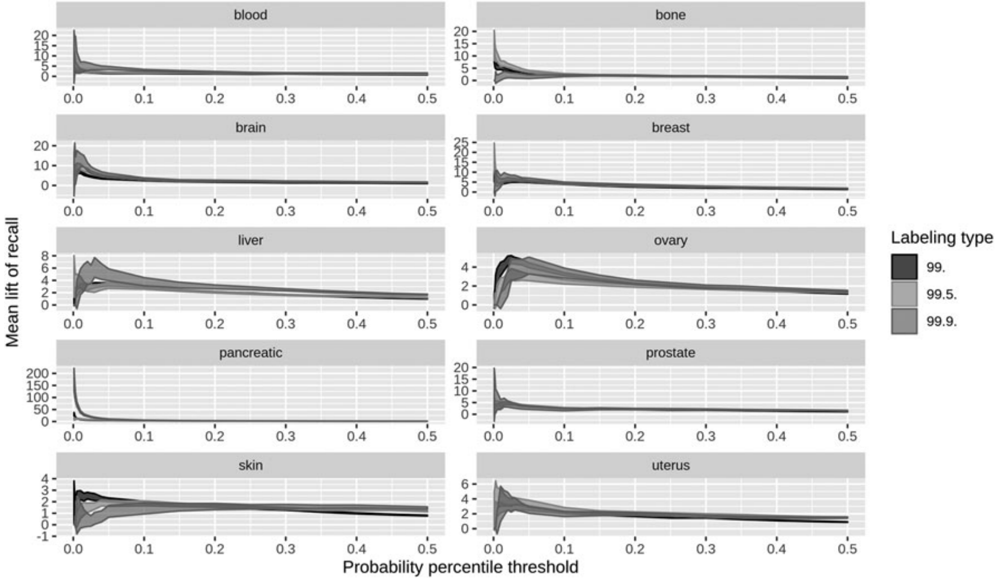

FIG. 4. Dependence of lift of recall from quantile threshold for different aggregation levels (see text for explanation).

Interestingly, for the breast cancer, all three labeling types are almost equally well predicted: they have relatively high lift of recall and differ slightly. Besides, for pancreatic cancer, 99.9% labeling type showed significant boost in both lift of precision and lift of recall.

3.3. *Prediction of hot spot breakpoints versus individual breakpoints*

The main question of the present study was to test how well recurrent and nonrecurrent breakpoints are predictable. We assume that the high-density breakpoint regions represent regions with recurrent breaks.

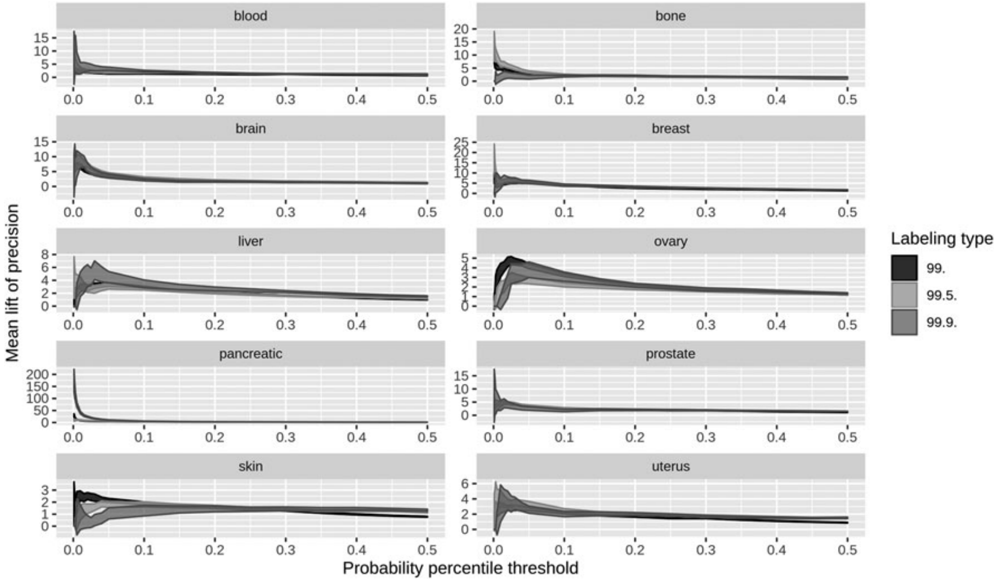

FIG. 5. Dependence of **(A)** lift of recall and lift of precision **(B)** from quantile threshold for different aggregation levels (see text for explanation).

First, we tested the predictive power of the machine learning models to recognize hot spots and separately individual breakpoints with filtered hot spots. Figure 6 shows distributions of test ROC AUC for hot spots (the best labeling type) and breakpoint prediction tasks.

The obtained results clearly demonstrate that breakpoint hot spots are considerably better predicted than individual breakpoints; however, there are two exceptions. One is for prostate cancer when the power of the models predicting breakpoints is almost the same as the power of the models predicting breakpoint hot spots and reaches almost 70% ROC AUC. The second is for breast cancer with the power of machine learning models predicting breakpoints is almost 76% ROC AUC; however, the power of predicting hot spots is 10% higher.

The quantitative estimate of the difference is given in Figure 7 and Table 2, which also reports the ratio of median test ROC AUC for hot spots and breakpoints prediction models. The highest ratio of the median test ROC AUC for hot spot prediction model to the median test ROC AUC for breakpoint prediction model is observed for the uterus and brain cancer (1.36 and 1.23, respectively), while for the prostate cancer, they are almost equal. For the other cancer types, the metric for hot spot model is 9%–18% higher than for the breakpoints. Nevertheless, breakpoints prediction remains slightly higher than random, and degree of randomness can be reflecting in test ROC AUC values. Also, it could be observed that variance of ROC AUC is significantly lower for breakpoints, and this could be a consequence of having a considerably higher number of positive examples for the model.

The low quality of models predicting breakpoints could be a consequence of constrains imposed by machine learning models approach where we manually assembled vectors of features. Although we tried to incorporate as many predictors as it is available from omics data comprising different groups, still we could miss the actual determinants. The group of cancer with moderately predicting breakpoints includes skin, liver, bone, blood, brain, and uterus cancer the median test ROC AUC around 0.6. In this case, we can conclude that the resulting set of breakpoints in those cancers is enriched in random breakpoints. And finally, in blood and uterus, the breakpoints are not predicted at all and corresponding models do not differ from random guessing.

The statistics of the lift of recall confirms the conclusion above (Fig. 8). All cancer types could be divided into two groups. For the pancreatic and skin cancer, breakpoints are unrecognizable as the lift of recall is very low (almost equal to zero). For ovary, breast, uterus, and prostate cancer, the metric hardly achieves 1 for the probability percentile threshold of 0–0.1 so that in these cases breakpoints are predicted as successfully as in the case of a random choice. For blood, bone, brain, and liver cancer, there are some probability thresholds for which the lift of recall is higher than 1 with the brain cancer model performing

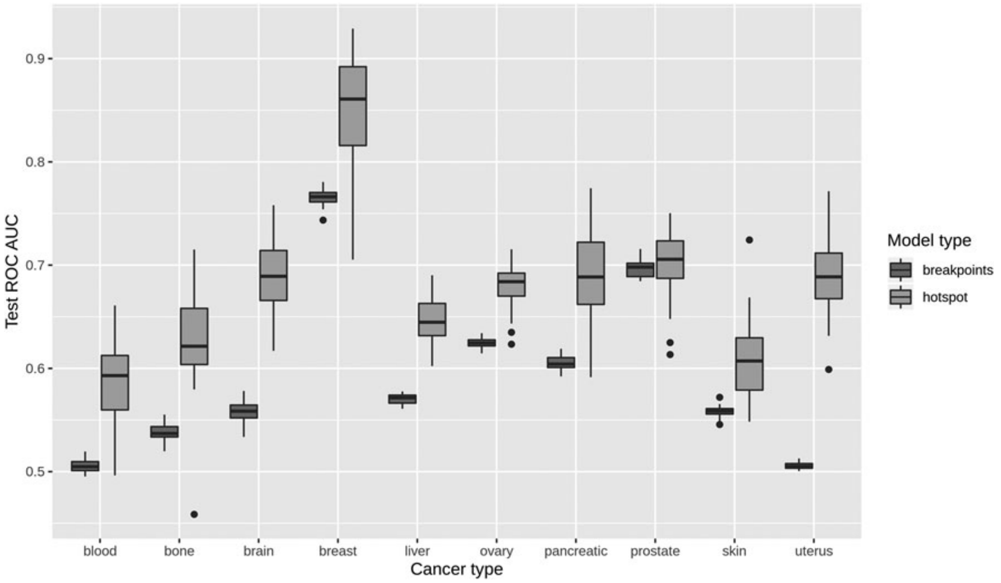

FIG. 6. Distribution of test ROC AUC for the best labeling hot spot profile and all breakpoints prediction.

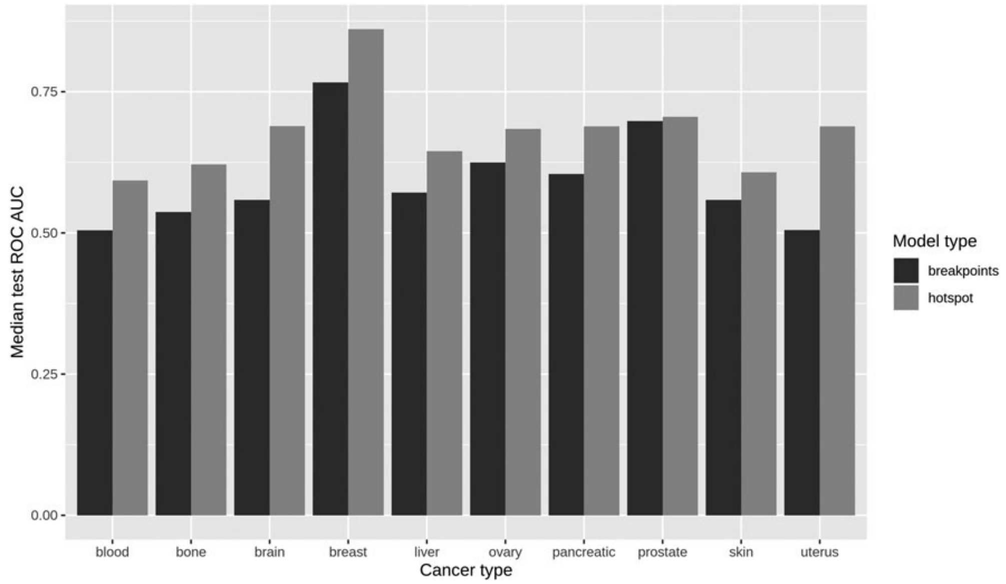

FIG. 7. Median test ROC AUC for best labeling hot spot profile and breakpoints prediction.

the best. In total, for six cancer types, the breakpoint prediction model quality does not significantly differ from a random choice, and only for four cancer types, the prediction is slightly better.

Additionally, it should be noted that for blood, bone, and brain cancer the lift of recall decreases with probability threshold. This could mean that for these cancer types some breakpoints are highly pronounced and could be more easily identified than the rest of breakpoints.

Besides, a set of cancer type models achieving the best performance for the task of breakpoint prediction according to the median ROC AUC (prostate, ovary, and breast) differs from a set determined by the lift of recall (blood, bone, brain, and liver). This difference outlines the fact that it is very important to choose the right performance metric for a given machine learning task. As the ROC AUC measures a quality of overall examples' ordering produced by the model and the lift of recall measures ordering quality of examples with the highest probabilities, they describe the model performance from different perspectives.

3.4. PU learning models

In case when the target variable heavily relies on the distribution of a specific variable, one has to deal with the data sufficiency problems. If breakpoint data are not representative, then the hot spot labeling may

Table 2. Median Test Area Under Receiver Operating Curve for Best Labeling Hot spot Profile and Breakpoints Prediction

Cancer type	*Breakpoints median test ROC AUC*	*Hot spots median test ROC AUC*	*Ratio of median test ROC AUC for hot spots and breakpoints prediction models*
Prostate	0.698	0.706	1.011
Skin	0.559	0.607	1.087
Ovary	0.625	0.684	1.095
Breast	0.766	0.861	1.124
Liver	0.572	0.645	1.128
Pancreatic	0.604	0.689	1.14
Bone	0.537	0.621	1.157
Blood	0.505	0.593	1.175
Brain	0.559	0.689	1.234
Uterus	0.505	0.689	1.363

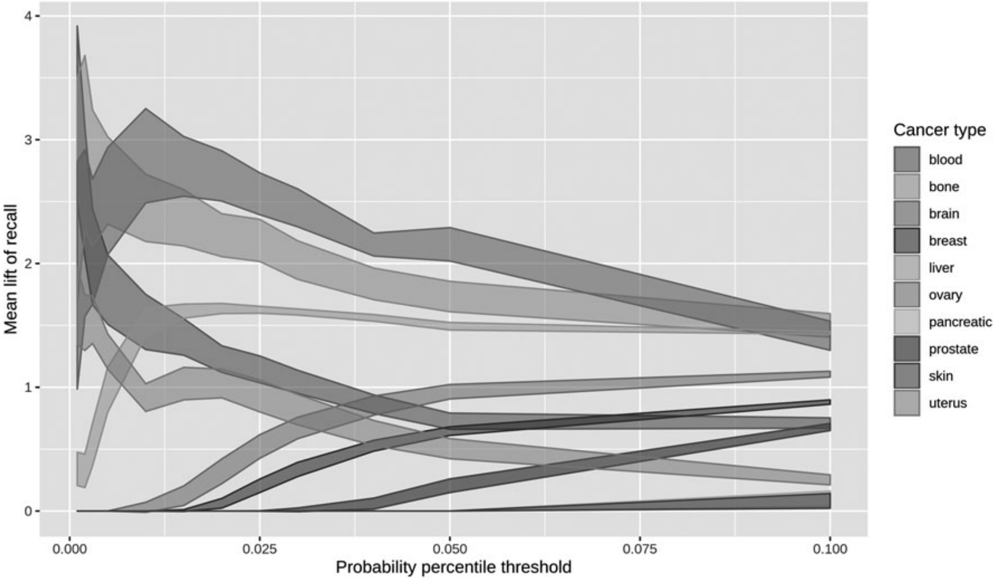

FIG. 8. Lift of recall for breakpoints prediction model.

be incorrect (due to the gaps some breakpoint counts could be underestimated and therefore some hot spot windows will not be labeled as hot spots). In this case, we can incorporate this uncertainty into the model using PU learning approach (Liu et al., 2002, 2003). In PU learning, a given sample is represented by positive data points and unlabeled data points, which could be positive or negative, but the true labels are unknown. The problem of building a binary classifier is approached using RN examples and RP examples. First, classification method is trained to separate positive and unlabeled sets, the latter is considered as the negative class. Then, the model generates prediction probabilities, and the RNs (and, optionally, the RPs) are determined based on thresholds, whereas some data could remain unlabeled. After that the model is trained only on the labeled data and the process is repeated several times until specific conditions are met (see Section 2).

Here, we implemented two modes of the algorithm: the RN mode (only the RNs are labeled at each iteration while positive data points are fixed) and the RP mode (sets of the RNs and RPs are adjusted at each iteration). We built models based on the final feature set for each cancer type (Supplementary Table S1).

Model performance comparison is presented in Figure 9 as the confidence interval for mean difference in lift of recall for RP and RN mode. It was observed that the sign of the difference depends on the number of breakpoints available for the cancer type. On one hand, the stable positive effect is observed for the cancer types, which belong to top-5 with the lowest number of breakpoints. In particular, the best results are demonstrated for the brain cancer, which has the minimal number of breakpoints. Here, we can conclude that inclusion of additional positive examples in the case of noisy data (target labeling could be noisy because of unrepresentative breakpoint data) helps to get higher model quality in the PU learning. On the other hand, the stable negative effect is observed for cancer types, which belong to top-4 with the maximal number of breakpoints. This could be explained by the fact that, in case of sufficient data, additional positive examples introduce noise.

Next, we compared PU learning model with the basic classification model. The mean difference in the lift of recall for PU learning and classification model and its confidence interval are presented in Figure 10 and the mean difference in test ROC AUC and its confidence interval in Figure 11. It could be noted that the quality of hot spot recognition in terms of lift of recall/precision is lower for PU learning procedure than for standard classification, and this could be explained by the loss of data, which leads to a higher uncertainty. However, for some cancer types, the mean difference in the test ROC AUC reached 0.01–0.04 for the RN mode, and this means that for these cancer types in PU learning setting, test positive data points was ranked higher, but the highest probabilities were assigned to the lower number of true hot spots.

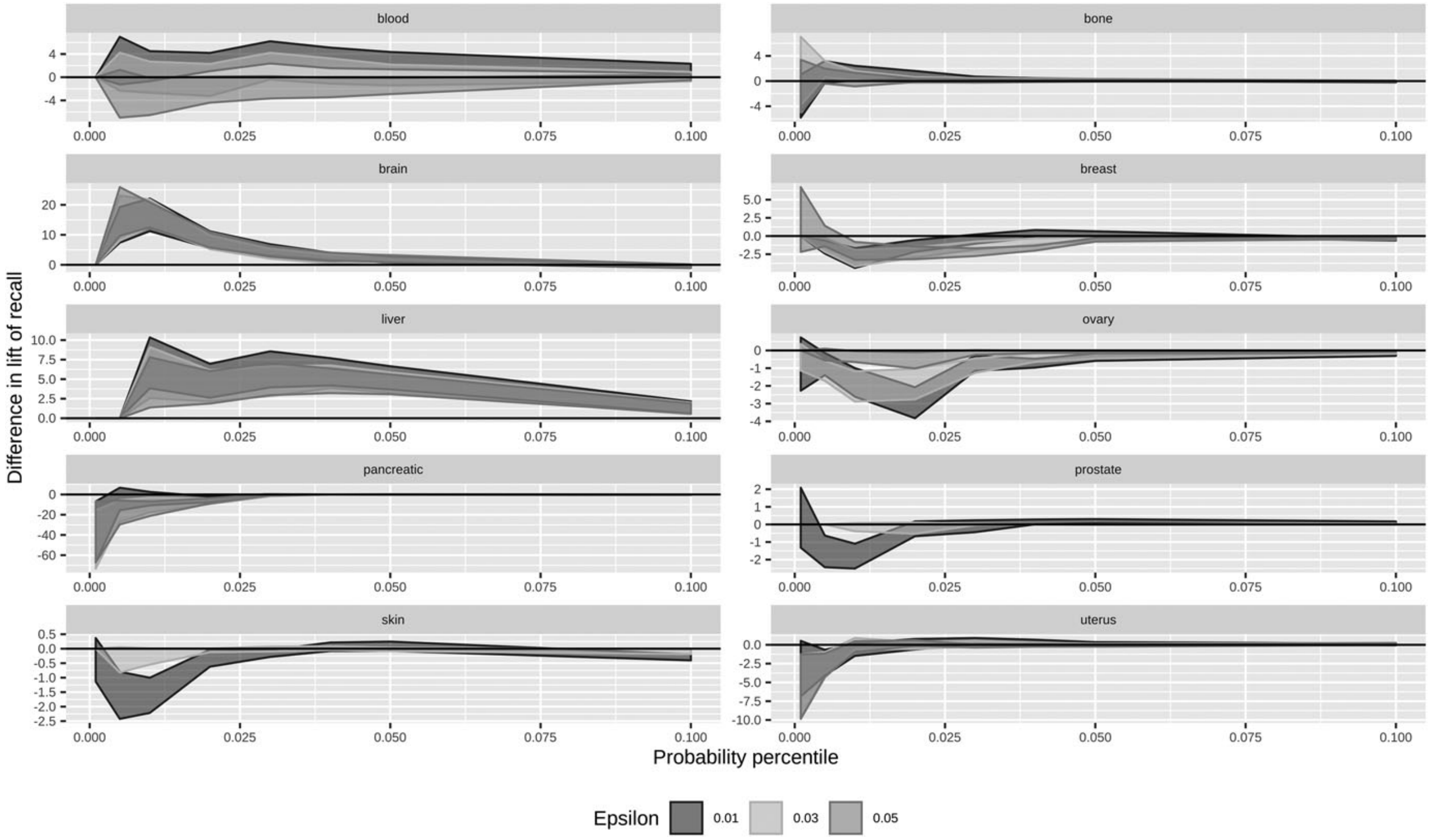

FIG. 9. Confidence interval for mean difference in lift of recall for RP and RN mode. RN, reliable negative; RP, reliable positive.

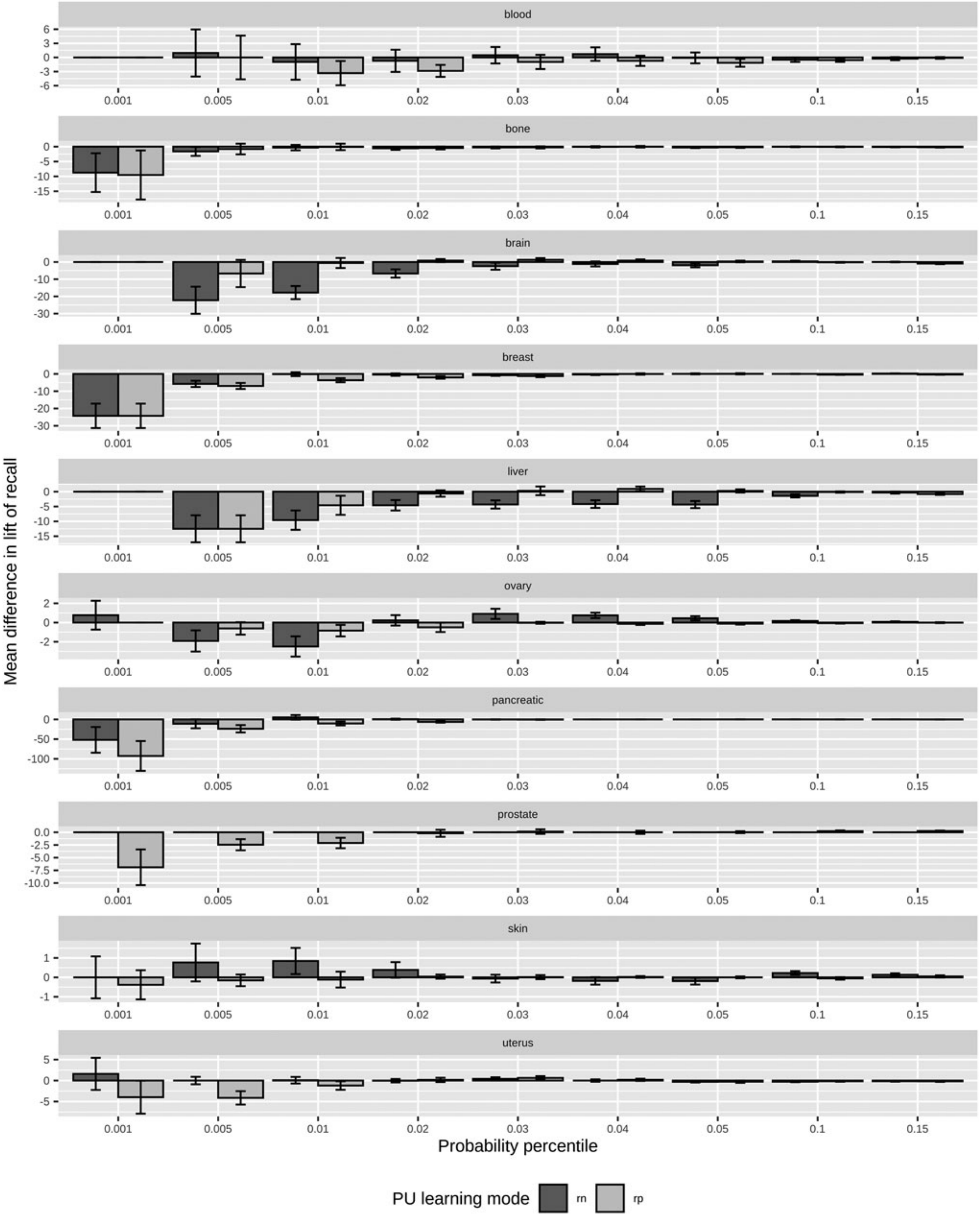

FIG. 10. Mean difference in lift of recall for PU learning and classification settings and its confidence interval. PU, positive-unlabeled.

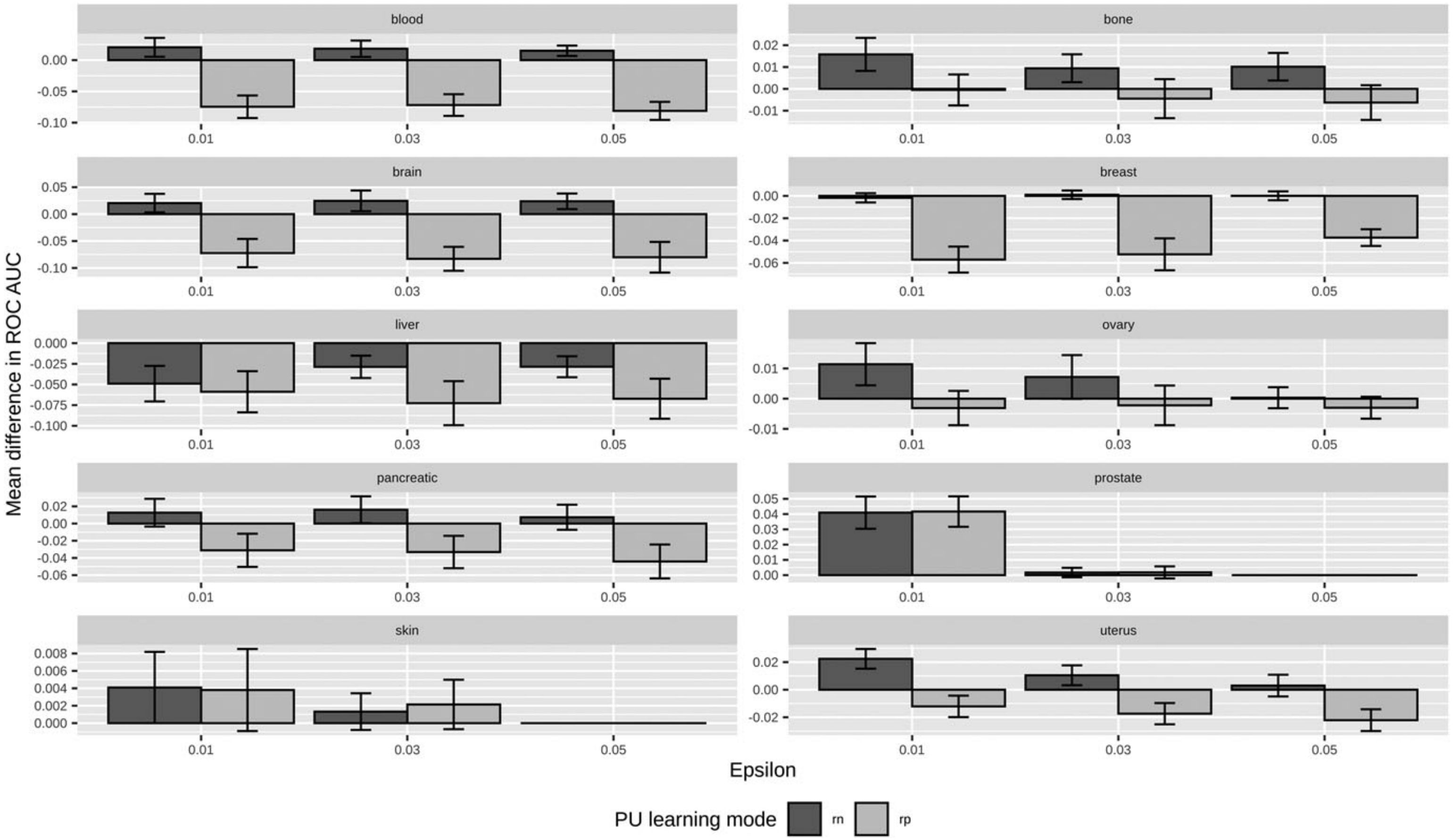

FIG. 11. Mean difference in test ROC AUC and its confidence interval for PU learning and classification settings and its confidence interval.

4. CONCLUSIONS AND DISCUSSION

In this study with machine learning approach, we addressed the problem of randomness in cancer breakpoint formation.

First, we systematically explored the effect of breakpoint hot spot selection criteria on breakpoint hot spots prediction in 10 types of cancer. We built machine learning models trained to predict hot spot breakpoints with three different labeling thresholds and built distributions of machine learning model performance metrics, such as the train and test ROC AUC, lift of precision, and lift of recall. The general increase of the model performance with selection of higher density threshold is not, however, the rule of thumb, because, for some cancer types, by increasing the threshold we also obtain an increase in variance and an increase in difference between the test and train ROC AUC. The suboptimal solution should be based on the combination of different metrics including the mean test ROC AUC, difference between train and test ROC AUC, variance in the distribution of test ROC AUC, mean lift of recall, and mean lift of precision. With a very high-density threshold, one can assemble a set of rare hot spot regions, although highly dense for a particular group of cancerous genomes. When decreasing the threshold, for some cancers, we trade off between less widespread but detectable by the same mutational mechanism, that is why it is possible to predict them with machine learning models.

Regions with increased density of breakpoint mutations, or regions with recurrent breakpoints, are better recognized by the machine learning models. This regularity holds true for almost all cancers with an exception of prostate cancer where prediction power of the models predicting individual breakpoints is as high as the power of the models predicting hot spots (with the median test ROC AUC close to 0.70). It is a subject for future research to answer the question if the mutation mechanisms underlying individual and recurrent breakpoint formation are the same in the observed prostate cancer genomes.

We would like to emphasize that, without actual tests of machine learning performance on individual breakpoints and breakpoint hot spots, it is not evident, which one of the two will have a higher prediction power. Indeed, breakpoint hot spots are regions enriched with breakpoints, and genomic features of these regions should explain their recurrent formation. On the opposite, rare events are often harder to predict except for the cases when strong predictors of a rare event are available. In the research, we posed a question whether the considered genome features identify breakpoints hot spots better than individual breakpoints, and whether individual breakpoints have also distinctive features that influence their formation. The comprehensive analysis of the features is out of scope of the present study is the subject of further systematic research.

In our specific research of cancer hot spot breakpoint mutagenesis with class imbalance problem, we paid much attention to two metrics: the lift of recall and lift of precision. The lift of recall and lift of precision signify how many times recall or precision is higher compared with a random choice. Likewise test ROC AUC, the distributions of the lift of recall and lift of precision confirmed that it is impossible to choose one breakpoint density threshold that would lead to the maximum prediction power of models for all types of cancer. Three groups of cancer with similar behavior according to the lift of recall and lift of precision were distinguished. For blood, brain, liver, and pancreatic cancer, it is beneficial to choose the highest density threshold; for ovary, prostate, and skin cancer, the third lowest, and for the remaining three—bone, breast, and uterus cancer—an intermediate threshold works the best. Again, it is an open question whether there exist common properties in cancer breakpoint mutagenesis in these three groups of cancer.

Selection criteria for the best hot spot labeling threshold based on the median test ROC AUC and the lift of recall and precision coincide only for three types of cancer—ovary, bone, and uterus. Moreover, concerning evaluation of prediction power of breakpoints, these metrics produce different results. As the ROC AUC and lift of recall measure quality of examples' ordering by the model at different scales (based on all examples and examples with the highest probabilities respectively), we recommend to use the lift of recall and lift of precision metrics to choose the hot spot thresholds.

We applied PU learning procedure to recognize data sets that are likely to have insufficient sample size and therefore susceptible to noise. For this purpose, it is often not enough to look only at a sample size. Thus, using PU learning, we found that data sets for the brain and liver cancer are insufficient, while for the blood and bone cancer, which have more breakpoints and samples than the brain but considerably less than liver (two times less samples and nine times less breakpoints), the data set is sufficient.

Comparison of breakpoint predictions and breakpoint hot spots with a chosen selection criterion based on the best machine learning model showed that the median test AUC is always higher for hot spots rather than for individual breakpoints. Overall, the results of our study showed that machine learning model prediction power depends on density threshold for cancer hot spots, and the threshold is different for different types of cancer. Besides, we demonstrated that although individual breakpoints are harder to predict than breakpoint hot spots, individual breakpoints can be predicted to a certain extent, and, moreover, in prostate cancer, they are predicted equally well as hot spots. While choosing a selection criterion, the test ROC AUC only is not enough to choose the best model, the lift of recall and lift of precision should be taken into consideration at the level of individual type of cancer.

The degree of randomness of individual breakpoints scattered over cancerous genomes can be quantitatively estimated through the power of predictive machine learning models. We can roughly distinguish three groups. The first group is when the breakpoints are not predicted at all or slightly better predicted than random; it includes the uterus (ROC AUC of 0.505), blood (0.505), bone (0.537), brain (0.559), and skin (0.559). The other group comprises cancers in which breakpoints can be predicted with a relatively high confidence (prostate with ROC AUC of 0.698 and breast with 0.766). And the remaining cancers—ovary (0.625), pancreatic (0.604), and liver (0.572)—represent a mixture of random noise with predictable breakpoints.

The issue of randomness in cancer breakpoint landscape is a complicated issue and cannot be addressed in full in the framework of the present study. However, some trends can be highlighted. As the result of machine learning analysis and predictive modeling, the emerging view is that in terms of predictability, breakpoint hot spots are superimposed to individual breakpoints. However, individual breakpoints are not all random noise, and some (and the size of that portion is to be estimated) are well predicted by machine learning models as well as hot spots. This can be explained by common underlying mechanisms of breakpoint mutagenesis in the particular cancer type and a particular sample. Our study is an illustration of how machine learning approaches can be applied to address the problem of randomness in cancer breakpoint genome distributions.

ACKNOWLEDGMENT

We thank the Centre for Fundamental Research of the Higher School of Economics, Moscow, Russia, for supporting this research project.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

SUPPLEMENTARY MATERIAL

Supplementary Table S1

REFERENCES

Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120.

Cheloshkina, K., and Poptsova, M. 2019. Tissue-specific impact of stem-loops and quadruplexes on cancer breakpoints formation. *BMC Cancer* 19, 434.

Consortium ITP-CAoWG. 2020. Pan-cancer analysis of whole genomes. *Nature* 578, 82–93.

Georgakopoulos-Soares, I., Morganella, S., Jain, N., et al. 2018. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* 28, 1264–1271.

Harewood, L., Kishore, K., Eldridge, M.D., et al. 2017. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* 18, 125.

Javadekar, S.M., and Raghavan, S.C. 2015. Snaps and mends: DNA breaks and chromosomal translocations. *FEBS J.* 282, 2627–2645.

Li, Y., Roberts, N.D., Wala, J.A., et al. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 112–121.

Liu, B., Dai, Y., Li, X., et al. 2003. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*. IEEE179–IEEE186.

Liu, B., Lee, W.S., Yu, P.S., et al. 2002. Partially supervised classification of text documents. In *ICML*. 387–394.

Nakagawa, H., and Fujita, M. 2018. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci.* 109, 513–522.

Nakagawa, H., Wardell, C.P., Furuta, M., et al. 2015. Cancer whole-genome sequencing: Present and future. *Oncogene* 34, 5943–5950.

Polak, P., Karlic, R., Koren, A., et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518, 360–364.

Salk, J.J., Fox, E.J., and Loeb, L.A. 2010. Mutational heterogeneity in human cancers: Origin and consequences. *Annu. Rev. Pathol.* 5, 51–75.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., et al. 2013. Cancer genome landscapes. *Science* 339, 1546–1558.

Zhang, K., and Wang, H. 2015. [Cancer Genome Atlas Pan-cancer Analysis Project]. *Zhongguo Fei Ai Za Zhi* 18, 219–223.

Address correspondence to:
Dr. Maria Poptsova
Laboratory of Bioinformatics
Faculty of Computer Science
National Research University Higher School of Economics
11 Pokrovsky Boulvar
Moscow 101000
Russia

E-mail: mpoptsova@hse.ru