Convolutional Neural Network-Based Classification of Histopathological Images Affected by Data Imbalance

Michał Koziarski(✉), Bogdan Kwolek, and Bogusław Cyganek

Department of Electronics, AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland
michal.koziarski@agh.edu.pl

Abstract. In this paper we experimentally evaluated the impact of data imbalance on the convolutional neural networks performance in the histopathological image recognition task. We conducted our analysis on the Breast Cancer Histopathological Database. We considered four phenomena associated with data imbalance: how does it affect classification performance, what strategies of preventing imbalance are suitable for histopathological data, how presence of imbalance affects the value of new observations, and whether sampling training data from a balanced distribution during data acquisition is beneficial if test data will remain imbalanced. The most important findings of our experimental analysis are the following: while high imbalance significantly affects the performance, for some of the metrics small imbalance. Sampling training data from a balanced distribution had a decremental effect, and we achieved a better performance applying a dedicated strategy of dealing with imbalance. Finally, not all of the traditional strategies of dealing with imbalance translate well to the histopathological image recognition setting.

Keywords: Convolutional neural network · Data imbalance · Histopathological image classification

1 Introduction

Due to the recent algorithmic advances, as well as a growing amount of data and computational resources, machine learning is becoming increasingly suitable option for the task of histopathological data processing. In particular, deep learning methods are becoming dominant technique in the field [4]. A significant amount of work has been done by the scientific community on the problem of using deep learning algorithms in the histopathological image recognition task. However, despite that, a little attention has been given to the issue of data imbalance in the histopathological setting, or more generally in the image recognition task. Data imbalance [9] can be defined as a situation, in which the number of observations from one of the classes (majority class) is higher than the number

© Springer Nature Switzerland AG 2019
X. Bai et al. (Eds.): FFER 2018/DLPR 2018, LNCS 11264, pp. 1–11, 2019.
https://doi.org/10.1007/978-3-030-12177-8_1

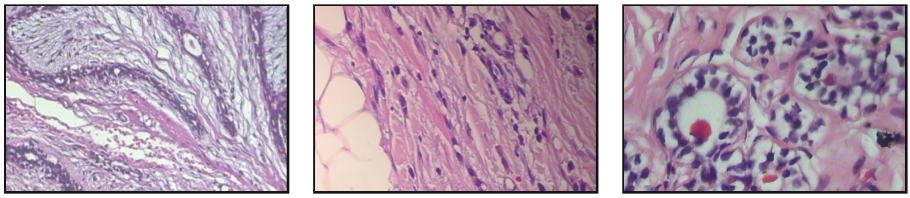
of observations from another class (minority class). Most of the existing machine learning algorithms assume a balanced data distribution, and perform poorly in an imbalanced setting, biasing predictions towards the majority class. Notably, data imbalance can be observed in various existing histopathological benchmark datasets, such as Breast Cancer Histopathological Database (BreakHis) [14]. It is, however, unclear to what extent data imbalance affects the performance of deep learning algorithms in the histopathological image recognition task, or what techniques of dealing with data imbalance are suitable in such setting. In a recent study Pulgar et al. [12] evaluate the impact of data imbalance on the performance of convolutional neural networks in the traffic sign recognition task. They conclude that data imbalance negatively affects the performance of neural networks. They do not, however, consider using any strategies of dealing with data imbalance. In another study by Buda et al. [1] the authors also evaluate the impact of data imbalance on the performance of convolutional neural networks, this time evaluating some of the existing strategies of dealing with imbalance. However, neither of the mentioned papers uses the histopathological data. Furthermore, in this study we consider additional questions related to the issue of data imbalance, namely the value of new observations in the imbalanced data setting and the choice of strategy of dealing with imbalance. Finally, it is worth mentioning a study by Lusa [11], in which the author experimentally evaluates the performance of one of the most prevalent strategies of dealing with data imbalance, SMOTE [2], on a high-dimensional data. Based on that study, SMOTE is not suitable for dealing with a high-dimensional data, such as images. It is not clear whether other strategies of dealing with imbalance translate well into the histopathological image setting.

In this paper we extend on the previous research, in particular focusing on the problem of histopathological image recognition. We experimentally evaluate various trends associated with data imbalance. First of all, we test to what extent data imbalance influences the classification performance. Secondly, we evaluate various strategies of dealing with data imbalance. Thirdly, we measure how data imbalance influences the value of new data. Finally, we test the hypothesis that artificially balancing the training distribution during data can be beneficial for performance, even if the test distribution is imbalanced.

2 Experimental Study

2.1 Set-Up

Dataset. We conducted our experiments on the Breast Cancer Histopathological Database (BreakHis) [14]. It contained 7909 microscopic images of breast tumor tissue, extracted using magnification factors 40X, 100X, 200X and 400X, with approximately 2000 images per magnification factor. Each image had the dimensionality of 700×460 pixels and an associated binary label, indicating whether the sample was benign or malignant. At each magnification factor the data was randomly divided into 5 folds, with approximately 70% of the samples



(a) 100X (b) 200X (c) 400X

Fig. 1. Sample images from BreakHis dataset at different magnification factors.

reserved for training, and 30% for testing. In our experiments we reused the random partitioning provided by the authors of the BreakHis dataset (Fig. 1).

By default, BreakHis dataset displayed the imbalance of approximately 2.0, with the malignant samples belonging to the majority class. During our experiments we performed undersampling of the data up to the point of achieving the desired imbalance ratio (IR). We considered IRs $\in \{1.0, 2.0, \ldots, 10.0\}$. Importantly, for each IR we used the same total number of samples, that is 676 training and 336 test images. It was the maximum amount of data allowing us to produce every considered IR. We decided to keep the same total number of samples for each IR, as opposed to decreasing the number of samples from the minority class and keeping the size of the majority class constant. It allowed us to avoid the issue of decreasing amount of data, which could be another factor affecting the classification performance.

Classification. For the classification we used the architecture of a convolutional neural network described in [13]. It consisted of 3 convolutional layers with filter size 5×5 and pooling size 3×3. The first layer used 32 channels and max pooling, the second layer used 32 channels and average pooling, and the third layer used 64 channels and average pooling. Afterwards, the network used two fully convolutional layers consisting of 64 and 2 channels, respectively. Each layer except the last used ReLU activation function.

For the training we used stochastic gradient descent with learning rate equal to 0.000001, momentum equal to 0.9, weight decay equal to 0.001 and batch size equal to 1. We used cross entropy as a loss function. Training lasted for 40000 iterations. During the training we augmented the images with a random horizontal flip and a random rotation by a multiple of 90°.

Prior to feeding the image to the network its size was reduced to 350×230. Additionally, a global per-channel mean was subtracted from every image. The network was supplied with a 64×64 image patches. During training they were selected randomly from the image. During evaluation multiple patches were extracted from the underlying image with a stride of 32, as well as a set of all of their possible augmentations. The individual patch predictions were averaged to obtain the final prediction for the whole image.

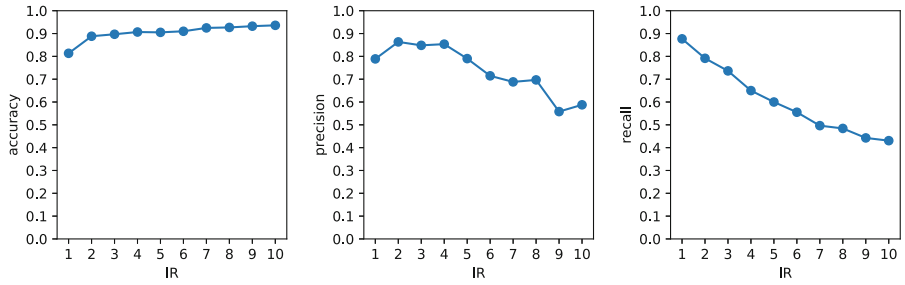
Strategies of Dealing with Imbalance. Various approaches to dealing with data imbalance have been proposed in the literature. They can be divided into

inbuilt mechanisms, which adjust the behavior of existing classifiers to better accommodate for data imbalance, and resampling strategies, in which either some of the majority samples are omitted (undersampling) or new minority samples are created (oversampling) to achieve a balanced training data distribution. In total, we evaluated 8 different strategies of dealing with data imbalance. Weighted loss (W. Loss), a strategy of assigning a weight associated with misclassification of an object based on its class. Specifically, we used a heuristic described in [3], and assigned the class weight as $w_i = \exp(-r_i)$, with r_i indicating the ratio of class i in the training data. Batch balancing (B. Balance), a strategy of randomly selecting an equal number of minority and majority samples for every batch. The batch size was increased to 2 in case of batch balancing strategy. Random oversampling (ROS), a technique of randomly duplicating some of the minority samples up to the point of achieving class balance. SMOTE [2], an approach in which instead of duplicating existing objects, a synthetic minority observations are produced. In this method new observations are generated by interpolating between original observations. CCR [8], an oversampling strategy that uses smaller local translations instead of interpolating between possibly far-away observations. In addition to oversampling, this method translates the existing majority observations to increase their distance from minority class boundary. RBO [7], another translation-based synthetic oversampling technique, that additionally considers the position of majority objects in process of oversampling. Random undersampling (RUS), a technique of randomly selecting only a subset of majority observations. And the Neighborhood Cleaning Rule (NCL) [10], a guided undersampling strategy, in which neighborhood-based approach is used to guide the process of data cleaning.

Evaluation. Since classification accuracy is not an appropriate metric to assess the classification performance in the imbalanced data setting, throughout the conducted experimental study we use five additional metrics: precision, recall, geometric mean (G-mean), F-measure and AUC. More detailed discussion on the choice of performance metrics can be found in [5] and [6].

2.2 The Impact of Data Imbalance on the Classification Performance

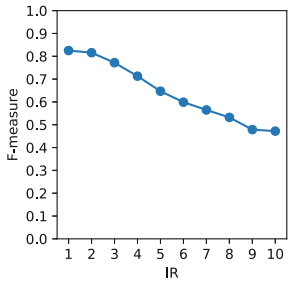
The goal of the first experiment was evaluating to what extent data imbalance affects the classification performance. To this end we undersampled the original BreakHis dataset up to the point of achieving the desired imbalance ratio (IR), at the same time keeping the total number of observations from both classes constant. We considered IR $\in \{1.0, 2.0, \ldots, 10.0\}$. Results of this part of the experimental study, averaged over all folds and magnification factors, were presented in Fig. 2. As can be seen, the accuracy is not an appropriate performance metric in the imbalanced data setting: it increases steadily with IR, despite the accompanying decrease in both precision and recall. On the other hand, all of the remaining measures indicate a significant drop in performance, especially for higher values of IR. For instance, for the balanced distributions we observed



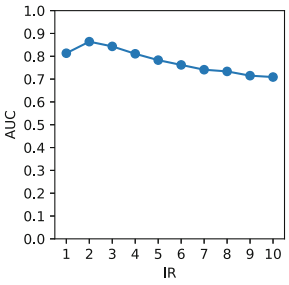
(a) accuracy

(b) precision

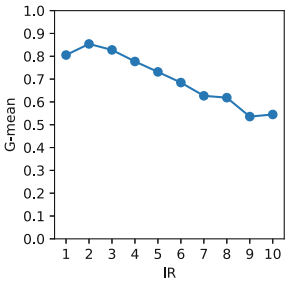
(c) recall



(d) F-measure



(e) AUC



(f) G-mean

Fig. 2. The impact of data imbalance ratio (IR) on the average values of various performance metrics.

average value of F-measure above 0.8, whereas for the IR = 10.0 it drops below 0.5, despite the total number of observations being the same. This indicates that data imbalance has a significant impact on the classifiers behavior and a noticeable decrease in performance can be expected for higher IR. It should be noted that for low values of IR, that is 2 and 3, we actually observed better precision, AUC and G-mean than for the balanced data distribution. This behavior may suggest that depending on our optimization criterion, slight data imbalance can actually be beneficial for the performance of the model. In the case of the histopathological data, especially if the majority class consists of the images of malignant tissue.

2.3 The Evaluation of Strategies of Dealing with Data Imbalance

The goal of the second experiment was comparing various strategies of dealing with data imbalance and assessing which, and under what conditions, lead to the best performance. In this experiment we considered the values of IR $\in \{2.0, 3.0, \ldots, 10.0\}$, and grouped the imbalance into four categories: low (2.0–4.0), medium (5.0–7.0), high (8.0–10.0) and any (2.0–10.0). For each category the results were averaged over the corresponding values of IR. We considered the

Table 1. Average ranks achieved by various techniques of dealing with data imbalance for the specified imbalance ratio (IR). Best performance observed for a given ratio was denoted in bold. The number of times a method achieved statistically significantly better or worse performance than one of the other methods was denoted in subscript with, respectively, a plus or a minus sign.

			Inbuilt mechanisms		**Oversampling strategies**				**Undersampling strategies**	
	IR	Baseline	W. Loss	B. Balance	ROS	SMOTE	CCR	RBO	RUS	NCL
Precision	2-4	**1.17**$_{+6, -0}$	2.83$_{+4, -0}$	7.33$_{+0, -3}$	6.42$_{+0, -2}$	4.92$_{+0, -1}$	5.08$_{+0, -1}$	6.42$_{+0, -2}$	7.50$_{+0, -3}$	3.33$_{+2, -0}$
	5-7	**2.33**$_{+4, -0}$	4.00$_{+0, -0}$	6.75$_{+0, -2}$	6.83$_{+0, -2}$	4.67$_{+0, -0}$	3.92$_{+0, -0}$	6.50$_{+0, -2}$	7.25$_{+0, -2}$	2.75$_{+4, -0}$
	8-10	**3.00**$_{+1, -0}$	3.67$_{+1, -0}$	6.00$_{+0, -0}$	5.58$_{+0, -0}$	5.00$_{+0, -0}$	4.58$_{+0, -0}$	5.83$_{+0, -0}$	7.25$_{+0, -2}$	4.08$_{+0, -0}$
	2-10	**2.17**$_{+6, -0}$	3.50$_{+4, -0}$	6.69$_{+0, -4}$	6.28$_{+0, -3}$	4.86$_{+1, -1}$	4.53$_{+2, -1}$	6.25$_{+0, -3}$	7.33$_{+0, -5}$	3.39$_{+4, -0}$
Recall	2-4	8.92$_{+0, -6}$	7.42$_{+0, -5}$	3.83$_{+2, -0}$	3.50$_{+2, -0}$	6.67$_{+0, -1}$	4.00$_{+2, -0}$	3.54$_{+2, -0}$	**2.62**$_{+3, -0}$	4.50$_{+1, -0}$
	5-7	8.75$_{+0, -5}$	7.33$_{+0, -4}$	3.54$_{+3, -0}$	2.62$_{+3, -0}$	5.75$_{+0, -1}$	4.38$_{+1, -0}$	**2.33**$_{+4, -0}$	3.17$_{+3, -0}$	7.12$_{+0, -4}$
	8-10	8.83$_{+0, -5}$	7.25$_{+0, -4}$	2.88$_{+3, -0}$	3.54$_{+3, -0}$	6.00$_{+0, -2}$	4.83$_{+1, -0}$	2.21$_{+4, -0}$	**2.04**$_{+4, -0}$	7.42$_{+0, -4}$
	2-10	8.83$_{+0, -7}$	7.33$_{+0, -5}$	3.42$_{+4, -0}$	3.22$_{+4, -0}$	6.14$_{+1, -4}$	4.40$_{+3, -0}$	2.69$_{+4, -0}$	**2.61**$_{+4, -0}$	6.35$_{+1, -5}$
F-measure	2-4	4.25$_{+0, -0}$	3.67$_{+0, -0}$	6.58$_{+0, -1}$	5.50$_{+0, -0}$	5.67$_{+0, -0}$	4.25$_{+0, -0}$	5.33$_{+0, -0}$	7.00$_{+0, -1}$	**2.75**$_{+2, -0}$
	5-7	7.17$_{+0, -1}$	6.42$_{+0, -0}$	5.08$_{+0, -0}$	4.25$_{+0, -0}$	4.92$_{+0, -0}$	**3.17**$_{+1, -0}$	4.25$_{+0, -0}$	5.08$_{+0, -0}$	4.67$_{+0, -0}$
	8-10	7.33$_{+0, -4}$	5.83$_{+0, -0}$	3.83$_{+1, -0}$	3.67$_{+1, -0}$	5.75$_{+0, -0}$	3.75$_{+1, -0}$	**3.42**$_{+1, -0}$	5.08$_{+0, -0}$	6.33$_{+0, -0}$
	2-10	6.25$_{+0, -1}$	5.31$_{+0, -0}$	5.17$_{+0, -0}$	4.47$_{+0, -0}$	5.44$_{+0, -0}$	**3.72**$_{+1, -0}$	4.33$_{+0, -0}$	5.72$_{+0, -0}$	4.58$_{+0, -0}$
AUC	2-4	6.50$_{+0, -0}$	4.58$_{+0, -0}$	5.96$_{+0, -0}$	4.58$_{+0, -0}$	5.50$_{+0, -0}$	4.29$_{+0, -0}$	4.67$_{+0, -0}$	5.17$_{+0, -0}$	**3.75**$_{+0, -0}$
	5-7	8.75$_{+0, -6}$	7.42$_{+0, -4}$	4.17$_{+1, -0}$	3.17$_{+2, -0}$	5.08$_{+1, -0}$	3.50$_{+2, -0}$	**3.12**$_{+2, -0}$	3.33$_{+2, -0}$	6.46$_{+0, -0}$
	8-10	8.58$_{+0, -5}$	7.17$_{+0, -4}$	2.75$_{+4, -0}$	3.42$_{+3, -0}$	6.25$_{+0, -3}$	4.50$_{+1, -0}$	**2.33**$_{+4, -0}$	2.67$_{+4, -0}$	7.33$_{+0, -4}$
	2-10	7.94$_{+0, -7}$	6.39$_{+0, -5}$	4.29$_{+2, -0}$	3.72$_{+3, -0}$	5.61$_{+1, -1}$	4.10$_{+2, -0}$	**3.38**$_{+4, -0}$	3.72$_{+3, -0}$	5.85$_{+1, -3}$
G-mean	2-4	7.00$_{+0, -0}$	5.42$_{+0, -0}$	5.08$_{+0, -0}$	4.58$_{+0, -0}$	5.92$_{+0, -0}$	4.08$_{+0, -0}$	4.33$_{+0, -0}$	4.67$_{+0, -0}$	**3.92**$_{+0, -0}$
	5-7	8.75$_{+0, -6}$	7.58$_{+0, -4}$	4.33$_{+1, -0}$	3.08$_{+2, -0}$	5.33$_{+1, -0}$	3.67$_{+2, -0}$	3.00$_{+2, -0}$	**2.92**$_{+3, -0}$	6.33$_{+0, -1}$
	8-10	8.67$_{+0, -5}$	7.33$_{+0, -4}$	2.67$_{+4, -0}$	3.50$_{+3, -0}$	6.25$_{+0, -3}$	4.58$_{+1, -0}$	**2.25**$_{+4, -0}$	2.58$_{+4, -0}$	7.17$_{+0, -4}$
	2-10	8.14$_{+0, -7}$	6.78$_{+0, -5}$	4.03$_{+2, -0}$	3.72$_{+4, -0}$	5.83$_{+1, -3}$	4.11$_{+2, -0}$	**3.19**$_{+4, -0}$	3.39$_{+4, -0}$	5.81$_{+1, -3}$

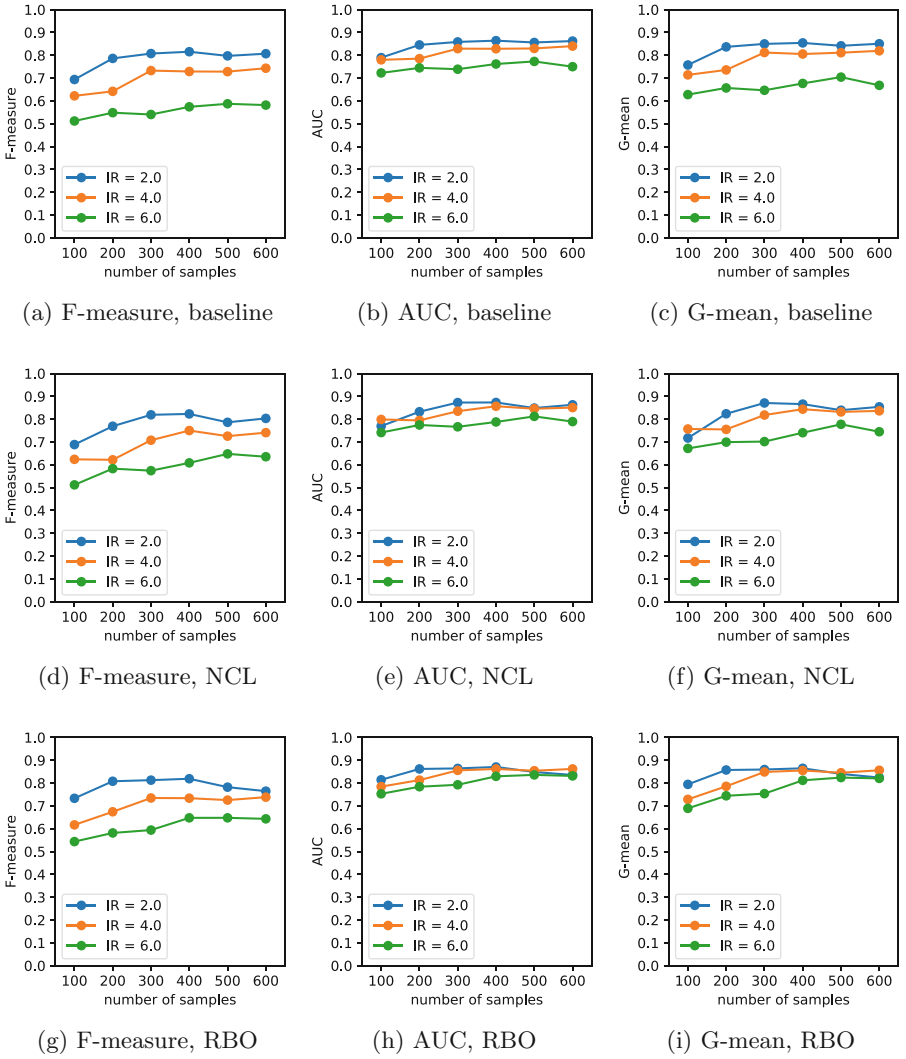
strategies described in Sect. 2.1, as well as the baseline case, in which no strategy was applied. To assess the statistical significance of the results we performed a Friedman ranking test with a Shaffer post-hoc analysis at the significance level $\alpha = 0.05$. The results were presented in Table 1. As can be seen, there was no single method that achieved best performance on all levels of imbalance and for all of the performance measures. In general, CCR, RBO, RUS and NCL methods achieved the highest rank in at least one of the settings. For low imbalance levels NCL achieved the best performance for all three combined metrics: F-measure, AUC and G-mean. However, in none of the cases did it achieve a statistically significantly better results than the baseline. For higher levels of imbalance RBO achieved the best rank in most cases, with statistically significant differences. While most of the approaches led to an improvement in performance compared to the baseline at least in some settings, two methods, weighted loss and SMOTE, achieved a noticeably worse performance than the other strategies.

2.4 The Value of New Data in the Presence of Data Imbalance

The goal of the third experiment was evaluating to what extent increasing the amount of training data improves the performance for various levels of imbalance. We considered the total number of training observations $\in \{100, 200, \ldots, 600\}$, and IR $\in \{2.0, 4.0, 6.0\}$. In addition to the baseline case, in which no strategy of dealing with imbalance was employed, we used two best-performing resampling techniques: NCL and RBO. The average values of the combined performance measures were presented in Fig. 3. As can be seen, in the baseline case data imbalance decreases the value of new observations. For the case of IR = 6.0, even after increasing the number of training samples six times, we did not achieve the same performance as the one observed for IR = 4.0, for any of the considered metrics. In other words, even when we used more training data from both minority and majority distributions, due to the inherent data imbalance we achieved a worse performance. To a smaller extent this trend is visible also between IR = 2.0 and IR = 4.0, especially when F-measure is considered. Using one of the resampling techniques prior to classification partially reduced this trend: in this case, after increasing the number of samples we were able to outperform the case with 100 training samples.

2.5 The Strategy of Balancing Training Distribution During Data Acquisition

In the previous experiments, while adjusting the imbalance ratio we modified both training and test data distributions. However, when dealing with real data we do not have an option of adjusting test distribution. Still, in some cases we can influence the imbalance of training data: for instance, in the case of histopathological images we can have at our disposal a larger quantity of unannotated images, and the main cost is associated with the annotation process. We can, therefore, select the images designed for annotation so that their distribution is balanced. The goal of the final experiment was evaluating whether



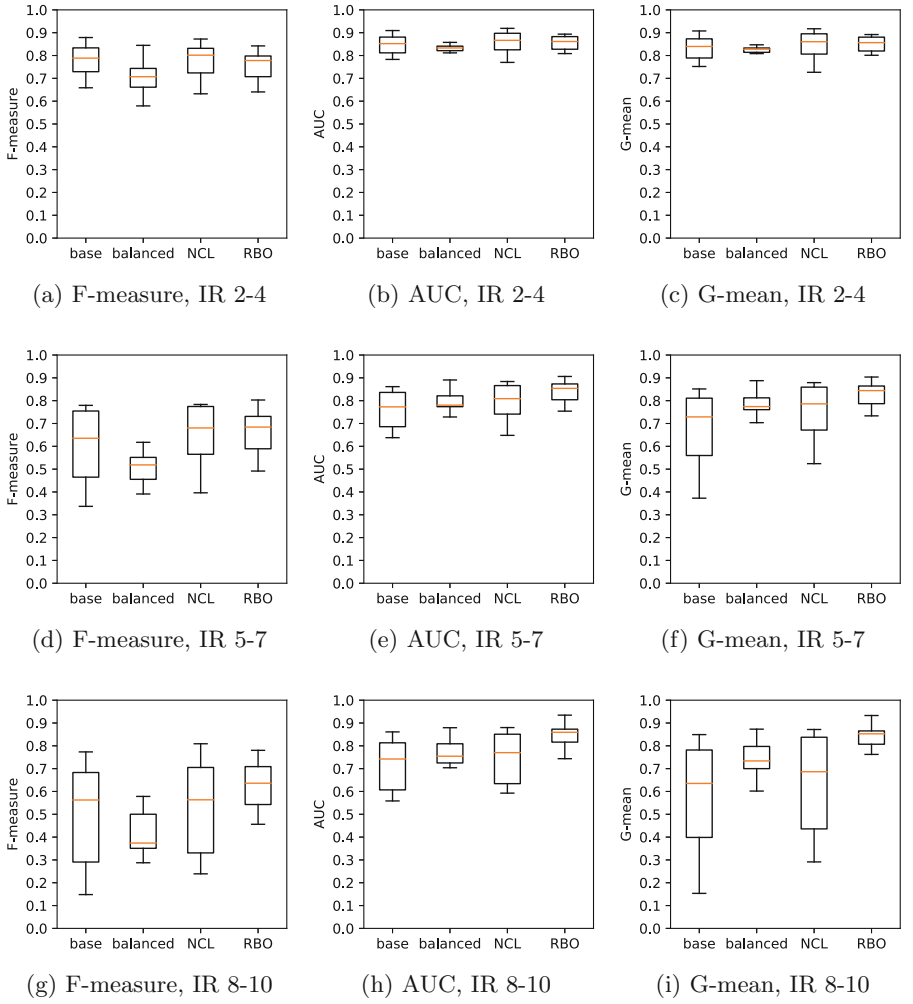
(a) F-measure, baseline (b) AUC, baseline (c) G-mean, baseline

(d) F-measure, NCL (e) AUC, NCL (f) G-mean, NCL

(g) F-measure, RBO (h) AUC, RBO (i) G-mean, RBO

Fig. 3. The impact of the number of training observations on average values of various performance metrics, either on the original data (top row), after undersampling with NCL (middle row) or oversampling with RBO (bottom row).

such data acquisition strategy is beneficial for the classification performance. To this end we evaluated two variants: the baseline case, in which both training and test data distribution were imbalanced with IR $\in \{2.0, 3.0, \ldots, 10.0\}$, and the balanced case, in which only test distribution was imbalanced and training data consisted of an equal number of samples from both classes. We presented the results of this experiment in Fig. 4. For reference, we also included the performance observed on data balanced with NCL and RBO. As can be seen, for



(a) F-measure, IR 2-4 (b) AUC, IR 2-4 (c) G-mean, IR 2-4

(d) F-measure, IR 5-7 (e) AUC, IR 5-7 (f) G-mean, IR 5-7

(g) F-measure, IR 8-10 (h) AUC, IR 8-10 (i) G-mean, IR 8-10

Fig. 4. Average values of various performance metrics. Baseline case, in which both training and test data was imbalanced, was compared with the case in which only test data was imbalanced. Performance for NCL and RBO was also included for reference.

low values of IR we actually observed a worse performance after balancing the training data according to all of the combined performance metrics. This trend was most noticeable for F-measure. Furthermore, the observed F-measure was also higher in the baseline case for higher IR. On the other hand, balancing training data improved the AUC and G-mean for medium and high levels of imbalance. In all of the cases, using the original, imbalanced training data distribution and balancing it with one of the considered resampling strategies led to an improvement in performance.

3 Conclusions

In this paper we experimentally evaluated the impact of data imbalance on the classification performance of convolutional neural network in breast cancer histopathological image recognition task. We conducted our analysis on the Breast Cancer Histopathological Database (BreakHis) [14]. The main findings of our experiments are the following:

- Medium and high data imbalance levels have a significant negative impact on the classification performance, irregardless of the chosen performance measure. However, for some of the considered measures, at low level of imbalance we observed an improved performance, which may suggest that small data imbalance can actually be beneficial in a specific settings. Especially the latter finding should be further confirmed on additional benchmark datasets.
- Some of the popular strategies of dealing with data imbalance, namely using weighted loss and oversampling data with SMOTE, significantly underperformed in the conducted experiments. Techniques that achieved the best results were NCL and RBO resampling algorithms. This leads us to a conclusion that developing a novel strategies of handling data imbalance, designed specifically for dealing with images, might be necessary to achieve a satisfactory performance in the histopathological image recognition task.
- Data imbalance negatively impacts the value of additional training data. Even when more data from both minority and majority class was used, due to data imbalance we were unable to achieve a performance observed for lower imbalance ratios. This can be partially mitigated by using an appropriate strategy of handling data imbalance.
- Depending on data imbalance ratio and the metric used to measure classification performance, balancing training data during acquisition can have a negative impact on the performance when compared to sampling training data with the same imbalance ratio as test data. In all of the considered cases, applying resampling on imbalanced data was preferable approach to balancing data during acquisition.

Since the conducted analysis based on a single benchmark dataset, further research should be focused on extending it to additional databases. Furthermore, a limited number of already proposed strategies dedicated to dealing with image imbalance should be included in the method comparison. Design of a novel methods is also likely necessary to be able to achieve a satisfactory performance.

Acknowledgment. This research was supported by the National Science Centre, Poland, under the grant no. 2017/27/N/ST6/01705 and the PLGrid infrastructure.

References

1. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. arXiv preprint arXiv:1710.05381 (2017)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
3. Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. arXiv preprint arXiv:1804.10851 (2018)
4. Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R.: Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. **47**, 45–67 (2018)
5. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
6. Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, Cambridge (2011)
7. Koziarski, M., Krawczyk, B., Woźniak, M.: Radial-based approach to imbalanced data oversampling. In: Martínez de Pisón, F.J., Urraca, R., Quintián, H., Corchado, E. (eds.) HAIS 2017. LNCS (LNAI), vol. 10334, pp. 318–327. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59650-1_27
8. Koziarski, M., Woźniak, M.: CCR: a combined cleaning and resampling algorithm for imbalanced data classification. Int. J. Appl. Math. Comput. Sci. **27**(4), 727–736 (2017)
9. Krawczyk, B.: Learning from imbalanced data: open challenges and future directions. Prog. Artif. Intell. **5**(4), 221–232 (2016)
10. Laurikkala, J.: Improving identification of difficult small classes by balancing class distribution. In: Quaglini, S., Barahona, P., Andreassen, S. (eds.) AIME 2001. LNCS (LNAI), vol. 2101, pp. 63–66. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-48229-6_9
11. Lusa, L., et al.: SMOTE for high-dimensional class-imbalanced data. BMC Bioinform. **14**(1), 106 (2013)
12. Pulgar, F.J., Rivera, A.J., Charte, F., del Jesus, M.J.: On the impact of imbalanced data in convolutional neural networks performance. In: Martínez de Pisón, F.J., Urraca, R., Quintián, H., Corchado, E. (eds.) HAIS 2017. LNCS (LNAI), vol. 10334, pp. 220–232. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59650-1_19
13. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: Breast cancer histopathological image classification using convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2560–2567. IEEE (2016)
14. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2016)