Handling Unbalanced Data in Deep Image Segmentation

Harriet Small
Brown University
harriet_small@brown.edu

Jonathan Ventura
University of Colorado, Colorado Springs
jventura@uccs.edu

***Abstract*— We approach the problem of training Convolutional Neural Networks (CNNs) for image segmentation tasks that involve unbalanced data—meaning that some of those classes we seek to identify and label occur with significantly less frequency than other classes represented in the dataset. We investigate alternative sampling strategies intended to increase the accuracy of the learned model itself and neutralize misclassifications arising from the unbalanced nature of the training data, and examine their efficacy in comparison to random sampling.**

Keywords: class imbalance, image segmentation, convolutional neural networks, machine learning

I. INTRODUCTION

One pervasive challenge in the field of deep image segmentation is the unbalanced distribution of classes in much training data [7], [8]. If pixels corresponding to a particular "majority" label are far more numerous than pixels of one or more "minority" class, the rarity of the "minority" class in the training data inhibits accurate learning and labeling, as the learned model will tend to classify most pixels as members of the "majority" class. As class imbalance in a data set increases, the performance of a neural net trained on that data has been shown to decrease dramatically [6].

The segmentation of MRI images is one notable application of deep learning in which such a class imbalance exists; in a typical MRI image of a brain tumor, the volume of healthy brain tissue is significantly greater than the volume of cancerous tissue [4]. For the purposes of this paper, we trained our neural networks on the BraTS Challenges 2013 dataset, which is comprised of MRI images of brain tumors.

Using these MRI images, we explore techniques for surmounting learning obstacles introduced by unbalanced training data. In particular, our focus is on modifying the procedure by which we sample batches to train a Convolutional Neural Network (CNN) intended to classify unbalanced data. We compare the performance of random sampling with two alternatives: a sampling protocol that generates batches containing each class in equal proportion, and a second protocol which re-introduces incorrectly classified (and borderline correctly-classified) samples from prior epochs into the batches for the current epoch. We evaluate the efficacy of these three methods by examining their effect on the correct labeling of small tumor substructures.

II. PRIOR RESEARCH

A variety of attempts to rectify the class imbalance problem have been made. In a survey study, López et al. identified

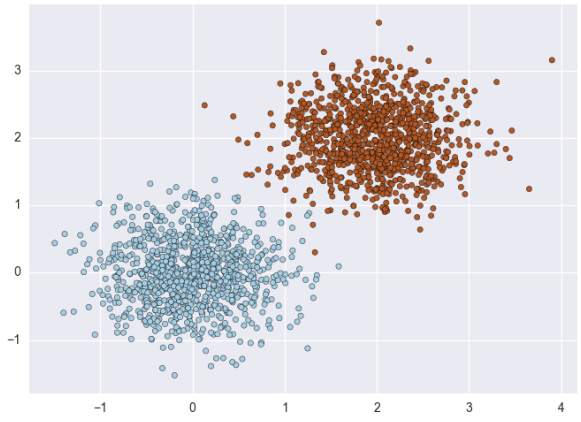

Fig. 1. A typical example of a balanced data distribution. Note that there are approximately the same number of examples of the red and blue classes. Compare with unbalanced distribution in next image. Source: https://svds.com/learning-imbalanced-classes/

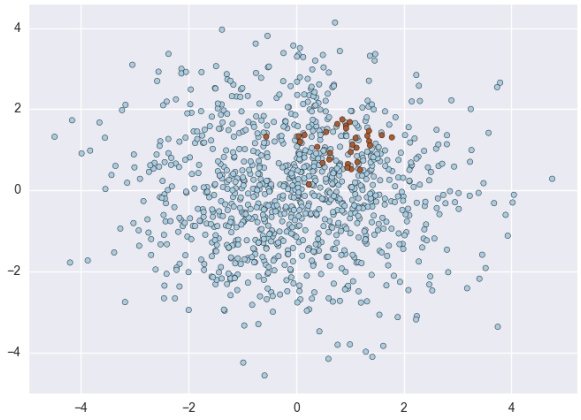

Fig. 2. An unbalanced data distribution; note that the vast majority of samples are of the blue class, and that there are comparatively few red examples. Source: https://svds.com/learning-imbalanced-classes/

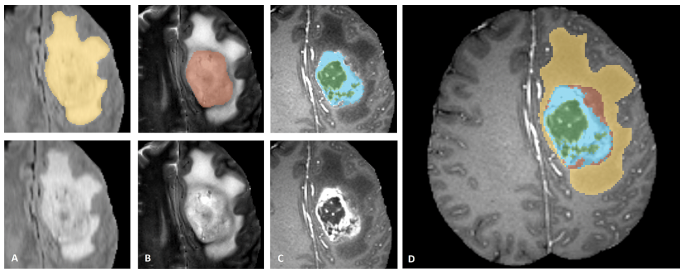

Fig. 3. A brain tumor image from BraTS annotated with four types of diseased tissue. Note that the number of pixels representing tumor tissue is much smaller than the total number of pixels [3].

three notable types of solutions: modification of data prior to learning via oversampling or undersampling, modification of the learning process itself, and application of cost-sensitive learning techniques, which weigh the relative "costs" of misclassification for each class against each other [8].

One notable approach—which falls into the first of the above categories—is the Synthetic Minority Over-Sampling Technique, or SMOTE. This technique involves generating synthetic samples of the minority class to train on, thus reducing the class imbalance by artificially inflating the size of the minority class itself [9]. This strategy, when tested alongside undersampling of the majority class, was shown to improve the performance of the trained model. Our methodology also focuses on modifying the class distribution in the dataset, although we will use only data from the original set rather than replicating additional minority samples.

A considerable amount of prior research has focused on the application of cost-sensitive learning techniques to the class imbalance problem. Often, the real-world misclassification cost of a minority sample is greater than the misclassification cost of a majority sample; when identifying a rare disease, for example, a false positive has the potential to be less damaging to the patient than a false negative. Researchers have incorporated this concern into learning algorithms by modifying the loss function at the center of the learning process to overvalue classification mistakes on the minority class, thus emphasizing the correct classification of minority samples at the expense of identifying the majority class. This type of cost-sensitive technique, while not a part of our approach, is certainly relevant in the area of tumor segmentation; the misclassification of cancerous tissue as healthy is far more costly to a patient than the misclassification of healthy tissue as cancerous [8].

One sampling-based attempt to counteract the negative effects of an imbalanced dataset was presented by Felzenszwalb et al. in their work on object detection [1]. Their dataset consisted of images with large amounts of negative space with interspersed objects; the relative rarity of the objects themselves demanded a nuanced approach. Their technique, "hard negative mining", involved identifying those examples of the majority class—the background—which the current classifier was not correctly labeling, and reintroducing those examples for further training. Focusing on "harder" negative examples in this manner allowed them

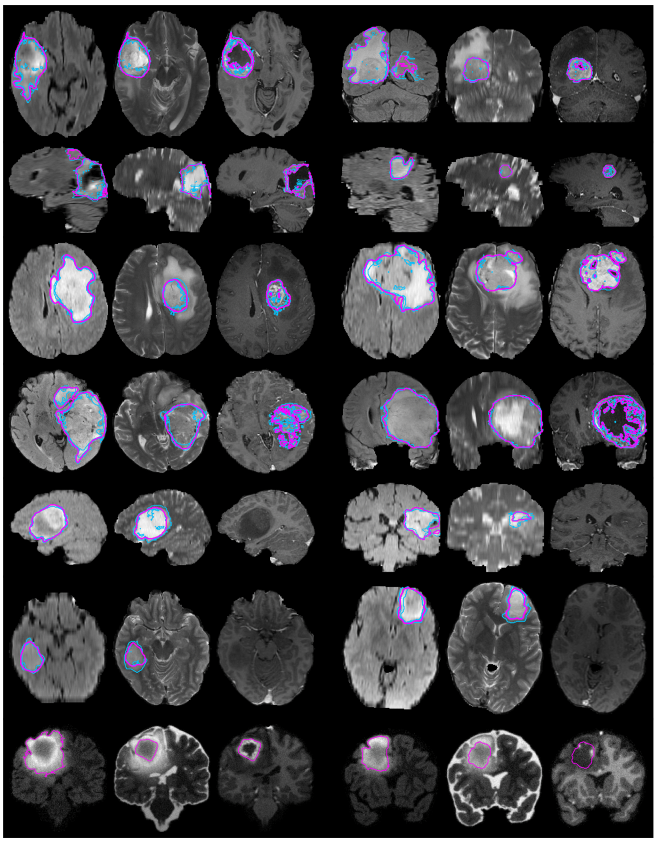

Fig. 4. A small subset of the MRI images from the BraTS 2013 Challenge Dataset, with expert annotations of the four tumor substructures shown in purple and blue. Image from [3].

to increase the fraction of minority samples used during the training process without sacrificing performance on the majority class.

Initially, our intent was to bring "hard negative mining" to bear on our own classifier, but experimentation revealed that increasing the fraction of minority class samples in training batches did not reduce accuracy on the majority class, and thus this corrective process was not required. However, we did attempt a modified version of the strategy (targeting minority classes rather than simply the majority class), an approach which is described in more detail in a later section of this paper.

III. EXPERIMENTAL DATA

For the purposes of our experimentation, we use the non-synthetic data from the 2013 edition of the Multimodal Brain Tumor Image Segmentation (BraTS) Challenge. This data set is a series of 65 MRI images of brain tumors which must be segmented into healthy tissue and four differing types of cancerous tissue herein referred to as tumor substructures. The substructures labeled in the training data are as follows:

- Edema
- Non-Enhancing (Solid) Core
- Necrotic (Fluid-Filled) Core
- Non-Enhancing Core

The challenge dataset is a typical example of an unbalanced class problem; the images in it contain a disproportionate number of healthy tissue examples and only small areas of cancerous tissue. See Fig. 4 for the subset of the challenge images themselves. A more detailed description of the dataset can be found in Havaei et al. [4].

IV. METHODOLOGY

In order to determine the optimal sampling method for this unbalanced dataset, we trained three separate convolutional neural networks with consistent architecture. The first was trained using random sampling, and the second and third using our modified sampling techniques. We then compared their performance on the identification and segmentation of minority classes to evaluate the sampling methods themselves.

During each training epoch, we refined the parameters of our CNNs using a series of batches of the training data, each consisting of a number of square sections from training images, herein referred to as patches. This type of mini-batch sampling—taking a few pixels (or, in our case, patches) from each of a diverse set of training images—has been demonstrated effective in pixel-labeling problems such as edge-detection and image segmentation [5]. This strategy reduces the computational load of processing each batch by taking advantage of the dependencies between neighboring pixels; adjacent pixels tend to have similar surroundings and therefore including many neighboring pixels in a batch is redundant [5]. Each of the networks discussed below was trained for exactly 50 epochs, each consisting of 1000 individual batches of training data, each containing 120 image patches that were 64 pixels square.

V. EVALUATION OF RESULTS

Metrics for evaluating the correctness of models trained and tested on unbalanced data using can yield misleading results [7], [8]. A model which is well-attuned to the features of a majority class but has poor performance when labeling a minority class, for example, might have high overall accuracy as the test set contains mostly pixels of the majority class. However, such a model cannot be considered successful.

Several alternative metrics sensitive to data imbalance have been proposed and used to evaluate models of unbalanced data. For the purposes of this project, we adopt the measuring scheme used for the BRATS challenge [3], which uses the Dice score to quantify the overlap between the ground-truth area of a particular tissue type and the area labeled as that type by our classifier. For some tissue class c, let P be a binary map of every pixel in the image to 1 if it is a member of class c, 0 otherwise. Furthermore, let T be the ground-truth binary mapping. Let P_1 and T_1 be the sets of all pixels mapped to 1 by P and T respectively. The Dice score is calculated thus:

$$Dice(P,T) = \frac{|P_1 \cup T_1|}{(|P_1| + |T_1|)/2}$$

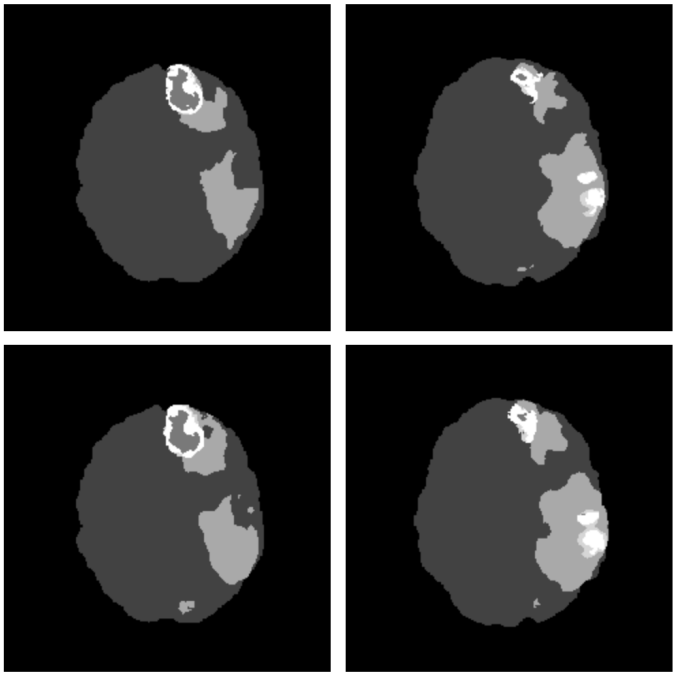

Fig. 5. The top two images are the ground truth labelings for two brain scans, the second row contains the segmentations done by the balanced-sampling model.

That is, the score for a particular class c is the size of the overlap between the predicted region and its true counterpart, normalized by the averaged size of the two regions.

The BRATS benchmark adapts the binary Dice score to the multi-class segmentation problem by choosing a subset of the set of minority classes and treating all tissue types in that subset as a single class. The three subsets under consideration are the entire tumor (containing all four cancerous tissue types), the tumor excluding edema, and the enhancing core region, which consists of a single tissue class [3].

VI. EXPERIMENTS IN BALANCED SAMPLING

Initially, we investigated the impact of forcing each batch of training data to contain the same number of examples of each class represented in the dataset. Fig. 2 contains segmentations performed by this model alongside the ground truth labeling and segmentations produced by our baseline random-sampling model.

Note that the random model's segmentations tend to underestimate the amount of cancerous tissue contained in the scan, especially tissue of the rarest substructure types (those with the lightest pigmentation in the segmentations). That is, this model displays a typical failing of a classifier trained on unbalanced data: a tendency to overclassify the majority class at the expense of one or more minority classes. The balanced model does not display this tendency; instead, segmentations produced through balanced sampling tend to overestimate the area covered by the rare cancerous tissue types.

The overclassification vs. underclassification problem described above is evident in quantitative as well as qualitative

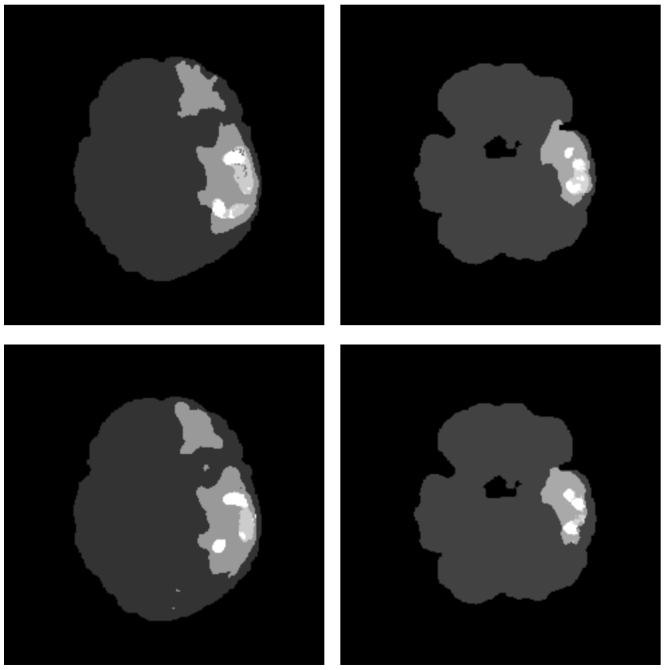

Fig. 6. The top two images are the ground truth labelings for two brain scans, the second row contains the segmentations done by the random sampling model.

evaluation of the generated segmentations. Below are the Dice Scores of the models trained using random and balanced sampling techniques.

	Whole Tumor	Core	Active Region
Random	0.863	0.776	0.786
Balanced	0.833	0.845	0.862

Note that although the average dice score of the random model over the entire area of the tumor is greater than the same score for the balanced model, the core and active-region scores for the balanced model are significantly higher. That is, those two scores most heavily dependent on the rarest tumor substructures are higher for the balanced model. However, this increased facility in identifying rare classes is coupled with a loss of facility in identifying more common classes. In an attempt to preserve the advantages of the balanced-sampling model and further improve its score on the full tumor region, we investigated the strategy of hard example mining.

VII. EXPERIMENTS IN HARD EXAMPLE MINING

In an attempt to improve upon the balanced sampling approach, we incorporated the concept of hard example mining. This technique was introduced by Felzenszwalb et al., who combat class imbalance in object detection with a technique they termed "hard negative mining" [1]. This technique involves constructing each training batch such that it contains a disproportionately large number of minority-class samples alongside a small subset of the majority-class samples that are deemed "hard" for the current classifier. A sample is considered "hard" with respect to a classifier if it was misclassified or only correctly classified by a small margin by that classifier in the previous epoch. The intent is to reduce the proportion of majority class examples without sacrificing the classifier's ability to identify the majority class by emphasizing those examples it fails to classify correctly.

However, our focus was on identifying and reintroducing hard samples of all classes rather than just hard majority samples. Adopting a balanced sampling approach (and thus necessarily reducing the frequency of the majority class) did not substantially reduce our classifier's ability to identify the majority class. The Dice score for healthy tissue in the model trained on balanced batches of patches was greater than 0.99 and only marginally less than that of the model trained on randomly selected batches. That is, capping the fraction of majority class samples per batch at just 20% of the batch size didn't negatively impact the model's ability to recognize that class. Therefore, Felzenszwalb et al.'s "hard negative" approach was not appropriate for our problem, because performance on the majority class did not need improvement. Instead, we brought their technique to bear on each of the minority classes—our rare tumor substructures—in hopes that it would further increase our classifier's ability to identify those classes.

To evaluate the efficacy of this hard example strategy, we trained a third CNN. Between each training epoch, we ran the current classifier on a randomly selected subset of each class and stored the indices of all samples in this subset which were misclassified, or correctly classified by a small margin (with less than 75% confidence). When generating batches for the following epoch, we drew first from these "hard" examples to fill out each batch. We continued to balance the proportions of each class in each batch, gathering a number of samples from each class equal to 20% of the overall batch size.

Ultimately, this approach proved less effective than simply using balanced sampling methods alone. The table below displays the comparative Dice scores for the three sampling strategies.

	Whole Tumor	Core	Active Region
Random	0.863	0.776	0.786
Balanced	0.833	0.845	0.862
Hard Example	0.810	0.801	0.835

Note that hard example mining fails to perform as well as balanced sampling at segmentation all of the three tumor categories. The segmentations themselves (images below) display the problem visually. Evidently, the model trained primarily on hard data has the tendency to overdraw the tumor region. Although the locality of the tumor itself is correct, this classifier exaggerates its size and fails to recognize its boundaries. Perhaps reintroducing hard examples to the learning process overemphasizes the those characteristics which result in a class being confused for another class,

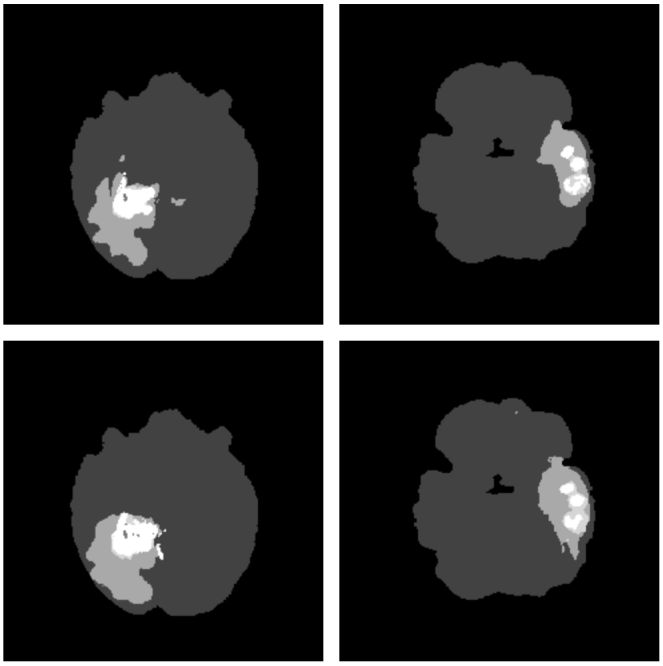

Fig. 7. The two bottom images are segmentations produced by the classifier trained on hard examples. Their corresponding ground truths are above.

ultimately blurring the boundary between the five classes in the dataset.

Although our experimentation with hard example mining failed to improve our classifier, the technique may merit further research, as discussed in the final section of this paper.

VIII. AREAS OF FURTHER EXPLORATION

Given the semi-promising results yielded by our hard example mining, I believe further exploration of this technique is certainly worthwhile. Further experiments could investigate the potential of hard example mining with differing ratios of majority and minority classes in training batches; for example, we could attempt to mimic the distribution of the minority classes in relation to each other in each training batch, while continuing to artificially skew the number of minority class samples in relation to the majority class, and combine this approach with hard example mining on those minority classes. It is also worth investigating whether this sampling technique is more effective in training on other datasets displaying the imbalance problem. If so, what characteristics of a dataset predict how useful the technique will prove in training a classifier for that data?

There is also further experimentation to be done in finding the ideal confidence threshold for identifying a "hard" example. It is possible that this sampling strategy proves more effective if only those samples which are actually misclassified are labeled hard, or, alternatively, when every sample not assigned to the correct class with 99% confidence is considered "hard."

A second promising further area of research is in mitigating the multi-class imbalance problem by training two separate classifiers, one for the majority class and another for the minority classes. The former would be trained to segment images into instances of the majority class and instances of any other class. The latter would be trained only on examples of the rarer classes and could then be used to further segment the non-majority pixels identified by the former into those rarer classes. In the context of the BraTS dataset, the first classifier would make a binary determination of whether each pixel represented healthy or diseased tissue, and the second classifier would identify the tumor substructures contained in the diseased patches identified by the first. Ideally, the second classifier would not be subject to the problem of class imbalance, as the diseased subset of the data has a much more balanced class distribution than the dataset as a whole.

Although much scholarship on the class imbalance problem exists already, the multitude of applications for powerful machine learning-based classifiers that can be trained on unbalanced data render further research and exploration in the area absolutely indispensable.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1659788 and Grant No. 1359275.

REFERENCES

[1] Felzenszwalb, Pedro, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model." In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1-8. IEEE, 2008.

[2] Zhou, Zhi-Hua, and Xu-Ying Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem." IEEE Transactions on Knowledge and Data Engineering 18, no. 1 (2006): 63-77.

[3] Menze, Bjoern H., Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging 34, no. 10 (2015): 1993-2024.

[4] Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. "Brain tumor segmentation with deep neural networks." Medical image analysis 35 (2017): 18-31.

[5] Bansal, Aayush, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. "Pixelnet: Towards a general pixel-level architecture." arXiv preprint arXiv:1609.06694 (2016).

[6] Mazurowski, Maciej A., Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance." Neural networks 21, no. 2 (2008): 427-436.

[7] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on knowledge and data engineering 21, no. 9 (2009): 1263-1284.

[8] López, Victoria, Alberto Fernández, Salvador Garca, Vasile Palade, and Francisco Herrera. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." Information Sciences 250 (2013): 113-141.

[9] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.