ADDRESSING THE CLASS IMBALANCE PROBLEM IN THE AUTOMATIC IMAGE CLASSIFICATION OF COASTAL LITTER FROM ORTHOPHOTOS DERIVED FROM UAS IMAGERY

D. Duarte [1,2*], U. Andriolo[2], G. Gonçalves [1,2]

[1] Department of Mathematics – University of Coimbra – Coimbra, Portugal (diogoavaduarte, gil)@mat.uc.pt
[2] INESC Coimbra, Coimbra, Portugal uandriolo@mat.uc.pt

Commission III / WG III/10

KEY WORDS: garbage, mapping, marine, convolutional neural networks, oversampling, class weighting, classifier thresholding

ABSTRACT: Unmanned Aerial Systems (UAS) has been recently used for mapping marine litter on beach-dune environment. Machine learning algorithms have been applied on UAS-derived images and orthophotos for automated marine litter items detection. As sand and vegetation are much predominant on the orthophoto, marine litter items constitute a small set of data, thus a class much less represented on the image scene. This communication aims to analyse the class imbalance issue on orthophotos for automated marine litter items detection. In the used dataset, the percentage of patches containing marine litter is close to 1% of the total amount of patches, hence representing a clear class imbalance issue. This problem has been previously indicated as detrimental for machine learning frameworks. Three different approaches were tested to address this imbalance, namely class weighting, oversampling and classifier thresholding. Oversampling had the best performance with a f1-score of 0.68, while the other methods had f1-score value of 0.56 on average. The results indicate that future works devoted to UAS-based automated marine litter detection should take in consideration the use of the oversampling method, which helped to improve the results of about 7% in the specific case shown in this paper.

1. INTRODUCTION

Coastal marine litter has a negative impact on marine ecosystems (Rochman *et al.*, 2016), marine life (Kühn *et al.*, 2015) coastal communities (Beaumont *et al.*, 2019) and human health (Werner *et al.*, 2016). It is therefore essential to find proper solutions for mapping and monitoring marine litter load, in order to identify the sources, to propose mitigation measures, and to support cleaning operations (Galgani *et al.*, 2013). Traditionally, the monitoring of marine litter is often based on in-situ visual census surveys (Cheshire *et al.*, 2009), which are expensive in term of human effort and logistically limited (Lavers and Bond, 2017).

Recent studies have proposed the use of Unmanned Aerial Systems (UAS) for detecting and mapping marine litter pollution on coastal environment (Fallati *et al.*, 2019; Gonçalves *et al.*, 2019). UAS offers a cost-efficient collection of the required high-resolution imagery for the detection of marine meso-litter items (comprised between 2.5 cm and 50 cm) on beaches and dunes. Furthermore, UAS allow a fast coverage of a wider area when compared to in-situ visual census. In spite of this capabilities, manual identification of marine litter items on UAS-derived orthophotos is subjective and time-consuming. In fact, the operator needs to manually localize and characterize each marine litter object on the orthophoto, which is a time-consuming task and its quality directly related with the experience of the operator. As a consequence, automated detection of marine litter items would be preferable, as it would allow a faster and robust image processing.

To date, a few approaches made use of image recognition algorithms to perform image classification or image segmentation methods to automatically identify litter on UAS images (Martin *et al.*, 2018; Fallati *et al.*, 2019) and derived orthophotos (Gonçalves *et al.*, 2019). Martin et al. (2018), used a random forest algorithm, obtaining a significant over-estimation of marine litter items. The authors indicated that for the binary classifier (litter and no litter), only 10% of the total number of image samples were from the litter class. Gonçalves et al. (2019) proposed a random forest algorithm based on colour features to segment litter instances on orthophotos. In this approach the authors used a pixel level approach, which required as input both the orthophoto and the masks delineating the litter objects. Also in this case, only a small percentage of the total study area contained litter.

Convolutional neural networks have recently shown better performance for image recognition tasks on remote sensing optical images (Huang *et al.*, 2017; Maggiori *et al.*, 2017) when compared with traditional machine learning algorithms such as random forests. Regarding marine litter mapping, Fallati et al. (2019), used a commercial software which follows an object detection framework using CNN. Other details from the approach are not available given the commercial nature of the software used for the image classification step.

Overall, all the works that considered an automated approach for marine litter detection on UAS-derived imagery experienced the issue of class imbalance, as only a very small percentage of the area contains litter in comparison with sand and vegetation classes (Martin *et al.*, 2018; Fallati *et al.*, 2019; Gonçalves *et al.*, 2019, 2020). However, the class imbalance issue has been overlooked and not properly analysed, despite the fact that class imbalance has been shown to negatively affect the performance of traditional classifiers such as Support Vector Machine, and Random Forest (Japkowicz and Stephen, 2002; Mazurowski *et*

al., 2008), and convolutional neural networks (Buda *et al.*, 2018).

The attempts proposed to address the class imbalance issue can be divided in two types of methods: a) data and b) algorithmic methods. Data level methods focus on the training data and on class distribution to reduce/eliminate class imbalance. Two main approaches can be distinguished: oversampling (Johnson *et al.*, 2013; Douzas *et al.*, 2019) and undersampling (Leichtle *et al.*, 2017; Buda *et al.*, 2018). These methods use new image samples generation (oversampling) or modify the subset sample selection (undersampling). Another approach regarding data level methods is the definition of weights for each of the samples according to its class. Hence, an image sample of the majority class will have a smaller input weight than a minority class when computing the training loss. Algorithmic methods instead adjust the training procedure and/or inference step. For instance, classifier thresholding (Buda *et al.*, 2018) modifies the pre-defined threshold probability of the classifier to consider a given example as positive. In a one-class classification, also known as novelty detection, a network is trained to recognize a given target class instead of discriminating between two classes (Deng *et al.*, 2018). Other algorithmic methods focus directly on the loss function which penalizes a certain type of errors (Dong *et al.*, 2019). Overall, oversampling is recurrently indicated as the best performing approach in systematic studies (Buda *et al.*, 2018) and for different remote sensing applications (Johnson *et al.*, 2013; Douzas *et al.*, 2019; Xia *et al.*, 2019).

This work proposes to analyse and address the class imbalance issue in the automatic image classification of marine litter on coastal environment. Image classification is performed with a CNN framework, and three approaches are tested to deal with the class imbalance problem which are compared against a reference experiment (baseline) where the class imbalance is not taken into account. The work is structured as follows. In section 2, we present the study area, dataset, a brief description of the CNN and where the approaches to address the class imbalance problem are described. Section 3 presents the results, followed by Discussion in section 4. Conclusions are reported in Section 5.

2. METHOD AND STUDY AREA

This section describes the methodological approach to address the class imbalance issue regarding the image classification of litter using orthophotos derived from UAS images. First, the study area and dataset are described in section 2.1. The used CNN is briefly described in section 2.2, while the focus of this paper, the methods to address the class imbalance problem, are defined in section 2.3.

2.1 Study site and dataset

The study area was Cabedelo beach, in Figueira da Foz, Portugal (Figure 1). The camera gimbal of the UAS (DJI Phantom 4) was set to −90° to capture photos perpendicular to the direction of the flight and the ISO, shutter speed and aperture were set to 100, 1/1250 s and f/3.2, respectively. This allowed to generate an orthophoto of 5 mm ground sampling distance (GSD).The orthophoto was generated using 432 UAS images with a resolution of 4864 × 3648 pixels which were acquired at 20 m flying height and overlapped with 80% front and 65% side rates in order to derive the corresponding digital surface model generated by dense image matching.

The area covered by the orthophoto was divided following a geomorphological procedure (Gonçalves et al. 2019), in order to obtain the beach and dune zones. Within each of these two zones, three areas were selected (and numbered) to train and test the automated algorithms. Areas 1-3-4-6 were used to train, while Area 2 and 5 were used to test the machine learning algorithms on beach and dune zones, respectively.

It can be observed how the beach zone background is mostly composed by sand, with some sectors that were covered by wood and grass, mostly stranded by waves. The dune zone background was instead more variable, with the presence of sand, different types of vegetation and wooden paths. Marine litter items were mostly found on beach zone (Gonçalves et al. 2019), however some items were also present on dunes. Some examples of marine litter objects visible on the orthophoto can be seen in Figure 2, where the different backgrounds of beach and dune can also be compared.

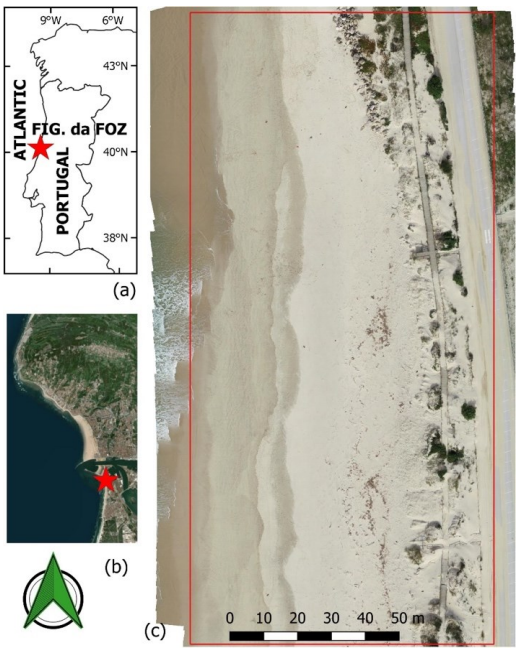

Figure 1 Study site map (a and b) and the orthophoto used in this work (c).

The marine litter objects visible in the training areas were digitalized by an experienced operator to create digital masks of the objects (Figure 3 a). The image Areas (1-6) were extracted from the orthophoto and divided in smaller 64x64px image patches, as required by CNN architecture (see section 2.2). The masks were used to categorize each of the extracted image patches as *litter* or *not litter*. If an image patch contained 8 or more *litter* pixels, this was considered as a *litter* image patch else it was considered as a *no litter* one. The threshold value of 8 pixels was set on the rational that, given the orthophoto GSD of 5.5 mm, the smallest marine meso-litter object (2.5 cm), which represent the lowest limit of our target, would be represented by 9 pixels.

Table 1 presents the final amount of image patches for training and validation areas and also discretized between dunes and beach. It is worth noting that the marine litter items were less than 1% of the total number of image samples. Hence, a highly imbalanced dataset. The ratio of samples used in validation is ~25% in the case of no litter class and ~37% in the case of the litter class.

2.2 Convolutional neural network

In this work it was used an adaptation of densenet121 (Huang *et al.*, 2017). Densenet121 capacity was greatly reduced given that the original network was developed for computer vision datasets like ImageNET, where hundreds of classes may be considered and where each class may contain thousands of samples. Hence, by decreasing the complexity of the network, overfitting is attenuated and the model is faster when considering a binary image classification problem.

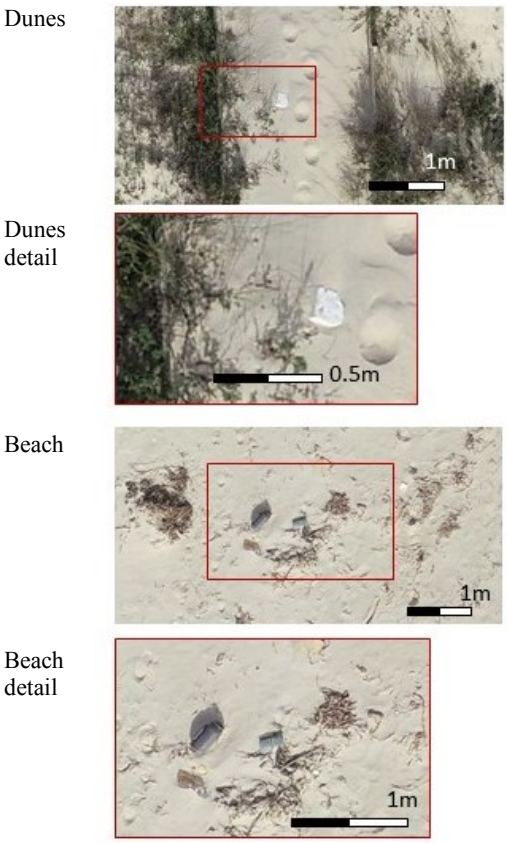

Figure 2 Details of the areas considered in this study, dunes and beach.

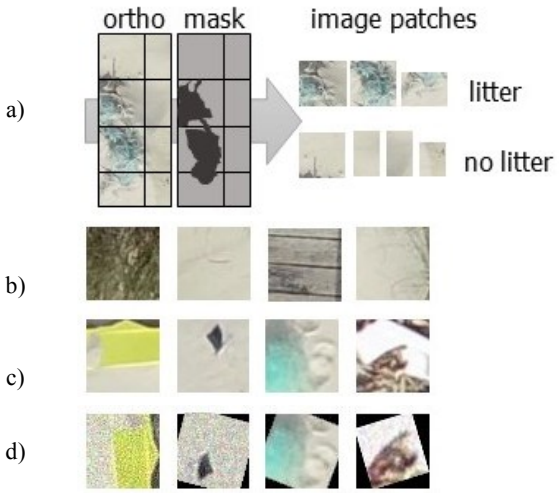

Figure 3. Examples of image samples extracted from the ortho. a) . Ortho and manually defined mask. On the right the resulting image patches, b) litter and c) oversampled litter patches (see section 3.3.2).

	Beach		Dunes		Total	
	No litter	litter	No litter	litter	No litter	litter
Training (1,3,4,6)	36352	307	3990	28	40342	335
Validation (2,5)	11165	171	1985	24	13150	195
Total	**47517**	**478**	**5975**	**52**	**53492**	**530**

Table 1 Total number of image samples for the training and validation and also considering separately the beach and dune areas.

The densenet is a network built by stacking dense blocks which in turn are composed by convolutional sets. The convolutional sets are concatenated within each dense block. This allows to keep feature information throughout the network, where between dense blocks there is a transitional block which reduces the size of the feature maps to a final size of 4x4px (from the initial 64x64px). Three dense blocks, composed each by 2, 3 and 6 convolutional sets each, formed the main body of the architecture. Hence, from the original network, one dense block was removed and the remaining had the number of convolutional sets reduced (Figure 4). In the classification step a sigmoid activation function was used, where the final output was the probability of a given patch to contain litter.

The batch size was considered as 600, which is the maximum the hardware would allow in order to have as many marine litter instances per batch as possible. Considering that the network had 530 litter and 53492 no litter image patches, choosing the batch size as 600 allowed to have about 6 litter patches per batch. Data augmentation was used with the objective of reducing overfitting (Krizhevsky *et al.*, 2012) where random translations, rotations and shear were applied during training. This was performed for all experiments.

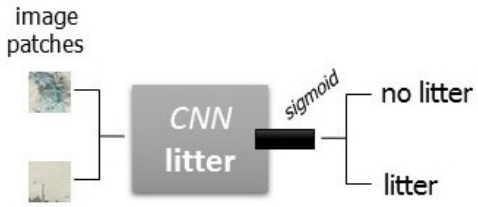

Figure 4 . General steps for the image classification of litter. Image patches are fed to the network where a densenet based CNN extracts features from the images which are then classified with a sigmoid activation function.

2.3 Addressing class imbalance

Three different approaches for dealing with class imbalance were tested in this work: oversampling, class weighting, and classifier thresholding. In general, the aim of the methods was to automate the image classification process where parameters could be drawn directly from the data. The reference experiment (baseline) does not consider any method to address the class imbalance problem as it currently happens in the literature.

2.3.1 Oversampling: While undersampling could also be used, it was assumed that valuable feature information could be lost due to this. Instead we used oversampling as already indicated as preferable by Buda et al. (2018). The authors also indicated that oversampling with CNN does not lead to overfitting as would happen with traditional image classification approaches. Hence, in this study the minority class (litter) was oversampled to match the same amount of no litter samples. This was only performed during training, where the validation contained only original samples. To perform this, random image transformations were applied to the images and new ones were generated, see Figure 3. These transformations were random noise, random rotation and zoom, which were applied randomly to the litter image patches. In this way the final number of samples for training was balanced regarding the number of image samples per class.

2.3.2 Class weighting: This aims at balancing the training procedure by changing the weight that each training sample has when computing the training loss. In this case, class weighting was introduced, since the weight of each sample was related to its class. The weights of a class were tied to its proportion of image samples in the dataset. The litter instances are only 1% of the whole dataset, hence a weight of 100 has been given to that class, while the no litter class was 1. This approach is named w100 from now on.

2.3.3 Thresholding: Thresholding is an approach in which the classification threshold of a classifier is modified. There are several ways of defining this threshold to accommodate class imbalance. (Richard and Lippmann, 1991; Lawrence *et al.*, 1998; Buda *et al.*, 2018). A common approach, which can be driven by the imbalance ratio is for the threshold to accommodate this imbalance. In this case the threshold was inversely proportional to the ratio of litter/no litter samples (only ~1% of litter instances are litter). Hence, when predicting, an image sample is considered to have litter only if its prediction score probability is higher than 0.99. The thresholding was applied to all the experiments (class weighting, baseline and oversampling).

3. RESULTS

Statistical metrics of each of the approaches are presented in (Table 2). Within brackets the results of the thresholding approach for each of the experiments is also presented. The results are presented for both the beach and dune areas. The total presented in the table was determined by adding the true positives, false positives and false negatives of both the areas and consequently determine the total precision recall and f1-scores.

While baseline, weighting and oversampling were standalone methods, the thresholding was applied to each of the experiments. Examples of the results on both the beach and dunes (Figure 5 and Figure 6) are presented, followed by the plotting of the precision-recall (pr) curve of the main set of experiments.

Overall, the oversampling method achieved the highest f1-score. This is more accentuated when considering only the beach area. On the other hand, the thresholding strategy was the worst performing experiment. In general thresholding improved recall but decreased precision, worsening the f1-score of the experiments. While in the beach area, the w100 is still comparable to the baseline experiment, this is not the case in the dunes area. The difference between the baseline and oversampling experiment is also bigger on the beach region, where that difference is shortened in the dunes area.

Beach

	f1	precision	recall
baseline(t)	0.56(0.21)	0.73(0.91)	0.45(0.12)
w100(t)	0.57(0.47)	0.59(0.83)	0.55(0.32)
oversampling(t)	**0.63**(0.57)	0.62(0.46)	0.65(0.77)

Dunes

	f1	precision	recall
baseline(t)	0.65(0.30)	1.00(1.00)	0.48(0.17)
w100(t)	0.52(0.65)	0.45(1.00)	0.61(0.48)
oversampling(t)	**0.68**(0.63)	0.67(0.92)	0.70(0.48)

Total

	f1	precision	recall
baseline(t)	0.57(0.35)	0.76(0.22)	0.46(0.84)
w100(t)	0.55(0.24)	0.55(0.14)	0.55(0.78)
oversampling(t)	**0.64**(0.38)	0.62(0.25)	0.66(0.80)

Table 2. Statistical measures of the experiments, f1 score, recall and precision for all the experiments and considering the beach, dunes and all the image data. Between brackets the thresholded version of the experiments (t).

Regarding the precision and recall of the experiments, the w100 maintains the same value for both of these metrics. Hence, the model is having a similar rate of false negatives and false positives. On the other hand, the baseline experiment contains a higher rate of false negatives when compared with the false positives. The oversampling approach is also balanced in terms of the false positive and true positive rate.

Figure 5 and Figure 6, present the results for the dune and beach area, respectively. For each experiment details of the classified orthophotos are presented. True positives (green), false negatives (blue) and false positives (red). In Figure 5, the baseline experiment was only able to detect 1 of the litter instances, while the oversampling detected them all. The weighting approach also detected one more litter instance, however, at the expense of having more false positives.

In Figure 6 the baseline experiment only detected one litter patch correctly. This experiment, while not having false positives, contains several false negatives. The oversampling and weighting present the same number of false positives; however, the oversampling approach correctly detects much more litter patches than the class weighting method.

The precision-recall curve shows the tradeoff between precision and recall for different classification thresholds. This gives us a general performance visualization of a given model, which is especially relevant when the aim is to compare performances. It is also indicated when the datasets are highly imbalanced (Saito and Rehmsmeier, 2015). The higher the curve is on the y axis the better the model performs. This performance can be quantified by calculating the area under the curve (AUC). In Figure 7, the precision-recall curves of each of the main experiments are assessed, baseline, weighting and oversampling. The AUC is 0.67 for the baseline, 0.66 for the oversampling and 0.55 for the weighting. Hence, this chart indicates the baseline experiment as the best model, even if just marginally. Moreover, it is visible from the figure that there are

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-439-2020 | © Authors 2020. CC BY 4.0 License.

regions in the chart where the baseline experiment is above the oversampling one.

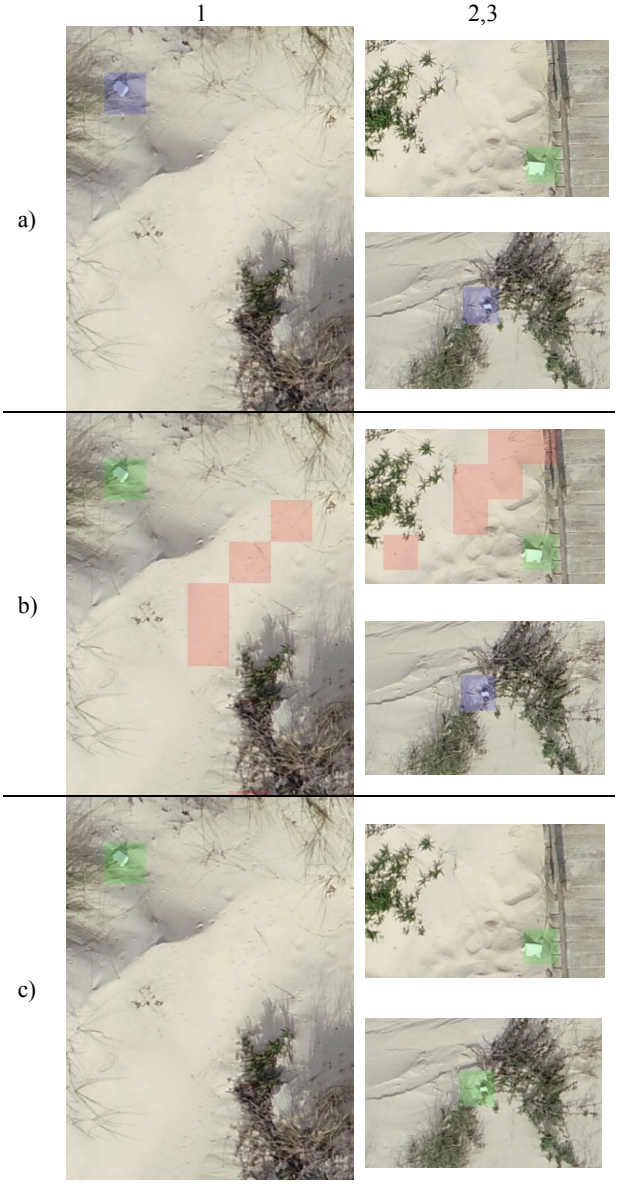

Figure 5. Details of the results for the dunes. Blue tiles are false negatives, red tiles false positives and green tiles true positives. Detail 1, 2 and 3 considering each of the approaches (a – baseline, b – w100 and c – oversampling).

4. DISCUSSION

Both the statistical metrics and the presented image patches with results point to the oversampling approach as the best approach to deal with the class imbalance problem. This happened when testing against classification thresholding, class weighting and against a baseline experiment where the class imbalance was not taken into account. These results are in agreement with the literature when addressing the class imbalance problem: oversampling approaches often are the best performing ones (Johnson *et al.*, 2013; Buda *et al.*, 2018; Douzas *et al.*, 2019), even if in this case a binary classification problem was considered and an imbalance ratio of around 1:100.

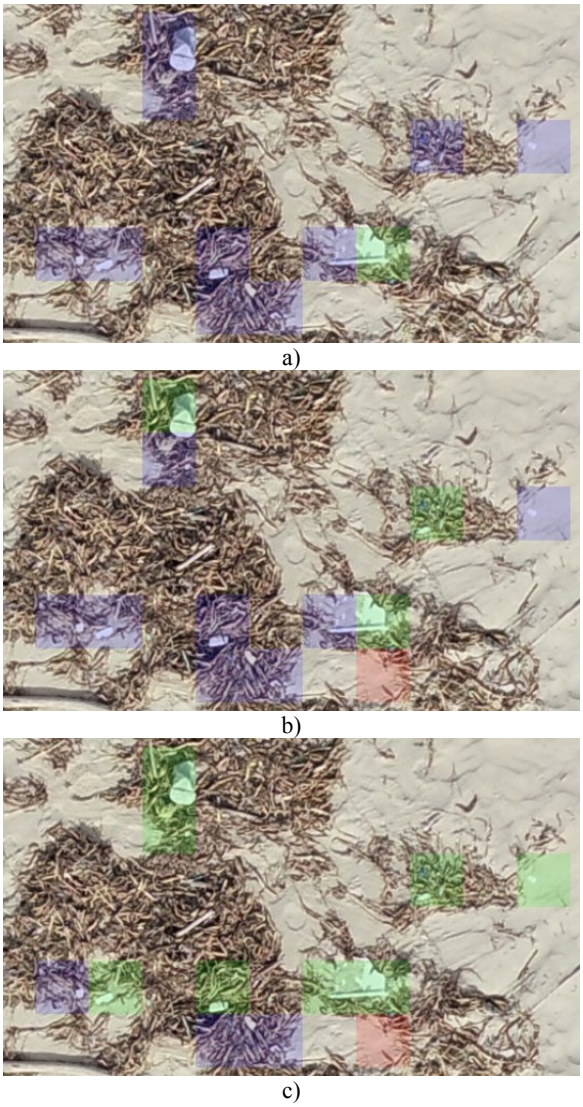

Figure 6. Details of the results for the beach. Blue tiles are false negatives, red tiles false positives and green tiles true positives. a) baseline, b) w100 and c) oversampling.

Nonetheless the precision-recall curve indicates a smaller difference between oversampling and baseline experiment where the latter even had a marginal improvement when it comes to the pr-AUC. Looking at the pr curve the baseline seems to perform in pair with the oversampling. This indicates that the oversampling approach may not have more recognition capabilities but instead allows to follow the common classification threshold of > 0.5 to consider an image sample as one containing litter. In the baseline approach the search for an optimal classification threshold would have to be empirically tested. This is translated in a more automated process, which does not rely on the *a posteriori* empirical search for the optimal classification threshold.

In fact, throughout this paper, the automation of the parameters was central. The parameters for both the weighting and thresholding scheme were derived from the data and the imbalance rate presented in the image data. Such thresholds are not optimal (e.g. weighting approach); however, they enable the automation of the training procedure, which is not dependent on empirically tested thresholds on a given set of data.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-439-2020 | © Authors 2020. CC BY 4.0 License.

While the oversampling approach worked better it was also the slowest during training given the higher amount of image samples. Nonetheless, testing time is the same given the methods to address the class imbalance were all either data or classifier level methods, where the network was always the same.

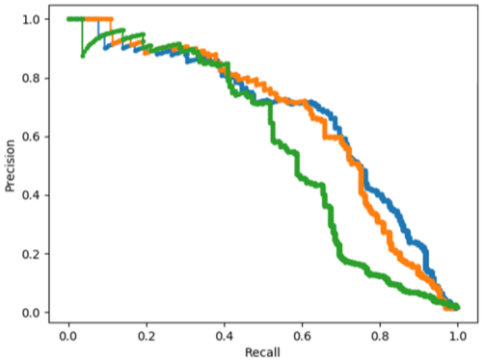

Figure 7. Precision-Recall curves for the 3 main experiments (baseline-blue, oversampling-orange and w100-green) considering all the validation data (from both the beach and dunes).

In this work a patch-based image classification approach was tested to derive a binary classifier to identify coastal litter. Such approach is less computationally expensive when comparing with pixel based approaches (Fallati *et al.*, 2019; Gonçalves *et al.*, 2019) and where the labelling can be more coarse than the delineation of each litter instance. This greatly reduces the effort in the labelling of the data to feed to the networks.

The binary nature of the approach also does not allow for the identification of each of the localized litter instances. However, given the data limitation this was not possible and needs to be addressed in a future work, when more image data is available. To this regard it may be useful to also consider several classes of no litter (Martin *et al.*, 2018), however, this would make the approach more location dependent (vegetation, sidewalks, rock formations, etc.). A one-class classification approach (Deng *et al.*, 2018) could also be tested given the highly imbalanced data (~1% of litter image samples) instead of the presented binary approach. Even more relevant if new advancements in computer vision such as generative adversarial networks are considered for one class classification (Akcay *et al.*, 2019).

Geographical transferability must also be tested. This test only addressed a single orthophoto from a given region and considering only an UAS survey. More image data from other captures in different beach-sand systems need to be captured in order to test the approach with different image characteristics, resolution and different beach-dune systems. This must be tested to assess the generalization power when applied to different beach-dune systems.

To note that even if using a GSD of 5 mm, it is often difficult to assess the difference between litter and other objects present in the scene, such as rocks and more dense vegetation.

5. CONCLUSIONS

The use of UAS have been recently applied for marine litter mapping on beach-dune environment. The related works have experienced the class imbalance issue when machine learning algorithms have been applied on marine litter objects detection, as sand and vegetation are predominant on the orthophoto.

In this work, we applied a Convolutional Neural Network to automatically detect marine litter object on a beach-dune systems. In the image data used, image patches containing litter instances were less than 1% of the total number of samples, thus a clear class imbalance was observed.

Three different approaches were tested to address the class imbalance, based on data level and algorithm level methods, namely oversampling, class weighting and thresholding. From the experiments, the oversampling approach achieved best results, whereas a priori class weighting inverse to the class presence performed worst. The classifier thresholding applied to the baseline, weighting and oversampling approaches, overall performed worst than the baseline approach.

Considering a monitoring application, more work is needed. For example, in this study only a dataset was used. More tests need to assess the transferability of the method when considering different surveys with different flying parameters, camera, lighting conditions, needs to be assessed. Moreover, the geographical transferability of the method should also be tested, given that it is time consuming to generate location specific training data.

ACKNOWLEDGEMENTS

This work was supported by the Portuguese Foundation for Science and Technology (FCT) and by the European Regional Development Fund (FEDER) through COMPETE 2020 - Operational Program for Competitiveness and Internationalization (POCI) in the framework of the strategic project UIDB/00308/2020 and the research project UAS4Litter (PTDC/EAM-REM/30324/2017).

REFERENCES

Akcay S, Atapour-Abarghouei A, Breckon TP. 2019. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In: Jawahar CV, Li H, Mori G and Schindler K (eds) *Computer Vision – ACCV 2018*. Springer International Publishing: Cham, 622–637.

Beaumont NJ, Aanesen M, Austen MC, Börger T, Clark JR, Cole M, Hooper T, Lindeque PK, Pascoe C, Wyles KJ. 2019. Global ecological, social and economic impacts of marine plastic. *Marine Pollution Bulletin*, 142: 189–195. https://doi.org/10.1016/j.marpolbul.2019.03.022.

Buda M, Maki A, Mazurowski MA. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Cheshire A, Adler E, Barbière J. 2009. *UNEP/IOC guidelines on survey and monitoring of marine litter*. United Nations Environment Programme, Regional Seas Programme ; Intergovernmental Oceanographic Commission, Integrated Coastal Area Management and Regional Programme: Nairobi : Paris.

Deng X, Li W, Liu X, Guo Q, Newsam S. 2018. One-class remote sensing classification: one-class vs. binary classifiers. *International Journal of Remote Sensing*, 39(6): 1890–1910. https://doi.org/10.1080/01431161.2017.1416697.

Dong Q, Gong S, Zhu X. 2019. Imbalanced Deep Learning by Minority Class Incremental Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6): 1367–1381. https://doi.org/10.1109/TPAMI.2018.2832629.

Douzas G, Bacao F, Fonseca J, Khudinyan M. 2019. Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the Geometric SMOTE Algorithm. *Remote Sensing*, 11(24): 3040. https://doi.org/10.3390/rs11243040.

Fallati L, Polidori A, Salvatore C, Saponari L, Savini A, Galli P. 2019. Anthropogenic Marine Debris assessment with Unmanned Aerial Vehicle imagery and deep learning: A case study along the beaches of the Republic of Maldives. *Science of The Total Environment*, 693: 133581. https://doi.org/10.1016/j.scitotenv.2019.133581.

Galgani F, Hanke G, Werner S, Oosterbaan L, Nilsson P, Fleet D, Kinsey S, Thompson RC, Van Franeker J, Vlachogianni T, Scoullos M, Veiga JM, Palatinus A, Matiddi M, Maes T, Korpinen S, Budziak A, Leslie H, Gago J, Liebezeit G. 2013. *Guidance on monitoring of marine litter in European seas*. Publications Office of the European Union: Luxembourg.

Gonçalves G, Andriolo U, Pinto L, Bessa F. 2019. Mapping marine litter using UAS on a beach-dune system: a multidisciplinary approach. *Science of The Total Environment*, 135742. https://doi.org/10.1016/j.scitotenv.2019.135742.

Gonçalves G, Andriolo U, Pinto L, Duarte D. 2020. Mapping marine litter with Unmanned Aerial Systems: A showcase comparison among manual image screening and machine learning techniques. *Marine Pollution Bulletin*, 155: 111158. https://doi.org/10.1016/j.marpolbul.2020.111158.

Huang G, Liu Z, Maaten L van der, Weinberger KQ. 2017. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE: Honolulu, HI, 2261–2269.

Japkowicz N, Stephen S. 2002. The class imbalance problem: A systematic study1. *Intelligent Data Analysis*, 6(5): 429–449. https://doi.org/10.3233/IDA-2002-6504.

Johnson BA, Tateishi R, Hoan NT. 2013. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. *International Journal of Remote Sensing*, 34(20): 6969–6982. https://doi.org/10.1080/01431161.2013.810825.

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1907–1105.

Kühn S, Bravo Rebolledo EL, van Franeker JA. 2015. Deleterious Effects of Litter on Marine Life. In: Bergmann M, Gutow L and Klages M (eds) *Marine Anthropogenic Litter*. Springer International Publishing: Cham, 75–116.

Lavers JL, Bond AL. 2017. Exceptional and rapid accumulation of anthropogenic debris on one of the world's most remote and pristine islands. *Proceedings of the National Academy of Sciences*, 114(23): 6052–6055. https://doi.org/10.1073/pnas.1619818114.

Lawrence S, Burns I, Back A, Tsoi AC, Giles CL. 1998. Neural Network Classification and Prior Class Probabilities. In: Orr GB and Müller K-R (eds) *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg: Berlin, Heidelberg, 299–313.

Leichtle T, Geiß C, Lakes T, Taubenböck H. 2017. Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 60: 83–98. https://doi.org/10.1016/j.jag.2017.04.002.

Maggiori E, Tarabalka Y, Charpiat G, Alliez P. 2017. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2): 645–657. https://doi.org/10.1109/TGRS.2016.2612821.

Martin C, Parkes S, Zhang Q, Zhang X, McCabe MF, Duarte CM. 2018. Use of unmanned aerial vehicles for efficient beach litter monitoring. *Marine Pollution Bulletin*, 131: 662–673. https://doi.org/10.1016/j.marpolbul.2018.04.045.

Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2–3): 427–436. https://doi.org/10.1016/j.neunet.2007.12.031.

Richard MD, Lippmann RP. 1991. Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation*, 3(4): 461–483. https://doi.org/10.1162/neco.1991.3.4.461.

Rochman CM, Browne MA, Underwood AJ, van Franeker JA, Thompson RC, Amaral-Zettler LA. 2016. The ecological impacts of marine debris: unraveling the demonstrated evidence from what is perceived. *Ecology*, 97(2): 302–312. https://doi.org/10.1890/14-2070.1.

Saito T, Rehmsmeier M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3): e0118432. https://doi.org/10.1371/journal.pone.0118432.

Werner S, Budziak A, Franeker J van, Galgani F, Hanke G, Maes T, Matiddi M, Nilsson P, Oosterbaan L, Priestland E, Thompson R, Veiga J, Vlachogianni T. 2016. *Harm caused by Marine Litter: MSFD GES TG Marine Litter - thematic report*. .

Xia W, Ma C, Liu J, Liu S, Chen F, Yang Z, Duan J. 2019. High-Resolution Remote Sensing Imagery Classification of Imbalanced Data Using Multistage Sampling Method and Deep Neural Networks. *Remote Sensing*, 11(21): 2523. https://doi.org/10.3390/rs11212523.