A DEBIASING VARIATIONAL AUTOENCODER FOR DEFORESTATION MAPPING

M. X. Ortega Adarme[1,4]*, P. J. Soto Vega[2], G. A. O. P. da Costa[3], R. Q. Feitosa[4], C. Heipke[1]

[1] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany - (ortega, heipke)@ipi.uni-hannover.de
[2] LaTIM, INSERM, UMR 1101, University Brest, Brest, France - sotovega@univ-brest.fr
[3] State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil - gilson.costa@ime.uerj.br
[4] Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil - raul@ele.puc-rio.br

Commission III, WG III/4

KEY WORDS: Debiasing Variational Autoencoder, Deforestation Detection, Deep Learning, Semantic Segmentation.

ABSTRACT:

Deep Learning (DL) algorithms provide numerous benefits in different applications, and they usually yield successful results in scenarios with enough labeled training data and similar class proportions. However, the labeling procedure is a cost and time-consuming task. Furthermore, numerous real-world classification problems present a high level of class imbalance, as the number of samples from the classes of interest differ significantly. In various cases, such conditions tend to promote the creation of biased systems, which negatively impact their performance. Designing unbiased systems has been an active research topic, and recently some DL-based techniques have demonstrated encouraging results in that regard. In this work, we introduce an extension of the Debiasing Variational Autoencoder (DB-VAE) for semantic segmentation. The approach is based on an end-to-end DL scheme and employs the learned latent variables to adjust the individual sampling probabilities of data points during the training process. For that purpose, we adapted the original DB-VAE architecture for dense labeling in the context of deforestation mapping. Experiments were carried out on a region of the Brazilian Amazon, using Sentinel-2 data and the deforestation map from the PRODES project. The reported results show that the proposed DB-VAE approach is able to learn and identify under-represented samples, and select them more frequently in the training batches, consequently delivering superior classification metrics.

1. INTRODUCTION

The popularity of Deep Learning (DL) techniques has increased enormously (Wang et al., 2020) in recent years. Indeed, DL-based approaches are associated with exceptional advances across a wide range of applications, in a diverse variety of fields, such as medical image analysis and diagnosis, biometry, environmental monitoring, autonomous driving, and natural language processing among others. These systems need to be carefully conceived and properly developed, ideally taking into consideration the concept of fairness (Osoba, Welser IV, 2017), as the solutions have a direct impact on society. In that context, providing safe and fair systems has become a relevant topic in a research community concerned with guaranteeing long-term, successful implementation of such systems, while minimizing potential negative side effects of their continuous use (Mehrabi et al., 2021, Amini et al., 2019).

Conventionally, DL approaches achieve good performances in scenarios with enough labeled training data (Vardi, 2022). In that case, they are able to generalize successfully on the unobserved test data. Nevertheless, the lack of large sets of labeled data is inherent to many real-world problems, and such a condition can lead to the creation of biased systems.

In addition, the class imbalance problem is quite common in different DL applications (Thabtah et al., 2020). In such cases, the trained classifier may be biased towards the majority classes, and it is possible to reach a high overall accuracy by just assigning those classes to the test samples (Ali et al., 2019). For the minority classes, however, the classification performance may be very poor. We observe that such issue is particularly significant in change detection, as usually the instances of the changed class are much scarcer than those of the unchanged samples. The capacity to properly learn from imbalanced data is, therefore, of paramount importance for change detection applications.

To address the fairness and the bias issues, several methods have been proposed thus far. Typically, those methods are categorized based on the stage where the bias mitigation is performed. Three different categories have been considered (Friedler et al., 2019): pre-process, in-process; and post-process.

Pre-process algorithms involve the transformation of the training data before feeding them into the model. For instance, the authors of (Naseriparsa et al., 2020, More, 2016) proposed resampling techniques to generate synthetic samples of the minority class. Those methods, however, do not properly consider within-class variability, and can also oversample uninformative or noisy samples (Jiang et al., 2021). The authors of (Kamiran, Calders, 2012, Luong et al., 2011) propose reweighing or changing the labels of some samples to train a classifier with non-discrimination constraints. The target samples in those works are those closest to the decision boundary, which are usually critical in the classification prediction. Similarly, works for learning the latent structure of data have also been studied, e.g., (Feldman et al., 2015, Calmon et al., 2017, Louizos et al., 2015). Those include latent SVM (Felzenszwalb et al., 2008) and fair dimensionality reduction with Principal Component Analysis (PCA) (Samadi et al., 2018). More recently, DL-based approaches for generating fair representations have been proposed. In that group, Variational Autoencoders (VAE) (Kingma, Welling, 2013, Louizos et al., 2015, Liu et al., 2022) and adversarial

* Corresponding author

training-based methods (Edwards, Storkey, 2015, Xie et al., 2017, Madras et al., 2018, Feng et al., 2019, Ruoss et al., 2020) are commonly used.

For in-process mechanisms, fairness is addressed during the training stage. The authors of (Kamishima et al., 2012) included a regularization term representing a trade-off between classification accuracy and fairness in the loss function. Likewise, (Bechavod, Ligett, 2017) included terms into the loss function that penalize unfairness by minimizing the differences between the false positive and false negative rates.

Finally, the post-process mechanisms make decision fairer by further processing the output scores of the classifiers. As an example, (Hardt et al., 2016) ensures a non-discriminatory prediction by flipping some decisions of a classifier to equalize predictor odds. Additionally, a decoupling system is proposed in (Corbett-Davies et al., 2017) to train independent classifiers for each class, and combine transfer learning approaches to learn from samples out of all classes.

In this work, we adapted the Debiasing Variational Autoencoder (DB-VAE), proposed by (Amini et al., 2019), for the task of semantic segmentation. Our approach is based on an end-to-end training scheme and employs the learned latent variables to adjust the individual sampling probabilities of data points during the training process. Thus, the method is able to identify types of samples that are under-represented in the training set, and to increase the likelihood that such instances are sampled during training. For semantic segmentation, we added an extra enconder-decoder module to the original architecture, composed of a series of convolutional operations followed by a softmax layer, which infers the probability of each class in a dense labeling fashion. We evaluated the proposed method on a deforestation detection application, using Sentinel-2 data from a region in the Amazon rain forest.

This paper is organized as follows: Section 2 describes the fully convolutional DB-VAE method used for deforestation detecting. Next, Section 3 presents the study area, the detailed experimental protocol, and the analysis of the obtained results. Finally, conclusions and final remarks are presented in Section 4.

2. FULLY DB-VAE

This section presents a brief description of the DB-VAE, originally proposed in (Amini et al., 2019), as well as the adaptation of the method to the semantic segmentation task.

Formally, let's consider a classification problem with d classes, and n training samples $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ indicates the features ($\mathbf{x} \in \mathbb{R}^m$), and $\mathbf{y}_i$ denotes the class label ($\mathbf{y} \in \mathbb{R}^d$) of the i-th sample. The goal is to find a nonlinear mapping $f_\theta : \mathbf{X} \rightarrow \mathbf{Y}$ parameterized by θ which minimizes a certain loss $\mathcal{L}$ over the entire training dataset. Then, given a new unobserved sample $(\mathbf{x}_i, \mathbf{y}_i)$, the classifier should ideally output $\hat{\mathbf{y}}_i = f_\theta(\mathbf{x}_i)$, where $\hat{\mathbf{y}}_i$ is "proximate" to $\mathbf{y}_i$, where the proximity is defined by the loss function. In addition, let's assume that each sample has an associated latent-space representation of dimension k ($\mathbf{z}_i \in \mathbb{R}^k$ and $k < m$), which compresses the feature information of the original input.

Inspired by (Amini et al., 2019), we propose a fully convolutional architecture for semantic segmentation. In particular, we focus on a change detection application, i.e., deforestation detection. Accordingly, the input of the neural network comprises a co-registered pair of images acquired at different dates: I_{t_0} and I_{t_1}. The images were concatenated along their spectral dimension, resulting in a tensor $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H and W refer to the spatial dimensions, and C to the number of image channels (two times those of each individual image).

Figure 1 shows the general design of the Fully DB-VAE. The encoder module of the VAE compresses the input samples and learns an approximation $q_\phi, (\mathbf{z}|\mathbf{x})$, parametrized by ϕ of the true distribution of the latent variables given a data point $\mathbf{x}$. The decoder reconstructs the input using the latent space representation by approximating $p_\psi(\hat{\mathbf{x}}|\mathbf{z})$, parametrized by ψ. VAEs parameterize the outputs as a normal distribution $\epsilon \sim \mathcal{N}(0, (I))$ and compute $\mathbf{z} = \mu(\mathbf{x}) + \sum^{1/2}(\mathbf{x}) \circ \epsilon$, where μ and σ are the mean and standard deviation of the latent variable distribution.

As we address a semantic segmentation approach, we added an extra encoder-decoder network to the architecture. That fully convolutional network is responsible for the pixel-wise classification, i.e., for assigning a class label to each input pixel location.

The whole model is trained in an end-to-end fashion; the loss function contains three components: a supervised latent loss (i.e., cross-entropy loss), a reconstruction loss (i.e., L2 norm), and a latent loss for the unsupervised variables (i.e., KL-divergence). The total loss $\mathcal{L}$ is a weighted sum of the three terms, and it is defined as follows:

$$\mathcal{L} = \lambda_s \left[\sum_{i=0}^{d} y_i log\left(\hat{y}_i\right) \right] + \lambda_r \left[\left\| \mathbf{x} - \hat{\mathbf{x}} \right\|_2 \right] + \lambda_u \left[\frac{1}{2} \sum_{i=0}^{k} \left(\sigma_i + \mu_i^2 - 1 - log\left(\sigma_i\right) \right) \right] \tag{1}$$

where λ_s, λ_r, and λ_u, represent the weights of the supervised latent loss, the reconstruction loss, and the unsupervised loss, respectively.

At each training epoch, we compute the histogram $\hat{\mathcal{Q}}_i$ of each latent variable $z_i(\mathbf{x})$ delivered by the VAE encoder. We further assume that the latent variables are statistically independent so that the probability of the latent representation $\mathbf{z}(\mathbf{x}) = [z_1(\mathbf{x}), ..., z_k(\mathbf{x})]$ of sample $\mathbf{x}$ is given by $\prod_i \hat{\mathcal{Q}}_i$.

The goal is that each subcategory of the training samples is equally represented in each batch. Therefore, the DB-VAE adopts the following probability of selecting a sample $\mathbf{x}$ for the next batch:

$$\mathcal{W}(\mathbf{z}(\mathbf{x}) \mid X) \propto \prod_i \frac{1}{\hat{\mathcal{Q}}_i\left(z_i(\mathbf{x}) \mid X\right) + \alpha} \tag{2}$$

where α is the debiasing parameter, which tunes the degree of debiasing during training. When $\alpha \rightarrow 0$, the samples of the training batch tend to follow a uniform distribution over the latent variables z. On the other hand, when $\alpha \rightarrow \infty$, the training batch follows a random uniform sampling of the original training dataset.

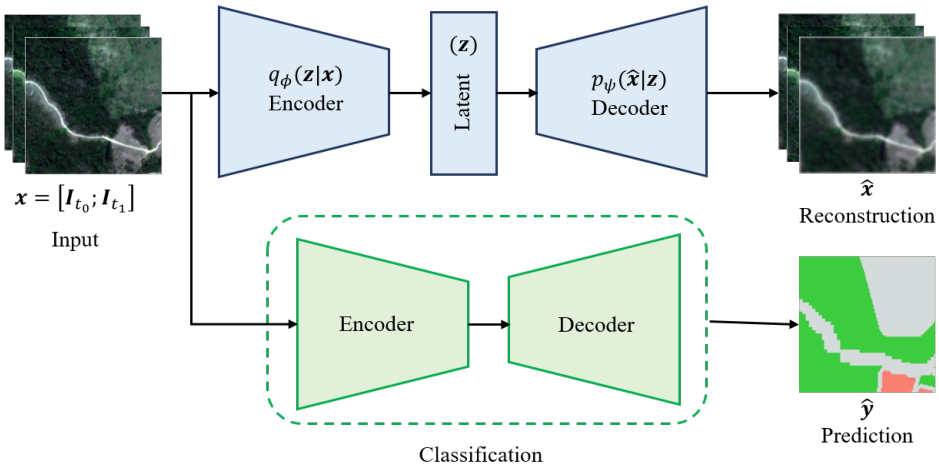

Figure 1. Fully Debiasing Variational Autoencoder (Fully DB-VAE) for semantic segmentation.

3. EXPERIMENTS

In this section, we report the results obtained with the Fully DB-VAE method in the deforestation detection application. We start by providing information about the dataset used in the experiments. Next, we describe the experimental setup, and finally we analyze the results in terms of classification metrics and visual interpretation. Furthermore, we compare the Fully DB-VAE with the conventional U-Net (Ronneberger et al., 2015), which was chosen as a baseline model.

3.1 Study Area

The study area comprises a Sentinel-2 scene, with a size of 17730×9200 pixels The region is located in Pará State, Brazil. That State recorded one of the highest deforestation rates in 2021, according to the PRODES monitoring system (Assis et al., 2019) from the Brazilian National Institute for Space Research (INPE). We used all Sentinel-2 bands with spatial resolutions of 10m and 20m. The 20m bands were resampled to 10m using nearest neighbor interpolation. The images were downloaded from the Google Earth Engine (GEE) platform (Gorelick et al., 2017). we used Level-1C images, which means they are orthorectified and contain top-of-atmosphere reflectance data. The change map was downloaded from the INPE web site, which is freely available at the PRODES database [1].

[1] Available at: http://terrabrasilis.dpi.inpe.br/map/deforestation

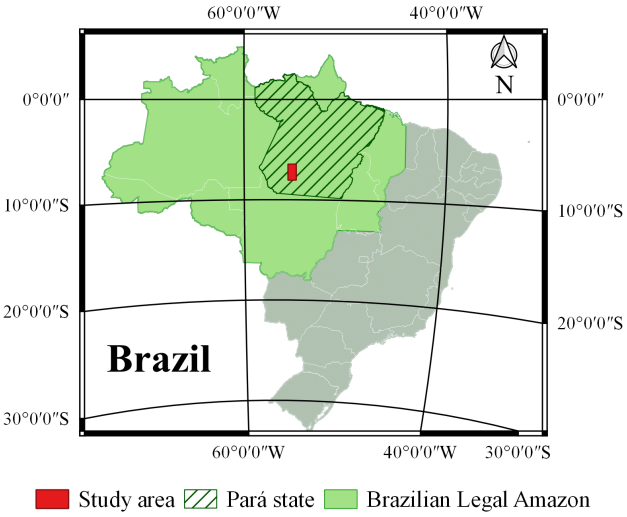

Figure 2. Geographical localization of the study area.

3.2 Experimental Setup

In all experiments, the input was a tensor resulting from the concatenation of a bitemporal image pair, acquired at dates t_0 and t_1, along the spectral dimension. Next, each band was normalized in the range of $[-1, 1]$. The study area was divided into 20 large tiles, following a distribution of $40\% - 10\% - 50\%$ for training-validation-test. The input to the encoder networks were patches of size 128×128 pixels, extracted using a sliding window procedure with a stride equal to 32.

Table 1 shows the acquisition dates of the Sentinel-2 data and the class distribution of the study area. We are interested in mapping the deforestation that occurred between 2020 and 2021. The dates of selected images are based on the PRODES project, which uses images during the dry season, with the minimum cloud cover. For date t_1, the image data is a mosaic composed of two Sentinel-2 scenes. The class deforestation represents samples mapped as forest at epoch t_0, and no forest at epoch t_1, the class no-deforestation corresponds to samples mapped as forest at epoch t_0 and t_1, and the class past-deforestation corresponds to all regions deforested before epoch t_0. It's worth mentioning that this application presents a

high level of class imbalance. Table 1 shows that less than 2% of the study area corresponds to the deforestation class.

Date t_0	July 15, 2020
Date t_1	July 25, 2021 August 4, 2021
Class deforestation (%)	1.86
Class no-deforestation (%)	56.40
Class past-deforestation (%)	41.74

Table 1. Image acquisition dates and class distribution of the study area.

Table 2 presents the Fully DB-VAE architectures, with information in the corresponding layers of the encoders, bottlenecks, and decoders. Furthermore, the following training parameters were defined: batch size equal to 16; Adam optimizer with learning rate equal to $2e^{-4}$, and β equal to 0.5. Also, the early stopping strategy was used to avoid over-fitting. The values of the loss function weights λ_s, λ_r, and λ_u were set to 1.0, 1.0, and $5e^{-4}$, respectively.

Following the PRODES methodology, during training and testing, we disregarded predictions within a two-pixel wide buffer in the internal and external edges of all polygons classified as deforestation in the reference change map. Similarly, pixels corresponding to the past-deforestation (before 2020) and polygons smaller than 625 pixels (6,25 ha) were ignored. The inference was carried out tile-wise, and each experiment was run five times.

3.3 Results and Discussion

In this section, we present the obtained results using the Fully DB-VAE. The classification performance was measured in terms of *Recall*, *Precision*, and *F1-score* for the deforestation class. Furthermore, for a visual interpretation, we present the deforestation probability maps.

Figure 3 shows the accuracy values obtained in each experiment. The first bar group represents the results for the baseline, which was a conventional U-Net, using a random sampling strategy for selecting the training samples. The other bar groups present the metrics obtained for the Fully DB-VAE with different values of α, starting from $\alpha = 1e{-}1$ to $\alpha = 1e{-}10$. According to the figure, one can notice that the Fully DB-VAE obtained the highest *Recall* scores. Also, the false deforestation predictions were lower than the baseline, but *Precision* was also lower. However, in all the cases, it exceeded 80%. In terms of *F1-score*, the best result was achieved for $\alpha = 1e{-}7$, yielding a score of 79.5, about 5% higher than the baseline.

As an additional experiment, we compared the performance of the Fully DB-VAE with the U-Net architecture using an arbitrary selection criterion in the training procedure: similar to (Ortega Adarme et al., 2020, Ortega et al., 2021), we only selected patches with at least 2% of pixels belonging to the class deforestation. Figure 4 summarizes the results of the U-net and the Fully DB-VAE with the two strategies for training sample selection. For the experiments with the Fully DB-VAE we defined $\alpha = 1e{-}7$, which presented the best performance in terms of the *F1-score* of previous results. We can observe the Fully DB-VAE still yields the best performance, although the results using the 2% selection criterion were better than with the random selection.

For a visual analysis, Figure 5 illustrates the deforestation probability maps of a snip in the test set. The maps represent the average prediction map. They were generated for all experiments using the U-Net as a baseline and the Fully DB-VAE with different values of α. In the first line, the RGB compositions of t_0 and t_1 are shown. Next, the class label map with classes deforestation, no-deforestation and past-deforestation is presented, as well as the outcome of the U-Net using the training samples from the random selection and from the 2% selection criteria. The second and third lines depict the outcomes of the Fully DB-VAE modifying the value of α. Red and blue colors represent the highest and lowest probability of belonging to the deforestation class. The black color symbolizes the past- deforestation class. Similar to the classification metrics, we note that the Fully DB-VAE with $\alpha = 7$ delivered the most accurate and reliable outputs, i.e., with probabilities close to one. In addition, we observed that the outcomes from the U-Net with random selection samples and the Fully DB-VAE with $\alpha = 1e^{-1}$ and $\alpha = 1e^{-2}$ are farther from the reference, delivering less reliable and less accurate definition of the polygons of the class deforestation.

4. CONCLUSIONS

This work presented an extension of the debiasing variational autoencoder (DB-VAE) architecture based on Fully Convolutional Neural Networks (FCN), so called Fully DB-VAE. The method was adapted to perform semantic segmentation in the context of detection of deforestation spots in a region of the Amazon Brazilian biome. In order to find the optimal value of the debiasing parameter α, an ablation study was carried out. The obtained results were compared with a conventional U-Net model using two strategies for training sample selection. The first one is a random approach, and the second one only patches with at least 2% of pixels belonging to the class deforestation were used. According to the experiments, the best performance in terms of $F1 - score$ was achieved with $\alpha = 1e^{-7}$. In that case, the Fully DB-VAE outperformed both baselines and delivered the most reliable values in the deforestation probability maps.

5. ACKNOWLEDGEMENTS

The authors would like to thank the German Academic Exchange Service (DAAD), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the financial support.

REFERENCES

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., Mushtaq, M. F., 2019. Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), 1560–1571.

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., Rus, D., 2019. Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 289–295.

Assis, F., Fernando, L., Ferreira, K. R., Vinhas, L., Maurano, L., Almeida, C., Carvalho, A., Rodrigues, J., Maciel, A.,

Architecture	Encoder	Bottleneck	Decoder	Output
Autoencoder	MP(C(3×3, 16)) MP(C(3×3, 32)) MP(C(3×3, 64)) MP(C(3×3, 128)) MP(C(3×3, 128))	Flatten() Dense(latentDim) Dense(latentDim*2)	Reshape(Dense(.)) US(C(3×3, 128)) US(C(3×3, 128)) US(C(3×3, 64)) US(C(3×3, 32)) US(C(3×3, 16))	Tanh (C(1×1, #IC))
Classifier	MP(C(3×3, 16)) MP(C(3×3, 32)) MP(C(3×3, 64)) MP(C(3×3, 128)) MP(C(3×3, 128))	-	US(C(3×3, 128) US(C(3×3, 128 US(C(3×3, 64) US(C(3×3, 32) US(C(3×3, 16)	Softmax (C(1×1, #CL))

Table 2. Network Architectures. Symbols: C (strided convolution), MP (max-pooling), US (bilinear up-sampling), IC (Input Channels), CL (Output classes). The values in parentheses refer to (kernel width x kernel height, number of filters).

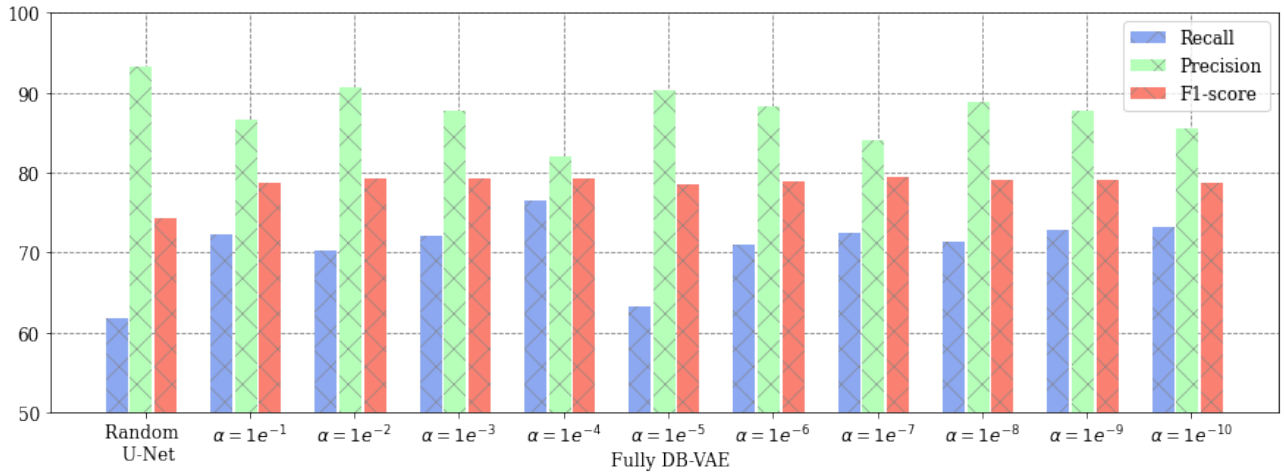

Figure 3. Classification metrics in [%], for the U-Net with random selection sampling and Fully DB-VAE with different values of α.

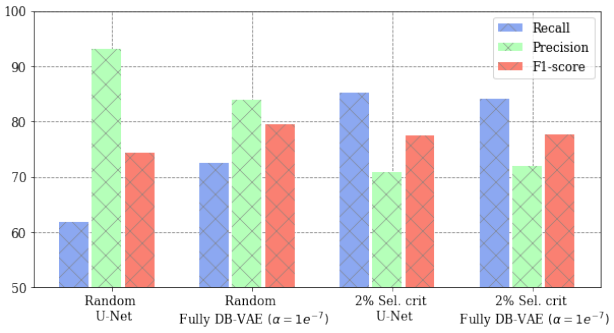

Figure 4. Classification metrics in [%], for the U-Net and Fully DB-VAE with $\alpha = 1e^{-7}$, using random samples and the 2% selection criteria.

Camargo, C., 2019. TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale Thematic Mapping. *ISPRS International Journal of Geo-Information*, 8(11), 513.

Bechavod, Y., Ligett, K., 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K. R., 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A., 2017. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Edwards, H., Storkey, A., 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., Venkatasubramanian, S., 2015. Certifying and removing disparate impact. *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. *2008 IEEE conference on computer vision and pattern recognition*, Ieee, 1–8.

Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., Wang, C., 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341*.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., Roth, D., 2019. A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the conference on fairness, accountability, and transparency*, 329–338.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202, 18–27.

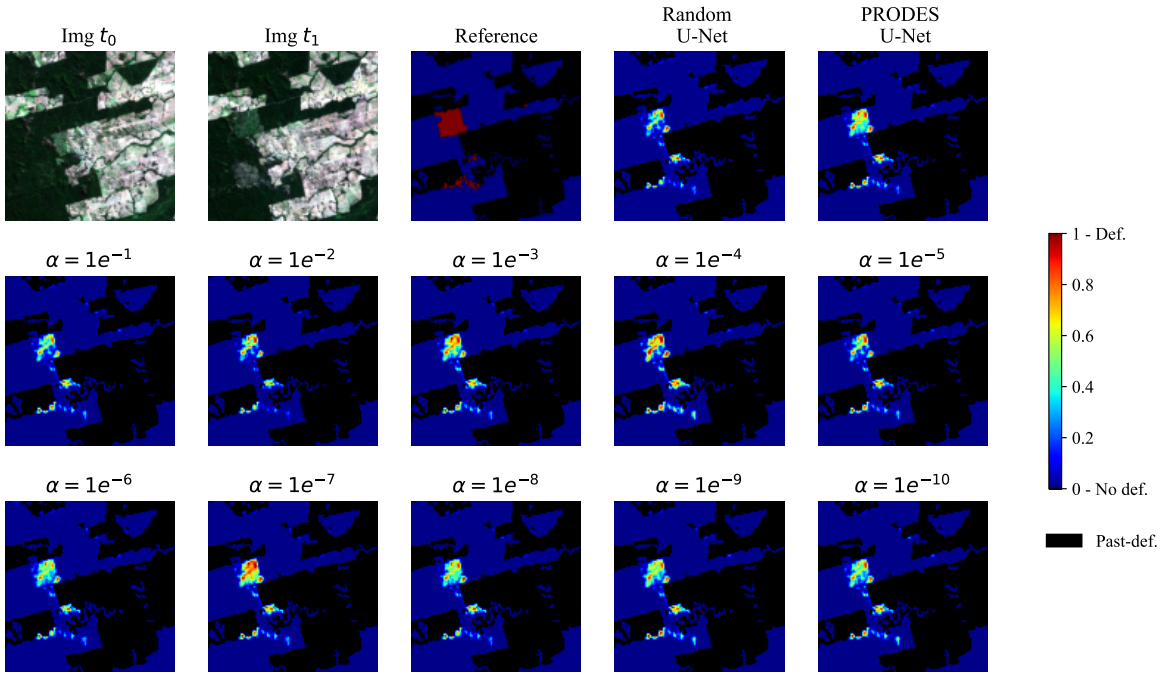

Figure 5. Visual example of a snip from the test set. The first line illustrates the RGB compositions of the snip in t_0 and t_1, the class label map, and the baselines, using the U-Net architecture with two strategies of training sample selection. The second and third lines presents the probability maps varying the values of α in the Fully DB-VAE. Red and Blue color represent the highest and lowest probability of belonging to the deforestation class. Black color symbolize the past deforestation.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Jiang, Z., Pan, T., Zhang, C., Yang, J., 2021. A new oversampling method based on the classification contribution degree. *Symmetry*, 13(2), 194.

Kamiran, F., Calders, T., 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1), 1–33.

Kamishima, T., Akaho, S., Asoh, H., Sakuma, J., 2012. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, Springer, 35–50.

Kingma, D. P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liu, S., Sun, S., Zhao, J., 2022. Fair Transfer Learning with Factor Variational Auto-Encoder. *Neural Processing Letters*, 1–13.

Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R., 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Luong, B. T., Ruggieri, S., Turini, F., 2011. k-nn as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510.

Madras, D., Creager, E., Pitassi, T., Zemel, R., 2018. Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, PMLR, 3384–3393.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.

More, A., 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Naseriparsa, M., Al-Shammari, A., Sheng, M., Zhang, Y., Zhou, R., 2020. RSMOTE: improving classification performance over imbalanced medical datasets. *Health information science and systems*, 8, 1–13.

Ortega Adarme, M., Queiroz Feitosa, R., Nigri Happ, P., Aparecido De Almeida, C., Rodrigues Gomes, A., 2020. Evaluation of deep learning techniques for deforestation detection in the Brazilian Amazon and cerrado biomes from remote sensing imagery. *Remote Sensing*, 12(6), 910.

Ortega, M. X., Feitosa, R. Q., Bermudez, J. D., Happ, P. N., De Almeida, C. A., 2021. Comparison of optical and sar data for deforestation mapping in the amazon rainforest with fully convolutional networks. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, 3769–3772.

Osoba, O. A., Welser IV, W., 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation.

Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 234–241.

Ruoss, A., Balunovic, M., Fischer, M., Vechev, M., 2020. Learning certified individually fair representations. *Advances in neural information processing systems*, 33, 7584–7596.

Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., Vempala, S., 2018. The price of fair pca: One extra dimension. *Advances in neural information processing systems*, 31.

Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A., 2020. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, 429–441.

Vardi, G., 2022. On the implicit bias in deep-learning algorithms. *arXiv preprint arXiv:2208.12591*.

Wang, X., Zhao, Y., Pourpanah, F., 2020. Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, 11, 747–750.

Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G., 2017. Controllable invariance through adversarial feature learning. *Advances in neural information processing systems*, 30.