Algal Bloom Forecasting in a Classification and Regression Setting

Implementing a UNet Architecture to evaluate the differences between both settings.

Rodrigo Alvarez Lucendo

Supervisor(s): Jan van Gemert, Attila Lengyel and Robert-Jan Bruintjes

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2023

Name of the student: Rodrigo Alvarez Lucendo
Final project course: CSE3000 Research Project
Thesis committee: Jan van Gemert, Attila Lengyel and Robert-Jan Bruintjes, Koen Langendoen

Abstract

Forecasting algal blooms using remote sensing data is less labour-intensive and has better coverage in time and space than direct water sampling. The paper implements a deep learning technique, the UNet Architecture, to predict the chlorophyll concentration, which is a good indicator for algal bloom in the Rio Negro water reservoirs of Uruguay. The research question focuses on the differences between classification and regression in algal bloom forecasting. The experiments show that the regression implementation achieves better accuracy and lower mean squared error than the classification implementation that uses cross-entropy loss and four pre-fixed bins. Different loss functions that account for the class imbalance in the data do not improve the model's performance. Finally, a quantile-based binning strategy that considers the data's underlying distribution achieves the highest accuracy in both settings.

1 Introduction

Over the last decades, there has been a dramatic escalation in the number of harmful algal blooms (HABs, commonly called "red tides") across the globe. HABs are caused by toxic or harmful algae blooms that can cause severe environmental and human health problems and economic impacts [1]. Forecasting algal bloom may be helpful in limiting the harmful effects of HABs.

Measuring algae concentrations traditionally relies on direct water sampling, a labour-intensive method that is limited spatially and temporally. Remote-sensing-based detection solves these two problems, but it often relies on estimated data such as chlorophyll (Chl-a) that may be unreliable estimates and not direct measurements [2]. The paper will use remote-sensing data to forecast the estimated chlorophyll concentration values (in μg/l). The estimation is based on a local algorithm developed by the Ministry of the Environment for the three reservoirs of the Río Negro in Uruguay: *Baygorria*, *Bonete* and *Palmar*.

The objective is to predict a continuous variable, the chlorophyll concentration, given remote-sensing data. The output is intuitive, but it does not explain the accuracy of the predictions. It is possible to frame the original regression problem as a classification problem to account for uncertainty by estimating the probability of a value belonging to a bin or a range of values.

Forecasting algal blooms is a challenging task due to algal's non-linear and non-stationary nature, especially for classical models such as linear regression. Machine learning approaches are known to work well with complex real-world data and have been applied to HAB predictions. Recent deep learning techniques such as Long-Short-Term-Memory (LSTM) networks can discover temporal patterns in the data and thus improve the prediction [3]. The research is not concerned about the accuracy of the predictions, so a simpler network that only takes into account the spatial information, the UNet Architecture, has been implemented [4]. Specifically, two different adaptations of UNet have been implemented. One predicts a discrete label corresponding to a range of values, and the other outputs a continuous value representing the chlorophyll concentration.

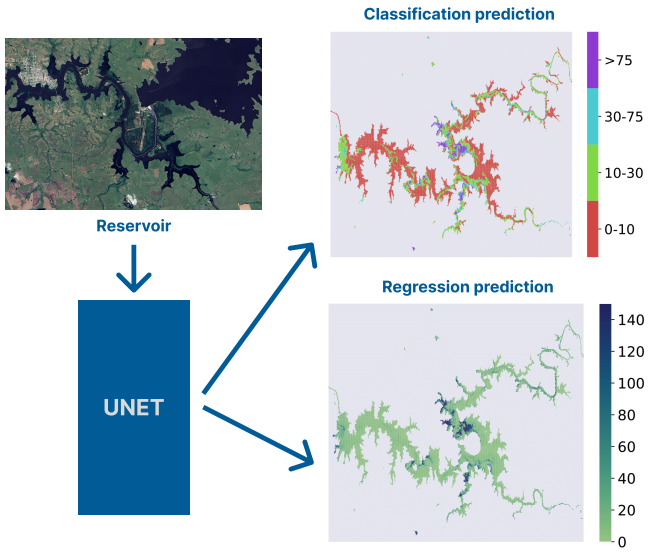

Figure 1. *The paper explores the difference between a regression and classification implementation of the UNet Architecture in the context of algal bloom forecasting.*

The paper will answer the following research question: *What are the differences between a classification and regression model for forecasting the chlorophyll-a concentration of a water reservoir?* The research question can be divided into three sub-questions:

1. What are the differences between a classification and regression implementation of the UNet Architecture?
2. How can class imbalance be mitigated using different loss functions?
3. What influence does the binning strategy have?

The research provides three main contributions. The regression implementation of UNet outperforms the simple classification implementation in terms of accuracy and mean squared error, but it misses vital information about the uncertainty of the predictions. Different loss functions that account for the class imbalance have little effect on the overall accuracy of the predictions. An appropriate binning strategy improves the performance in the classification task and gives the best accuracy and mean squared error.

2 Related Work

Machine learning methods have already been developed to forecast the occurrence of algal blooms using remote sensing data. In a study, the Gradient Boosted Descent Tree (GBDT)

algorithm was the most effective approach to predict such phenomenon [5]. The data used in HAB forecasts contains relevant spatiotemporal information. Long Short-Term Memory (LSTM) networks are typically used in time series data where the order of the observations is essential. Convolutional Neural Network (CNN) models are used to work with images. An approach to work with sequential images is the ConvLSTM [6]. Although, these methods provide excellent performance in predicting HABs, the research is concerned with the differences between regression and classification, so a more straightforward method, the UNet Architecture, has been implemented.

One way of handling the class imbalance problem is to use an appropriate loss function. A simple heuristic is to set class weights inversely proportional to the class frequency. However, as the imbalance increases, loss functions based on overlap measures, such as dice score, have been proven to be more robust than weighted cross entropy loss [7]. Dice loss penalizes the pixel-wise mismatch between prediction and ground truth and is widely used in image segmentation tasks to solve the class imbalance problem [8]. Other loss functions, like focal loss, force the network to focus on hard samples that are not easily discriminated from others and are often misclassified. Focal loss can enable Convolutional Neural Network (CNN) models to be less biased towards the majority class and achieve higher F_1-scores than models augmented with normal cross-entropy loss [9]. Another candidate loss function, Dynamically Weighted Balanced (DWB) loss, adapts its scale according to the reliability of confidence estimates [10]. The paper uses focal, dice and weighted cross entropy loss to account for the class imbalance of the chlorophyll concentration and leaves DWB loss as future work.

3 Method

This section describes the methods used to answer the research question and the three sub-questions.

3.1 Classification vs Regression

In the classification task, the continuous chlorophyll-a concentration values are assigned to a label according to the following ranges: label 1 [0, 10), label 2 [10, 30), label 3 [30, 75) and label 4 for values above 75 mg/mL. The data is skewed towards low chlorophyll-a concentration values since most of the points are set to labels 1 and 2 while only a few get assigned to label 4. Applying a Softmax Activation Function to the model's output gives the probabilities of each label for a given input. The label with the highest probability becomes the predicted label.

The label goes through the same transformations as the input in the regression implementation. The Mean Squared Error (MSE) is used as the loss function. Since the least popular labels have high chlorophyll-a concentration values, the loss function will penalise these more, correcting partly for the class imbalance. No other technique has been used to fight the class imbalance in the regression setting. After the standardisation has been reverted, the model's output becomes the predicted continuous value.

3.2 Loss Functions

The data is distributed unevenly across the different classes. Training on such an imbalanced dataset results in a model that performs well in majority classes but cannot predict the minority classes. Cross Entropy Loss treats each class equally and is expected to perform poorly on such datasets. Different loss functions that account for class imbalance have been implemented to answer the second research question.

Focal loss is a dynamically scaled cross-entropy loss.

$$\text{FocalLoss} = -\sum (1 - p_i)^{\gamma} \log(p_i), \tag{1}$$

where the term $-\log(p_i)$ is normal cross entropy loss and the term $(1 - p_i)^{\gamma}$ is the regulating factor that focal loss introduces [11]. The idea is to reduce the loss more in well-classified examples than in less confident misclassified samples. If the model is confident, a small portion of the cross entropy loss is taken, and when the model is less confident, it is penalised more by taking a larger portion of the cross entropy loss. The focusing parameter, γ, is set to five in the experiments.

Class-balanced loss introduces a weighing factor σ to account for the class imbalance. Precisely, a class-balanced loss function based on the effective number of samples [12] has been used in the experiments and has the following weighting factor.

$$\sigma_i = \frac{1 - \beta}{1 - \beta^{n_i}}, \tag{2}$$

where n_i refers to the number of samples in the ground truth label i and β is a parameter that can be tuned but is set to 0.99.

Dice loss is a measure of similarity between two samples, where zero means the labels are predicted perfectly, and one means none of the predictions matches the labels. The idea is to maximise the overlap or correctly classified labels while minimising the union of the ground truth and the prediction. The definition of dice loss is

$$\text{DiceLoss} = 1 - \frac{2 \sum \hat{y} y + \epsilon}{\sum \hat{y} + y + \epsilon}, \tag{3}$$

where ϵ is a small number to avoid division by zero, $\hat{y}$ are the predictions and y are the labels.

Compound Loss is obtained by summing over different types of loss functions.

3.3 Binning Strategies

Two different binning strategies have been implemented to convert the original regression problem into a classification one. Figure 2 visualises the class ranges used by the different binning strategies against the chlorophyll concentration distribution.

In *fixed-width binning*, values are assigned to bins according to some predefined range of values based on some domain

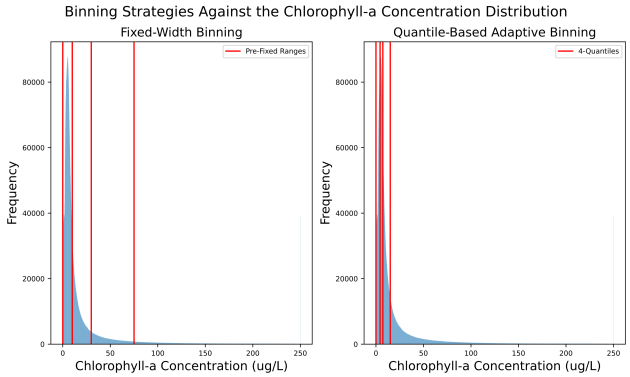

Figure 2. *Visualisation of the estimated chlorophyll-a concentration distribution and the class ranges used in different binning strategies. Fixed-width binning results in irregular bins, while quantile-based adaptive binning results in equal-sized bins by considering the data distribution.*

knowledge. A drawback of this approach is that it can lead to irregular bins that contain different numbers of points. The Uruguay government considers the chlorophyll-a concentration high when it is above 80 μg/L and is interested in the following ranges of values: [0, 10, 30, 75, 150].

If a specific range of values is not necessary, *adaptive binning* is a safer strategy because it is based on the underlying distribution of the data. Specifically, the paper implements quantile-based binning, where *q-Quantiles* can be used to partition the data into q equal partitions [13]. In the experiments, 4-Quantiles binning is used, which results in the following bin ranges: [0.0, 4.34, 7.24, 15.04, 150.0]. These ranges are fine-grained in the lower chlorophyll-a concentration values since the data is skewed towards lower values.

4 Dataset

The experiments were run on data from the *Palmar* reservoir. A data loader has been used to load the data into the model. Additionally, before training the model, some pre-processing steps were applied to the data.

4.1 Data loader

The data loader allows specifying a window size and a prediction horizon. The window size controls the number of past observations taken into account when training the model. A window size larger than one is more sensible in an architecture designed to learn temporal patterns, so it is set to one in the experiments. The prediction horizon represents how far into the future a prediction is made in terms of days. A prediction horizon of one is used in the experiments. The training loader samples 200 random crops of the reservoir every epoch in batches of size 4. The validation loader contains the same 37 samples, measurements that take place after December 31^{st} of 2021, across epochs.

4.2 Data Processing

The research discards the meteorological data and only uses biological, i.e., chlorophyll, turbidity and cdom, and water

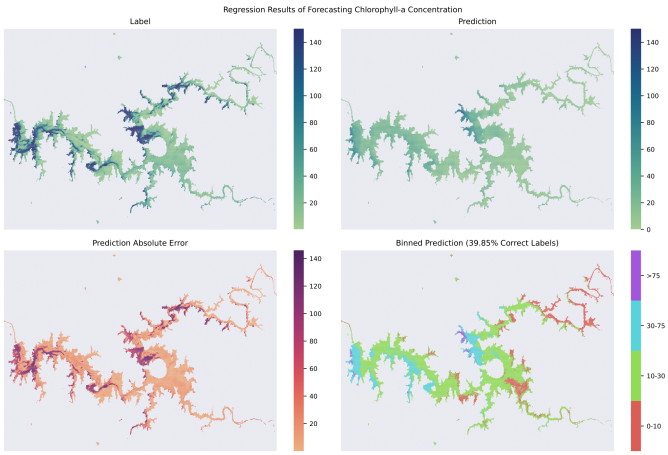

Figure 3. *Forecast of the chlorophyll concentration for a specific sample of the validation set in the regression setting. The label and the prediction are continuous values between 0 and 150, representing the chlorophyll concentration in μg/l for a specific reservoir location. The grey area represents missing labels due to cloud coverage. The absolute prediction error is also plotted, and the binned regression output is for a more direct comparison with the regression setting. The binned prediction plot contains the accuracy of the sample in the title.*

temperature features for efficiency. The missing values in the input are replaced by zero. The data contains significant outliers that may hurt the performance of the model. The values are clipped to a maximum value of 150 to limit the effect of outliers. Furthermore, a Yeo–Johnson transformation has been applied to the biological data to make the biological data more normal distribution-like. Finally, the data is standardised to have mean zero and unit variance.

5 Experiments

The average mse loss and accuracy of all the training and validation steps are logged per epoch. The loss is computed without considering predictions with missing labels, so these do not contribute to the gradient. Similarly, the accuracy ignores missing labels. The training uses the Adam optimizer and a learning rate of 10^{-4}. Gradients exceeding a value of 1.0 are clipped to avoid the exploding gradient problem. The network can overfit in one batch and predict all the labels. These checks verify that the network has been implemented and adapted correctly to the data. Multiple runs are needed to reduce the possible effects of the random initialization of weights and biases at the beginning of the experiment. Therefore, each experiment has been run five times.

Difference between regression and classification

The experiment answers the question: *what are the differences between a classification and regression implementation of the UNet Architecture?* Figure 3 shows the regression results for a specific sample of the validation set and Figure 4 shows the classification results for that same sample. One

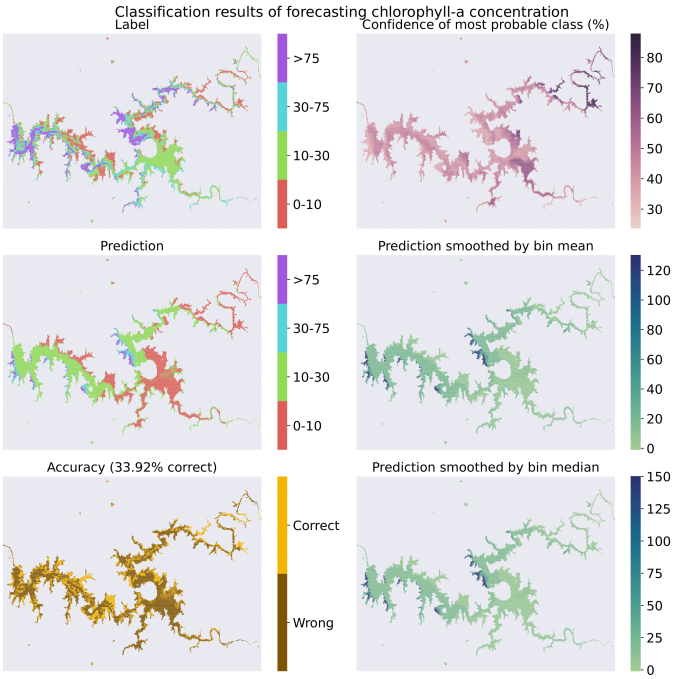

Figure 4. *Forecast of the chlorophyll concentration for a specific sample of the validation set in the classification setting. The label and the prediction are discrete labels that correspond to ranges of chlorophyll concentration values in µg/l. The figure shows the reservoir locations where the prediction matches the label and the confidence of the prediction to account for model uncertainty. The bins are smoothed to continuous values by taking the mean and median of each bin for a direct comparison with the regression output.*

big difference is that confidence in the prediction can only be measured and plotted in the classification setting. Binning the regression output makes it possible to obtain accuracy scores in the regression setting. Smoothing the classification output by the mean and median of each bin allows the computation of the mean squared error in the classification setting. Table 1 summarizes the average and standard deviation of the accuracy and mse in the regression and classification setting. The regression implementation has higher accuracy and lower error on average than the classification implementation that uses normal cross-entropy loss and four pre-fixed bins.

METHOD	ACCURACY (%)	MSE (μG/L)
regression	43.2 ± 0.7	1953 ± 8
classification	41.5 ± 0.9	2154 ± 35

Table 1. *Average and standard deviation of samples' accuracy and mean squared error in the validation set in the regression and classification task over five runs. The regression implementation has better accuracy and lower mean squared error than the classification implementation.*

LOSS	ACCURACY (%)	MSE (μG/L)
cross-entropy	41.5 ± 0.9	2154 ± 35
balanced cross-entropy	41.6 ± 0.1	2114 ± 37
focal	40.8 ± 0.4	2166 ± 44
dice	36.9 ± 0.3	2436 ± 6

Table 2. *Average and standard deviation of the accuracy and mean squared error of the validation set for different loss functions. There is no significant difference in the overall accuracy using different loss functions.*

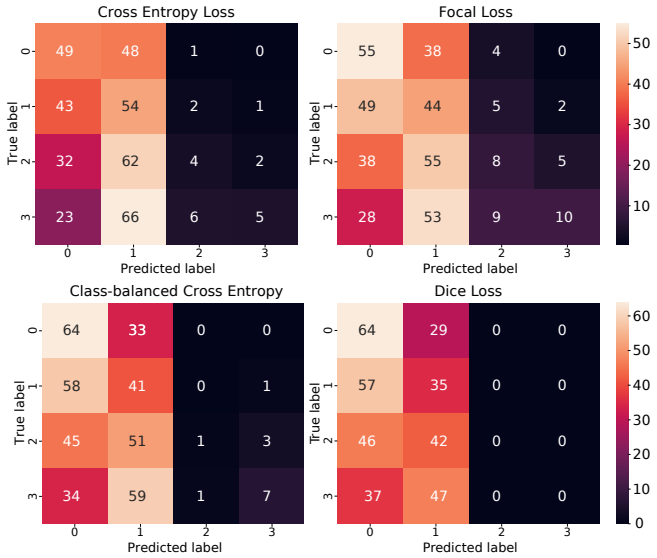

Figure 5. *Normalized confusion matrices of the classification predictions using different loss functions. The accuracy does not improve for the minority classes by using different loss functions.*

Choosing an appropriate loss function

The following experiment answers the question: *how can class imbalance be mitigated using different loss functions?* Table 2 shows that changing the loss function to account for class imbalance does not significantly impact the overall accuracy of the predictions in the validation set. Additionally, Figure 5 shows that the accuracy does not improve for the minority classes except for the model augmented with focal loss, which shows a slight improvement in the prediction of the minority classes.

Binning strategy analysis

This experiment answers the question: *what influence does the binning strategy have?* Figure 6 shows that fixed-width binning can predict roughly half of the labels in the majority classes but almost none in the minority classes, while quantile-based binning predicts all the labels more accurately, specifically label 3. On average, Table 3 shows that quantile-based binning results in higher accuracy but a bigger mean squared error. Most adaptive binning misclassifications are assigned to label three instead of label 0, as happens on fixed-width binning. Since label 3 has a range of higher chloro-

BINNING	ACCURACY (%)	MSE (μG/L)
fixed-width	41.5 ± 0.9	2154 ± 35
adaptive	52.6 ± 0.6	2753 ± 10

Table 3. *Average and standard deviation of samples' accuracy and mean squared error in the validation set using different binning strategies. The adaptive-binning strategy results in higher accuracy but a bigger mean squared error.*

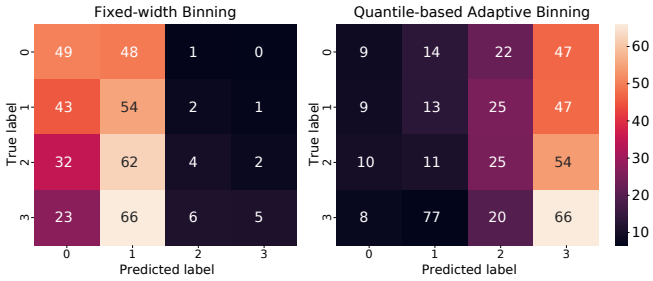

Figure 6. *Normalized confusion matrices for different binning strategies. Fixed-width binning predicts the majority classes well, labels 0 and 1, but it performs poorly on the minority classes, labels 2 and 3. Quantile-based binning reports better accuracy in all labels.*

phyll concentration values, the error is expected to be bigger in adaptive binning, even though the accuracy is better than in fixed-width binning

6 Responsible Research

The experiment has been made more reproducible by using a popular PyTorch implementation of the UNet architecture that is widely available on GitHub [14]. Additionally, data processing steps and parameters, such as the learning rate or the batch size, have been documented in the paper. However, the data used to train the model is not publicly available, so the experiment's reproducibility is limited to those with access to the data.

Algal blooms are a serious phenomenon that can have severe health, environmental and economic effects. For those reasons, correctly forecasting algal blooms is important. On the contrary, providing the wrong predictions may worsen the effects of harmful algal blooms. The research has provided a classification implementation of UNet that gives information about the confidence of the predictions, which can be used by the authorities to make more educated decisions in the possible presence of algal bloom.

7 Discussion

Two different variants of the UNet Architecture have been implemented to predict the chlorophyll concentration of a reservoir using remote sensing data in a regression and classification setting. The regression task uses mean squared error, achieves higher accuracy and has a lower error than the classification task, which uses cross-entropy loss and four prefixed bins. If a concrete set of ranges is not required, an adaptive binning strategy based on the quantiles of the chlorophyll concentration distribution yields the highest accuracy score in both settings. The classification implementation can provide information about the uncertainty of the predictions. The class imbalance in the data was tackled using different loss functions such as focal, dice and class-balanced loss. None of the loss functions showed a significant improvement in the overall accuracy of the predictions compared to cross-entropy loss. However, focal loss did achieve a slightly higher accuracy on the minority classes.

The experiments have been limited to manual tuning, and the UNet model is unsuitable for accurate algal bloom forecasts. The same experiments can be run in future work with a more adept model and hyperparameter tuning to see if the same conclusions are reached.

Acknowledgements

We thank the government of Uruguay for providing access to the remote sensing data of the Rio Negro reservoirs.

References

[1] Donald M. Anderson. Approaches to monitoring, control and management of harmful algal blooms (habs). *Ocean Coastal Management*, 52(7):342–347, 2009. Safer Coasts, Living with Risks: Selected Papers from the East Asian Seas Congress 2006, Haikou, Hainan, China.

[2] Paul R. Hill, Anurag Kumar, Marouane Temimi, and David R. Bull. Habnet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3229–3239, 2020.

[3] Muyuan Liu, Junyu He, Yuzhou Huang, Tao Tang, Jing Hu, and Xi Xiao. Algal bloom forecasting with time-frequency analysis: A hybrid deep learning approach. *Water Research*, 219:118591, 2022.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[5] Peixuan Yu, Rui Gao, Dezhen Zhang, and Zhi-Ping Liu. Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, 123:107334, 2021.

[6] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, 2015.

[7] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi,

Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, Cham, 2017. Springer International Publishing.

[8] Yue Zhang, Shijie Liu, Chunlai Li, and Jianyu Wang. Rethinking the dice loss for deep learning lesion segmentation in medical images - journal of shanghai jiaotong university (science), Jan 2021.

[9] Kitsuchart Pasupa, Supawit Vatathanavaro, and Suchat Tungjitnob. Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification - journal of ambient intelligence and humanized computing, Feb 2020.

[10] K. Ruwani M. Fernando and Chris P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2022.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

[13] Dipanjan (DJ) Sarkar. Continuous numeric data, Mar 2019.

[14] Milesial. Milesial/pytorch-unet: Pytorch implementation of the u-net for image semantic segmentation with high quality images.