Journal of Dentistry 148 (2024) 105063

Contents lists available at ScienceDirect

Journal of Dentistry

journal homepage: www.elsevier.com/locate/jdent

Conquering class imbalances in deep learning-based segmentation of dental radiographs with different loss functions

Martha Büttner [a,c], Lisa Schneider [a,c], Aleksander Krasowski [a], Vinay Pitchika [b], Joachim Krois [c], Hendrik Meyer-Lueckel [d], Falk Schwendicke [b,c,*]

[a] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité – Universitätsmedizin Berlin, Germany
[b] Clinic for Conservative Dentistry and Periodontology, Ludwig-Maximilians-University Munich, Germany
[c] ITU/WHO Focus Group AI4Health
[d] Department of Restorative, Preventive and Pediatric Dentistry, zmk Bern, University of Bern, Switzerland

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Computer Vision

ABSTRACT

Objective: The imbalanced nature of real-world datasets is an ongoing challenge in the field of machine and deep learning. In medicine and in dentistry, most data samples represent patients not affected by pathologies, and on imagery, pathologic image areas are often smaller than healthy ones. Selecting suitable loss functions during deep learning is essential and may help to overcome the resulting imbalance. We assessed six different loss functions for one exemplary task, tooth structure segmentation on bitewing radiographs, for their performance.
Methods: Six different loss functions (Focal Loss, Dice Loss, Tversky Loss and hybrid losses of Cross-Entropy and Dice Loss, Focal and Dice Loss, Focal and Generalized Dice Loss) were compared on a tooth structure segmentation task of 1,625 bitewing radiographs. Training was performed using three different model architectures (U-Net, Linknet, DeepLavbV3+) over a 5-fold cross-validation. Tooth structures consisted of the classes (occurrence in% of samples/captures areas measured on pixel level) enamel (100 %/25 %), dentin (100 %/50 %), root canal (100 %/10 %), filling (81 %/8 %) and crown (28 %/5 %).
Results: Hybrid loss functions significantly outperformed standalone ones and provided robust results over the different architectures for the classes enamel, dentin, root canal and filling. Specifically, the Dice Focal loss reached high performance to conquer both image level and pixel level class imbalance, respectively.
Clinical Significance: In dental use cases it is often important to predict minority classes such as pathologies accurately. Using specific loss function may be an effective strategy to overcome data imbalance when training deep learning models.

1. Introduction

Image diagnostics has emerged as one of the most prevalent research fields for Machine Learning and, specifically Deep Learning (DL), in dentistry [1–4]. Nevertheless, challenges of DL-based systems in the field of dentistry remain [5]. Among them is the often-imbalanced nature of real-world datasets, which is likely to degrade the performance of DL models [6]. Imbalance in a dataset may occur in several ways, depending on the task at hand; for example, images with certain pathologies may be much less frequent than images of healthy conditions (we here refer to this as image-level class imbalance). In segmentation tasks, where a DL model assigns pixels to specific classes (like being healthy or affected by a pathology), imbalance may additionally occur pixel level, with pathologies capturing smaller areas than healthy area (or, generally, the background class).

As any DL model is gaining its knowledge from the dataset, it is likely to learn better how to distinguish the majority class than the minority class, as it is confronted with significantly more examples from the former than the latter [6]. If no appropriate actions are taken by DL researchers, the model may run the risk of neglecting the minority class, treating it as noise, and instead focusing on maximizing its performance on the majority class to optimize its objective function. Notably, from a clinical perspective it is often specifically important to predict the minority class (e.g. a certain pathology) correctly. There is great need to address dataset imbalances in DL in dentistry.

Different methods aim to achieve this, often grouped into data and

* Corresponding author at: Clinic for Conservative Dentistry and Periodontology, Ludwig-Maximilians-University Munich, Goethestr. 70, Munich 80336, Germany.
E-mail address: falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

https://doi.org/10.1016/j.jdent.2024.105063
Received 7 March 2024; Received in revised form 30 April 2024; Accepted 9 May 2024
Available online 10 May 2024

algorithm level approaches [7]. Data-level approaches attempt to achieve class balance by modifying the training dataset, often through resampling (over- or undersampling of certain images) [8]. Another data-level method evolves around the generation of synthetic data for minority classes, for example by using Generative Adversarial Networks (GANs) [9] and Stable Diffusion Models [10].

Algorithm-level methods use another approach and may involve employing different loss functions. The loss function quantifies the difference between the predictions of a DL model and the actual ground truth. DL aims to minimize the loss and therefore the differences between the prediction and the ground truth. For this, the outcome of the loss functions iteratively guides the learning process of the model. Algorithm-level methods may emphasize the importance of the minority classes in two different approaches. First, one may incorporate different weightings for the classes in the loss function. Providing larger weightings to minority classes in the loss function assigns a higher weight on the errors made on samples of the minority class. This helps to steer the learning process towards the minority class. Secondly, one may adjust the loss function itself to overcome pixel-level class imbalance. Different loss functions were proposed with the goal of handling such imbalances. Dice Loss [11] is a popular loss function that captures the overlap of the predicted pixel area and the ground truth area and is particularly useful to conquer the imbalance between the background class and that of interest (e.g., pathology). By default, the Dice Loss does not address the imbalance between minority and majority classes, that is, difficult classes due to limited number of image examples in the data set. This ability was specifically aimed for when the Focal Loss [12] was proposed, which guides the model towards improving on those example it currently predicts wrong rather than those it can predict already with high confidence. The Focal Loss is based on the Cross-Entropy loss [13], which captures the difference between the predicted probability distribution and the ground truth values. Another loss function aiming to solve data imbalance challenges is the Tversky loss [14]. It is also built on the Sørensen-Dice Coefficient, but adds two parameters, α and β, which allow to penalize false negative pixels or false positive pixels, respectively.

As some of the loss functions address different types of imbalance, it is also frequent practice to employ hybrid loss function based on the combination of two loss functions such as the Focal Dice Loss [15], Cross-Entropy Dice Loss or the Generalized Dice Focal loss. The latter one is a combination of the Focal and Generalized Dice Loss [16]. The Generalized Dice Loss extends the Dice loss by adding a weight to each class, which is typically inversely proportional to the squared volume of the class in the ground truth. This ensures that all classes contribute equally to the loss, independent of their size. A graphical display and mathematical summary of the loss functions can be found in the appendix.

Research around data imbalance in the field of dentistry is highly limited. We here aimed to benchmark different loss functions regarding their ability to handle class imbalance on an exemplary task, segmenting tooth structures, namely enamel, dentin, root canal, filling, crown, on bitewing radiographs. We hypothesized that hybrid loss functions yield better performance than single loss functions. We additionally assess the impact of different loss functions when combined with different model architectures.

2. Materials and methods

2.1. Study design

The present study involved several experiments. First, it was evaluated whether certain loss functions reached universally better results on the underlying imbalanced segmentation task than others. Second, it was assessed whether hybrid loss functions such as the Dice Focal loss performed better than standalone loss functions like the Dice loss. Finally, it was analyzed whether hybrid loss functions provided more robust results over different architectures than standalone loss functions. Fig. 1 gives an overview of the study design.

Six loss functions, namely Focal Loss, Dice Loss, Tversky Loss as well as the hybrids Cross-Entropy Dice Loss, Dice Focal Loss and Generalized Dice Focal Loss were employed and compared regarding their ability to overcome class imbalance in a tooth structure segmentation task of bitewing radiographs. The analysis was based on 90 experiments conducted with three different model architectures U-Net [17] (with Densenet121 [18] backbone), Linknet [19] (with ResNet152 [20] backbone) and DeepLabV3+ [21] (with ResNet152 [20] backbone) in a 5-fold cross-validation. Train, validation and test datasets for each fold consisted of proportions of 60 % (3 folds), 20 % (1 fold), and 20 % (1 fold), respectively. Hyperparameters were automatically tuned for each configuration of architecture and loss.

2.2. Dataset

A dataset of 1625 dental radiographic bitewings with a maximum of 8–9 teeth per image, which were collected during routine care at XXX between 2019 and 2020, were utilized in this study under ethical approval (EA4/080/18). The descriptive statistics of the dataset included a mean (SD, min, max) age of 35.6 (15.5, 11, 83) years and a gender ratio of 52 % to 48 % of males and females, respectively. The samples originated from radiographic machines of Dentsply Sirona (51.5 %) and Dürr Dental (47.9 %). For 0.6 % of the images, there was no information about machinery available.

For the annotation of the tooth structures in a pixel-wise manner, one dental expert performed the annotation and a second dental expert reviewed it regarding its validity and correctness. Annotations were performed in a standardized custom-built annotation tool that has been used in several previous studies [22,23]. All examiners were trained and calibrated on how to conduct the segmentations. The prevalence of the classes in the dataset, which reflected the image-level imbalance were enamel (100 %), dentin (100 %), root canal (100 %), filling (81 %) and crown (28 %). The pixel-level imbalance was quantified through the amount of foreground pixels assigned to each class over the whole dataset were enamel (25 %), dentin (50 %), root canal (10 %), filling (8 %) and crowns (5 %). The resolution of the samples was downsampled to 224×224.

2.3. Hyperparameter tuning

To provide a fair comparison of the different loss functions for the different model architectures, an extensive hyperparameter search was conducted. Since hyperparameters have a strong impact on the model performance, and to base the comparison only on the best performing model, the best training settings for each loss function with every architecture were identified. Hyperparameters were the optimizer with or without regularization, learning rate and batch size. The optimizer choices included Adam and SGD and the batch size was chosen from 4,8,16 and 32. The learning rate options were 0.001, 0.005, 0.01, 0.05 and 0.08.

Further, we provided different options for loss specific attributes e.g., the weighting of Focal and Dice Loss in the Dice Focal loss. Additional weights for balancing of classes were provided: First, equal weighting of all classes, second; weightings based on the share of the dataset ([0.048, 0.067, 0.049, 0.154, 0.680]); third, two sets of weights that were manually created ([0.15,0.15,0.15,0.2,0.35], [0.1,0.1,0.1,0.25,0.45]). The full overview of the loss specific attributes is listed in the Appendix.

We randomly sampled 100 configurations from the listed options for each combination of loss and model architecture and evaluated the results based on the validation set. We used the Asynchronous Successive Halving Algorithm (ASHA) algorithm, which allows performing massively parallel hyperparameter optimization with early stopping to avoid unnecessary computational efforts. Hyperparameter tuning was solely performed on one of the cross-validation splits to reduce

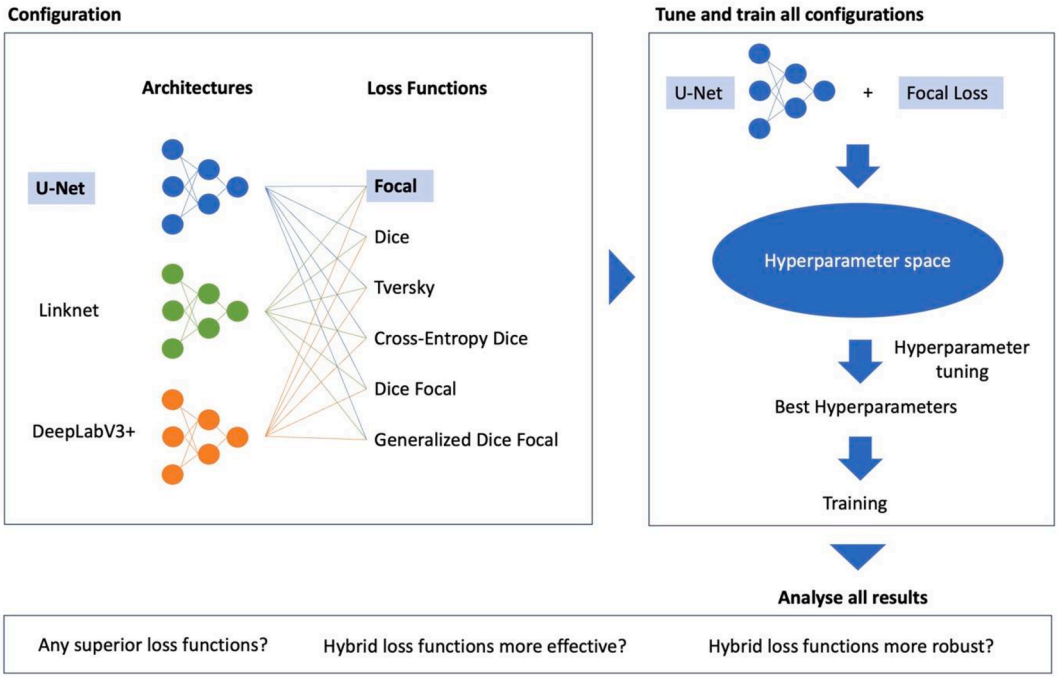

Fig. 1. Overview of the study design. All three model architectures were combined with all six loss functions in independent experiments. Through automatic hyperparameter tuning the best setting of hyperparameters were selected before training was conducted. Finally, all results were compared and hypotheses were assessed with statistical testing.

computational efforts. Tuning was performed with Ray 2.6.1 on four NVIDIA A100 40 G GPUs.

2.4. Training

The model architectures U-Net [17] (with Densenet121 [18] backbone), Linknet [19] (with ResNet152 [20] backbone) and DeepLavbV3+ [21] (with ResNet152 [20] backbone) were trained to solve the tooth structure segmentation task by utilizing the Dice Loss, Focal Loss, Tversky Loss, as well as combinations of Cross-Entropy and Dice Loss, Focal and Dice loss, Focal and Generalized Dice Loss, as laid out. Training was performed with the respectively tuned hyperparameters to prevent biases within the comparison. Models were utilized from Segmentation Models Pytorch [24]. Training was implemented with Pytorch 2.0 and MONAI 1.2 and was performed on four NVIDIA A100 40 G GPUs.

2.5. Performance metrics and statistical analysis

Model performance was primarily quantified by the harmonic mean of recall (specificity) and precision (positive predictive value (PPV)), also known as F1-score. The computation was based on the sum of sum of true positives, false positives, and false negatives over all channels of segmentation masks to reach unbiased results [25]. Secondary metrics were precision, sensitivity and specificity. Due to the non-normal distribution of the results, statistical analysis was performed with non-parametric tests, which were applied to different groups of results, depending on the focus of the analysis. To assess whether certain loss functions generally perform better than others, a Kruskal-Wallis test [26] was employed to test for significant differences between the results reached with the different loss functions. In case of significant differences, a post-hoc Dunn's test [27] was applied to find the ranking of the performances reached. P-values were adjusted using the Benjamini-Hochberg method [28] to account for the multiple comparisons.

The hypothesis that hybrid loss functions outperform standalone loss functions was tested with the non-parametric Mann-Whitney-U-Test [29]. Finally, it was assessed whether hybrid loss functions provide more robust results than standalone losses by considering the standard deviation reached by different losses across the different architectures. The collection of standard deviations of hybrids was tested against the standard deviations of the standalones by the Mann-Whitney-U-Test. Statistical testing was implemented with statsmodels 0.14, scikit-posthocs 0.7 and scipy 1.11.

3. Results

A detailed report of the tuned hyperparameters is provided in Appendix Table S1. Training was performed with the selected hyperparameters for all three model architectures with the six different loss functions in a 5-fold cross-validation and results were summarized as mean F1-scores for each class as represented Fig. 2.

In alignment with the individual comparison of loss functions, it was observed that hybrid loss functions universally outperformed single loss functions with statistically significant difference for all classes ($p < 0.01$/Mann-Whitney-U-Test). Focal loss was outperformed by all other losses for classes enamel, dentin and root canal ($p < 0.01$/Kruskal-Wallis with post hoc Dunn's). For the class filling, performances from standalone loss functions ranged between 0.59 (Dice) and 0.66 (Tversky) in F1-score. Even the Tversky loss improved the results in comparison to the Focal loss, the hybrid loss functions Generalized Dice Focal Loss and the Dice Focal loss outperformed all standalone loss functions ($p < 0.01$).

For the minority class with the lowest occurrence, crowns, performance differed more widely between the loss functions and ranged between an F1-score of 0.05 (Dice) and 0.75 (Focal Dice). The Focal Dice Loss outperformed all other losses ($p < 0.05$). All detailed results are reported in Appendix Table S2. Secondary metrics were represented in Appendix Table S3. Examples of the predictions provided for an exemplary radiographic bitewing with the U-Net architecture for the different loss functions are represented in Fig. 3.

Further, measured on the standard deviation of the performances reached over the different architectures with hybrid and standalone loss

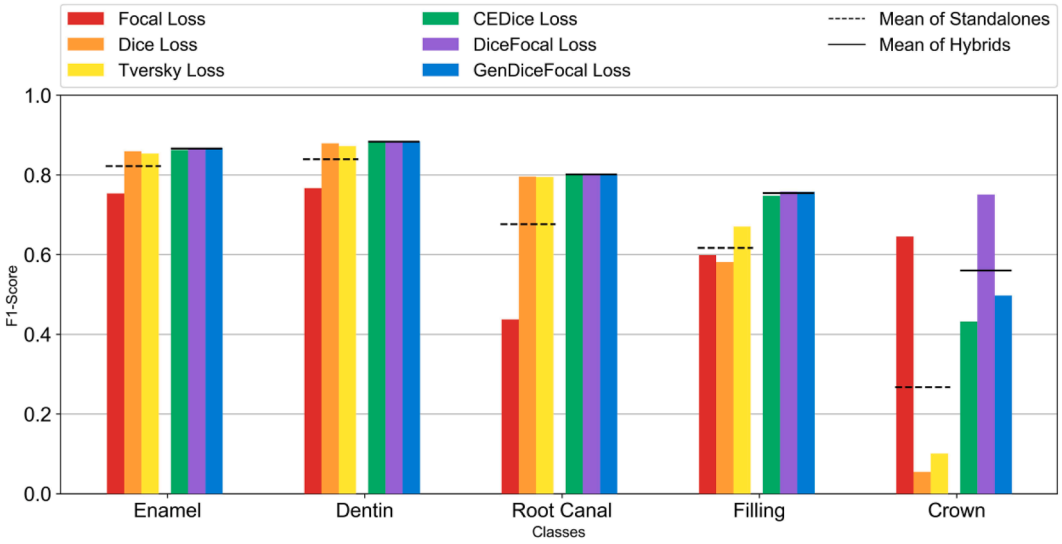

Fig. 2. Mean F1-scores reached with the different loss functions, namely Focal Loss, Dice Loss, Tversky Loss, Cross-Entropy Dice Loss (CEDice), Dice Focal Loss (DiceFocal) and Generalized Dice Focal Loss (GenDiceFocal). Mean values were aggregated from the results of the 5-fold cross-validation with different model architectures.

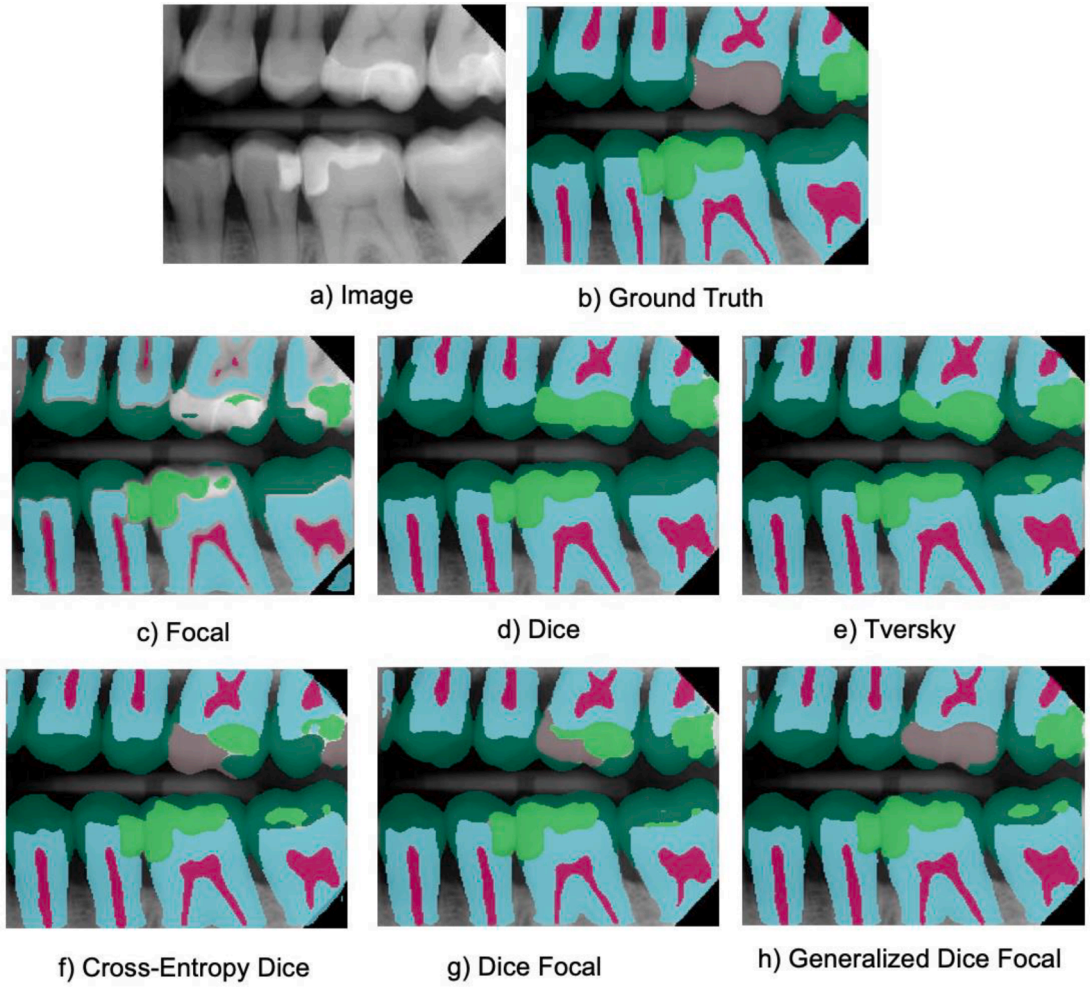

Fig. 3. Exemplary representation of a bitewing radiograph (a) with its Ground Truth (b) and predictions provided by models based on the U-Net architecture trained with the six different loss functions (c-h).

functions, hybrid loss functions reached more robust results for classes enamel, dentin, root canal and filling ($p < 0.01$). Only for the class with the least occurrence in the dataset, crowns, there was no significant difference between the robustness of single and hybrid losses ($p > 0.2$).

4. Discussion

In the present study, we compared six different loss functions regarding their performance on an imbalanced segmentation task in dentistry. We found significant differences when different loss functions were employed; these differences appeared consistently over different model architectures employed and also, by large, across different classes. Notably, performances were highest for the most common classes, enamel and dentin, and lower for less prevalent classes, filling and crown. Hybrid loss functions universally outperformed single loss functions with statistically significant difference for all classes ($p < 0.01$/Mann-Whitney-U-Test). Hence, our hypothesis that hybrid loss functions outperformed standalone loss functions was accepted. Our findings require further discussion.

For Dice and Tversky loss, their ability to address the imbalance between the background class and the classes of interest seems to lead to good performance for the majority classes. However, the lack of targeting the imbalance between individual was evident in the low performance of the minority classes, where they reached significantly worse performances than hybrid losses. Focal loss showed a slightly different behavior: It was already inferior for the majority classes but reached the second-best performance for the class with the lowest occurrence. This exactly aligns with the functionality of the focal loss explained above. By Down Weighting of the majority classes, its performance on these classes degrades while minority class segmentation is improved. Hybrid losses were able to bridge the shortcomings of different loss functions; the hybrid Dice Focal loss reached consistent results over all classes, for examples. This is in line with previous studies, which found that optimization for multiple objectives in hybrid loss functions improves the convergence of the training process and allows the model to find more stable parameters [30].

We further assessed the impact of different loss functions when using different model architectures. As different model architectures have different complexity, feature sensitivity and learning robustness may lead to different robustness towards class imbalances. The good performance of the hybrid loss function for the classes enamel, dentin, root canal, filling regardless of the model architecture leads us to the assumption that the choice of loss function is more important than the choice of the architecture when dealing with class imbalance. The superior performance of the focal loss for the class crown supports this assumption. Our findings have implications for researchers in the field. If facing imbalances, testing different loss functions should be attempted for optimizing performances. However, there is no one-size-fits-all solution and hence, we suggest experimentation with varying loss functions for each individual task. Moreover, the context and clinical relevance of identifying different classes should be considered. Combining this experimentation with different architectures may be of less relevance than assessing the impact of loss functions for different classes and tasks, as demonstrated in our case.

This study was subjected to a range of strength and limitations. It is the first holistic evaluation of the ability of different loss functions to overcome class imbalance in dentistry and may provide guidance for dental researchers in the selection of loss functions for the training process of their DL models. The inclusion of different model architectures as well as the extensive hyperparameter tuning process was another strength. However, as this analysis was only conducted on one task, a tooth structure segmentation task on bitewing radiographs, we cannot claim generalizability to other tasks. In addition, data stemmed only from one center, and transferability to other datasets may not be fully given. A similar constraint pertains to different data modalities (e.g. other radiographs, or generally other imagery). Further, data imbalance was limited on image-level (but more pronounced on pixel-level). For more aggravated class imbalance, combining the optimal loss function with other means of addressing imbalance (e.g., the use of GANs or Diffusion Models) may be needed. Lastly, all employed architectures were based on Convolutional Neural Networks and, hence, may not be applicable to other architecture types, e.g. those based on Vision Transformers.

5. Conclusions

In the present study, six different loss functions were assessed regarding their ability to tackle class imbalance on the exemplary task of tooth structure segmentation on bitewing radiographs trained with three different model architectures. Hybrid loss functions such as the Cross-Entropy Dice loss, Focal Dice loss and Generalized Dice loss reached overall better performances than single loss functions e.g., Dice loss, Tversky loss and Focal loss. This superiority of hybrids was confirmed for four out of five classes and different model architectures. Modelers faced with class imbalance should test for the impact of different loss functions in addition to other, more common strategies for tackling imbalance.

CRediT authorship contribution statement

Martha Büttner: Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Conceptualization. **Lisa Schneider:** Writing – original draft, Visualization, Resources, Methodology, Formal analysis, Conceptualization. **Aleksander Krasowski:** Writing – review & editing, Supervision, Resources, Methodology. **Vinay Pitchika:** Writing – review & editing, Validation, Methodology. **Joachim Krois:** Funding acquisition, Data curation, Conceptualization. **Hendrik Meyer-Lueckel:** Writing – review & editing, Supervision, Investigation, Funding acquisition. **Falk Schwendicke:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Declaration of competing interest

FS and JK are co-founders of the startup dentalXrai GmbH. dentalXrai GmbH did not have any role in conceiving, conducting or reporting this study. The authors are solely responsible for the contents of this paper.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number: 445925495.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2024.105063.

References

[1] L.T. Arsiwala-Scheppach, A. Chaurasia, A. Müller, J. Krois, F. Schwendicke, Machine learning in dentistry: a scoping review, J. Clin. Med. 12 (2023), https://doi.org/10.3390/jcm12030937.
[2] J. Krois, T. Ekert, L. Meinhold, T. Golla, B. Kharbot, A. Wittemeier, C. Dörfer, F. Schwendicke, Deep learning for the radiographic detection of periodontal bone loss, Sci. Rep. 9 (1) (2019) 1–6.
[3] T. Ekert, J. Krois, L. Meinhold, K. Elhennawy, R. Emara, T. Golla, F. Schwendicke, Deep learning for the radiographic detection of apical lesions, J. Endod. 45 (7) (2019) 917–922.
[4] J.-H. Lee, D.-H. Kim, S.-N. Jeong, S.-H. Choi, Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm, J. Dent. 77 (2018) 106–111.
[5] F. Schwendicke, W. Samek, J. Krois, Artificial intelligence in dentistry: chances and challenges, J. Dent. Res. 99 (7) (2020) 769–774.

[6] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259, https://doi.org/10.1016/j.neunet.2018.07.011.
[7] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (2016) 221–232, https://doi.org/10.1007/s13748-016-0094-0.
[8] R. Mohammed, J. Rawashdeh, M. Abdullah, Machine learning with oversampling and undersampling techniques: overview study and experimental results, in: Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, 2020, pp. 243–248.
[9] V. Sampath, I. Maurtua, J.J. Aguilar Martin, A. Gutierrez, A survey on generative adversarial networks for imbalance problems in computer vision tasks, J. Big Data 8 (2021) 1–59.
[10] L.W. Sagers, J.A. Diao, M. Groh, P. Rajpurkar, A.S. Adamson, A.K. Manrai, Improving dermatology classifiers across populations using images generated by large diffusion models, arXiv Preprint (2022).
[11] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Ieee, 2016, pp. 565–571.
[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[13] K.P. Murphy, Machine learning: a Probabilistic Perspective, MIT press, 2012.
[14] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: Proceedings of the International Workshop on Machine Learning in Medical Imaging, Springer, 2017, pp. 379–387.
[15] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, Y. Pan, Rethinking dice loss for medical image segmentation, in: Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 851–860.
[16] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, Springer, 2017, pp. 240–248. September 14, Proceedings 3.
[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional Networks for Biomedical Image Segmentation, in: Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a.
[18] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[19] A. Chaurasia, E. Culurciello, Linknet: exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2017, pp. 1–4.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[22] M. Büttner, L. Schneider, A. Krasowski, J. Krois, B. Feldberg, F. Schwendicke, Impact of noisy labels on dental deep learning–calculus detection on bitewing radiographs, J. Clin. Med. 12 (2023), https://doi.org/10.3390/jcm12093058.
[23] L. Schneider, R. Rischke, J. Krois, A. Krasowski, M. Büttner, H. Mohammad-Rahimi, A. Chaurasia, N.S. Pereira, J.-H. Lee, S.E. Uribe, S. Shahab, R.B. Koca-Ünsal, G. Ünsal, Y. Martinez-Beneyto, J. Brinz, O. Tryfonos, F. Schwendicke, Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs, J. Dent. 135 (2023) 104556, https://doi.org/10.1016/j.jdent.2023.104556.
[24] P. Iakubovskii, Segmentation Models Pytorch, GitHub Repository (2019). https://github.com/qubvel/segmentation_models.pytorch.
[25] L. Schneider, P. Dave, L. Arsiwala-Scheppach, F. Schwendicke, J. Krois, Exploring bias in F-score computation methods of multi-class segmentation models, in: Proceedings of the 2021 The 5th International Conference on Video and Image Processing, 2021, pp. 76–84.
[26] P.E. McKight, J. Najab, Kruskal-wallis test. The Corsini Encyclopedia of Psychology, 2010, p. 1.
[27] O.J. Dunn, Multiple comparisons using rank sums, Technometrics 6 (1964) 241–252.
[28] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Series B Stat. Methodol. 57 (1995) 289–300.
[29] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. (1947) 50–60.
[30] T. Lodkaew, K. Pasupa, Hybrid loss for improving classification performance with unbalanced data, in: H. Yang, K. Pasupa, A.C.-S. Leung, J.T. Kwok, J.H. Chan, I. King (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2020, pp. 807–814, https://doi.org/10.1007/978-3-030-63820-7_92.