BMC Oral Health

RESEARCH

Open Access

Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning

Sung-Jae Lee[1†], Hyun Jun Oh[2†], Young-Don Son[1,3], Jong-Hoon Kim[3,4], Ik-Jae Kwon[5,7], Bongju Kim[6], Jong-Ho Lee[2,6*] and Hang-Keun Kim[1,3*]

Abstract

Background Oral potentially malignant disorders (OPMDs) are associated with an increased risk of cancer of the oral cavity including the tongue. The early detection of oral cavity cancers and OPMDs is critical for reducing cancer-specific morbidity and mortality. Recently, there have been studies to apply the rapidly advancing technology of deep learning for diagnosing oral cavity cancer and OPMDs. However, several challenging issues such as class imbalance must be resolved to effectively train a deep learning model for medical imaging classification tasks. The aim of this study is to evaluate a new technique of artificial intelligence to improve the classification performance in an imbalanced tongue lesion dataset.

Methods A total of 1,810 tongue images were used for the classification. The class-imbalanced dataset consisted of 372 instances of cancer, 141 instances of OPMDs, and 1,297 instances of noncancerous lesions. The EfficientNet model was used as the feature extraction model for classification. Mosaic data augmentation, soft labeling, and curriculum learning (CL) were employed to improve the classification performance of the convolutional neural network.

Results Utilizing a mosaic-augmented dataset in conjunction with CL, the final model achieved an accuracy rate of 0.9444, surpassing conventional oversampling and weight balancing methods. The relative precision improvement rate for the minority class OPMD was 21.2%, while the relative F_1 score improvement rate of OPMD was 4.9%.

Conclusions The present study demonstrates that the integration of mosaic-based soft labeling and curriculum learning improves the classification performance of tongue lesions compared to previous methods, establishing a foundation for future research on effectively learning from imbalanced data.

Keywords Tongue cancer, Class imbalance, Deep learning, Mosaic augmentation, Curriculum learning

†Sung-Jae Lee and Hyun Jun Oh contributed equally to this work.

*Correspondence:
Jong-Ho Lee
leejongh@snu.ac.kr
Hang-Keun Kim
dsaint31@gachon.ac.kr
Full list of author information is available at the end of the article

Introduction

Oral cavity cancer accounted for approximately 377,000 new cases and 177,000 related deaths worldwide in 2020 [1], highlighting its significance as a public health issue. Tongue cancer is frequently diagnosed in many countries, making it an important area of focus. While oral cavity cancer is associated with high morbidity and mortality rates [2], oral potentially malignant disorders (OPMDs) can also increase the risk of developing this type of cancer, including tongue cancer [3, 4]. Therefore, it is essential to accurately and easily diagnose oral cavity cancers and OPMDs to prevent their progression. Accurate and accessible diagnosis techniques can lead to timely treatment and reduce cancer-specific morbidity and mortality [5–7].

Recently, the development of artificial intelligence has led to the use of deep learning to detect oral cavity cancers and OPMDs [8–15]. VGG [16], ResNet [17], and EfficientNet [18] techniques were commonly utilized in these studies. Most studies [11–14] have focused on binary classification, classifying oral lesions as either malignant or benign. Only a few studies [15] have investigated multi-class classification. Recent work showed that it is recommended to use a moderately complex convolutional neural network (CNN) with a data-bypassing architecture when working with a limited dataset. Nevertheless, one of the remaining challenges is addressing the class imbalance [19, 20]. Class imbalance, a prevalent issue in medical imaging applications, especially for cancer detection, occurs when certain classes are disproportionately represented in a dataset. This disparity can degrade classifier performance by neglecting minority classes [21–24]. In this study, the dataset was imbalanced, with the cancer and OPMD categories having fewer samples.

While the representativeness of a dataset is crucial for the effectiveness of deep learning algorithms, class imbalance is a frequent challenge in medical imaging applications, making it difficult to acquire representative datasets. Consequently, it is imperative to develop methods that enable the effective training of deep learning models on imbalanced and limited datasets. The goal is to ensure these models can achieve performance levels comparable to those trained on representative datasets. Several approaches have been proposed to address this issue, including data-level, algorithm-level, and hybrid methods [20, 25]. An effective data-level method is data augmentation, which can increase the diversity of a training dataset by applying data transformations. Examples of these augmentation techniques include Cutout [26], CutMix [27], Random Image Cropping and Patching (RICAP) [28], and Mosaic augmentation [29]. Cutout and CutMix techniques can enhance the performance of machine learning models by manipulating important parts of the input data. This helps the model to learn more robust features, resulting in better performance. These techniques have shown promise when applied to various models and datasets, making them a promising area for future machine learning research. RICAP is an augmentation technique that enhances the diversity of training datasets. It uses four random crops from the input images and patches them together to form a single image. This approach enables the model to learn from more diverse data and helps prevent overfitting. However, a significant limitation of RICAP, particularly in tongue cancer detection, is its propensity to lose lesions in the produced images. This problem occurs mainly due to the small size of the lesions and their frequent placement on the lateral edges of the tongue, which may be accidentally excluded during the random cropping process. Mosaic augmentation combines four image patches and resizes them into a single synthesized image to detect objects that may not be easily recognizable in their normal context owing to differences in scale [30]. Although Mosaic augmentation has proven to be effective in detecting objects, it is important to note that it was originally designed for object detection and may not be directly applicable to classification tasks.

In this study, we introduced an effective technique called "mosaic-based soft labeling" augmentation combined with curriculum learning (CL) [31]. CL begins with teaching the model using simpler training dataset (or task) and then gradually introduces more complex training dataset (or task). This method mimics human learning, allowing the model to build on previously learned concepts from simpler dataset, thereby facilitating the understanding of more complex concepts more accurately. Using this approach, augmented data with different levels of complexity are trained step-by-step accordingly. The purpose of this study was to evaluate the performance of mosaic-based soft labeling and CL compared to other methods, such as oversampling and weight balancing, in the class-imbalanced tongue lesion dataset.

Methods

Dataset

From January 2006 to December 2020, a total of 1,810 tongue images were acquired from patients aged over 20 years who visited the Department of Oral and Maxillofacial Surgery at Seoul National University Dental Hospital in Seoul, Republic of Korea, for diagnostic purposes or periodic check-ups. The clinical photographs were consistently captured using a single-lens reflex camera (D750, Nikon, Japan) by a single researcher under the supervision of the author (JH Lee). The captured images were saved in JPEG (Joint Photographic Experts Group)

format. The Institutional Review Board of the Seoul National University Dental Hospital approved the collection and use of this dataset (ERI22034).

The images were categorized into three distinct groups with an average resolution of 723×734 pixels: cancer, OPMD, and noncancerous lesions (Table 1). The images were categorized into three distinct groups with an average resolution of 723×734 pixels: cancer, OPMD, and noncancerous lesions (Table 1). Notably, the dataset used in this study consists of cases from the cancer class (20.6%), OPMD class (7.8%), and noncancerous lesion class (71.6%), with a significant imbalance between classes. Tongue cancer was defined as squamous cell carcinoma confirmed through pathological examination. OPMDs were diagnosed in accordance with the World Health Organization consensus report classification [3]. Lesions including pathologically confirmed oral lichen planus, leukoplakia, or erythroplakia were classified as OPMDs. Noncancerous lesions encompassed healthy tongue tissues and benign conditions such as hemangioma, fibroma, or aphthous ulcer. Two licensed surgeons with a minimum of 5 years of clinical experience (IJK and JHL) were responsible for the categorization of tongue images. For the test set, 10% (180 images) of the entire dataset was randomly selected to reflect the data population. The remaining dataset was used as training and validation datasets at 70% and 30%, respectively.

Data augmentation: mosaic-based soft labeling

To address the class imbalance and limited quantity of training data, we proposed a new augmentation technique named "mosaic-based soft labeling" (Fig. 1). This technique was motivated by two existing methods: *RICAP* [28] and *mosaic data augmentation* [29]. However, in contrast to these approaches, we used Gradient-weighted Class Activation Mapping (Grad-CAM) [32] to extract the representative regions of each image. Grad-CAM calculates the relationship between output and input images by using a gradient. This produces a heatmap that emphasizes the areas of the input image used by the network to determine the target class. Otsu's thresholding converts the Grad-CAM importance map into a binary map consisting of representative and nonrepresentative regions. Next, we determined the bounding box that covers the largest representative region in the resulting binary map. Finally, we combined the representative patches using cropping and resizing, which helped train the network to recognize representative patterns in the imbalanced data. The resulting synthesized image forms a mosaic-based soft labeling dataset.

In Mosaic augmentation, four images are combined to create a single synthesized image. However, because our goal was to classify the data into three categories, we selected one patch from each class and randomly chose the remaining patches using random sampling with replacement. The four selected representative patches were then combined to form a synthesized image. The label was calculated using the area covered by each class, with a sum of 1. The portion of the area occupied by each class can be interpreted as a probability associated with that specific class, intrinsically utilizing the soft labeling technique [33]. The model is trained to accurately predict the value of each component in the soft label vector. The soft labeling technique replaces the one-hot ground truth with smoothed labeling, which is known to improve classification performance by decreasing overconfidence and increasing generalization.

Table 1 Categories and number of tongue lesion samples

Categories	Number of samples
cancer	372
OPMD	141
noncancerous lesion	1,297
Total	1,810

Training: CL using the mosaic-based soft labeling

CL consists of several different training stages, each of which becomes progressively more challenging. In our study, we used CL consisting of two stages. Stage 1 is a conventional training using an original dataset with the weight balancing method. 1,630 images from the dataset are utilized, excluding the 180 images reserved for the test set. Mosaic-based soft labeling dataset is created using Grad-CAM with the model trained in Stage 1. In Stage 2, the model is retrained using a newly added mosaic-based soft labeling dataset and the original dataset. 1,200 mosaic-based soft labeling images (synthesized images) are utilized. Note that the test set of 180 images is not used for training of either stage. An overview of our CL is shown in Fig. 2.

The details of our CL stage 1 are as follows. The EfficientNet model was used as a feature extraction model for classification. This model outperforms other existing CNNs with similar computational costs by utilizing a scaling method that uniformly scales up the width, depth, or resolution [18]. The EfficientNetB0 model was used in this study. Deep learning models must be trained from scratch when the training data have different feature spaces and data distributions, which leads to an inefficient situation. When the available training dataset is relatively small, transfer learning is an effective method for training deep models without overfitting, and it has

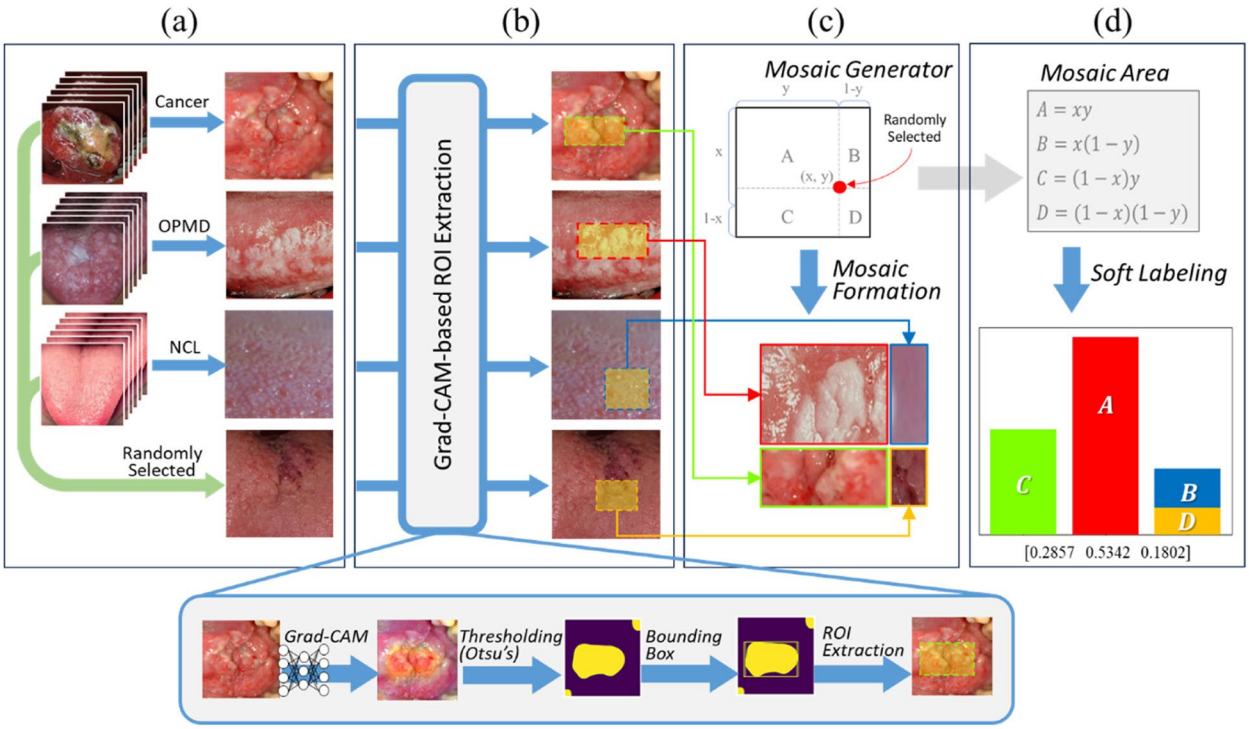

Fig. 1 A detailed description of mosaic formation and soft labeling process. **a** A set of four images consisting of a mosaic image. One image is selected for each class and the last one is additionally and randomly selected from all classes (in this case, the noncancerous lesion). **b** A Grad-CAM-based ROI extraction proposes a representative region of each image. A class activation map of each image is extracted from the trained model without mosaic dataset. **c** The mosaic generator randomly chooses a point within a square to form a 2×2 grid so that one segment should be, at least, larger than half of the synthesized image. A mosaic image is synthesized with four cropped and rescaled images from each Grad-CAM-based ROI. The position of each image is randomly assigned each time a synthesized image is created. Also, the oversampling rate is adjustable, allowing for larger areas to be allocated to user-defined classes. (in this case, A: OPMD, B: noncancerous lesion, C: cancer, D: noncancerous lesion) **d** The soft label is calculated with the areas of each class within the grid. The model is trained to accurately predict the value of each component in the soft label vector

been used in many studies [34–37]. Transfer learning involves reusing learned knowledge from a base dataset; typically, a large-scale dataset such as ImageNet is used [38]. In this study, transfer learning was performed using pretrained weights from ImageNet [38]. Conventional data augmentation was applied to the training dataset to increase its diversity. The Inputs were zoomed in by 80%-100% at random and flipped horizontally and vertically. They were rotated at random angles ranging from –30° to 30°. In addition, they were shifted to the left or right at random between -10% and 10% of the total width, and vertically in the same manner. Outside the input boundaries, the points were filled with the nearest pixels. Finally, fine-tuning was performed to train the network on the tongue cancer dataset. After fine-tuning, the network was retrained entirely on the tongue cancer dataset with a low learning rate owing to the significant differences between the ImageNet and tongue cancer datasets. This approach is commonly used when applying deep neural networks to medical applications and helps overcome the scarcity of medical datasets [35, 39]. Furthermore, because of the highly imbalanced class distribution of the datasets, a weight balancing method [40] was utilized. This method heavily penalizes misclassified predictions from the minority class and is adopted to address the class imbalance issue. At the end of Stage 1, the mosaic-based soft labeling dataset is created in the manner of the Mosaic-based soft labeling section.

In Stage 2 of our CL, the model resulting from Stage 1 was retrained using an augmented dataset that included both the original tongue cancer dataset and the mosaic-based soft labeling dataset. The weights are initialized and trained from the beginning. For the mosaic-based soft labeling dataset, the rotation and flipping were applied to eliminate the effect of cross-shaped lines between images in a synthesized image. The rotation range was set at random angles between -5° and 5°.

Test: prediction of the trained model

In the field of machine learning, including deep learning, prediction refers to the process of using a trained model to make a classification or regression on new

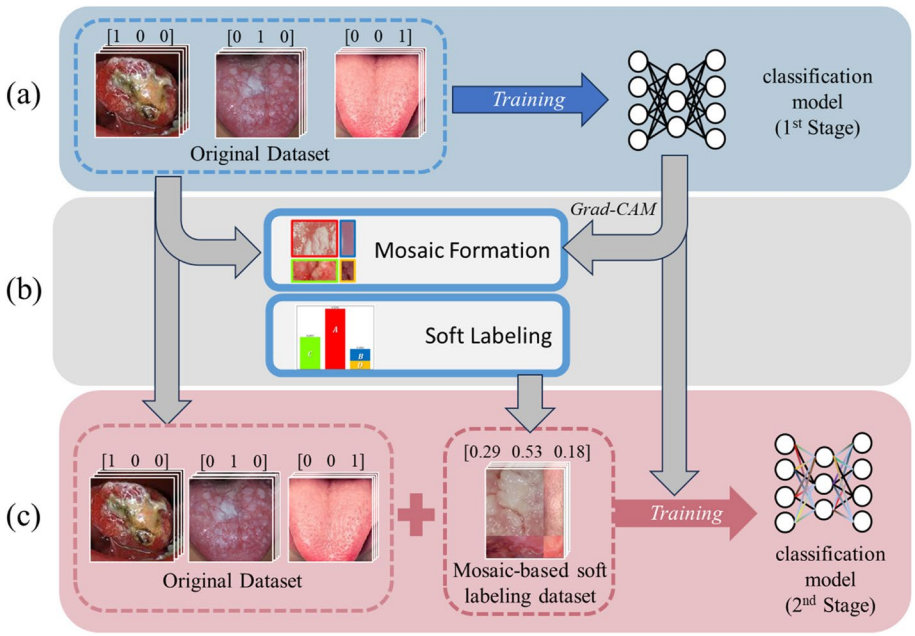

Fig. 2 An overview of mosaic-based soft labeling with curriculum learning. Our curriculum learning is consisted of two stages. **a** In stage 1, a conventional training is performed using an original dataset with the weight balancing method. Transfer learning, data augmentation, and fine-tuning was also used during training. **b** Using the trained model in Stage 1, a mosaic dataset and the corresponding soft labels can be obtained as described in Fig. 1. **c** In stage 2, the final model is trained with the original dataset and a newly added mosaic-based soft labeling dataset

data that was not included in the training set [41]. Prediction is also called inference. In this study, we have trained three deep learning models, each sharing the same underlying architecture but differing in the applied training technique. The first model employs a weight balancing technique, the second utilizes oversampling technique, and the third is trained using our proposed method, which integrates mosaic-based soft labeling and CL.

To assess performance, all three models are tested using a uniform test set, with comparative metrics derived from the outcomes of this test, which fundamentally involves an inference process. As described in the Dataset section, the test set comprises images that are not utilized during the training. It is noted that the mosaic images generated through mosaic-based soft labeling are not included in the test set.

Figure 3 shows how the prediction works in our proposed method. During the prediction stage, where the models are fed only real images, each model produces a result vector for each input image. Subsequently, each image is classified into the class corresponding to the highest value component in its result vector. In contrast, the training stage uses both real and synthetic images. The result vector itself is considered the final output of each model. This vector is then compared with the corresponding label — a soft label for synthetic images and a one-hot encoding label for real images — to calculate the loss function, which are then minimized through continued training.

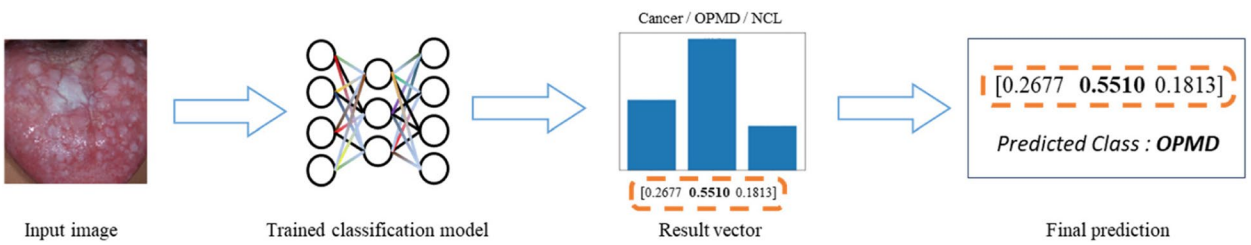

Fig. 3 Prediction process of the proposed method. During the prediction stage, where the models are fed only real images, the model produces a result vector for each input image. The image is classified into the class corresponding to the highest value component in its result vector

Metrics and statistical analysis

We computed the accuracy, precision, recall, and F_1 score using a test set to assess model performance. Accuracy is the ratio that indicates how accurate the classification was among all classification attempts. Accuracy ranges from 0 to 1. A score of 0 suggests no accuracy, meaning all classifications are incorrect, while a score of 1 indicates perfect accuracy with all classifications being correct. Accuracy is a commonly used metric for evaluating classification models because of its simplicity. However, it may not be suitable for imbalanced datasets [42]. Therefore, additional metrics to complement basic accuracy need to be considered. Precision is a measure that indicates the accuracy of positive predictions, specifically for a particular class. It shows how many of the predicted instances for that class are actually true instances of that class. On the other hand, recall, also known as sensitivity, is a measure that indicates the ability to detect all instances of a specific class through predictions. It shows how many of the total instances of that specific class are correctly identified. F_1 score is a single metric that combines both precision and recall into a single value, providing a balanced measure of a model's performance in classification tasks, especially when there is an imbalance between the classes. It is particularly useful when you want to find a balance between precision and recall, as they are often in tension with each other. Precision, recall, and F_1 score range from 0 to 1, with 0 indicating the worst possible performance and 1 indicating the best possible performance.

In this paper, we performed multi-class classification with three categories. Therefore, we calculated weighted average and macro average (Eq. 3) for precision, recall, and F_1 score, based on the averages of each class.

$$\text{macro average} = \frac{1}{3} \times (\text{score}_{\text{cancer}} + \text{score}_{\text{OPMD}} + \text{score}_{\text{noncancerous lesion}})$$

$$\text{weighted average} = \text{W}_{\text{cancer}} \times \text{score}_{\text{cancer}} + \text{W}_{\text{OPMD}} \times \text{score}_{\text{OPMD}} + \text{W}_{\text{noncancerous lesion}} \times \text{score}_{\text{noncancerous lesion}}$$

where

W_{cancer} = number of cancer samples divided by number of total samples
W_{OPMD} = number of OPMD samples divided by number of total samples
$\text{W}_{\text{noncancerous lesion}}$ = number of noncancerous lesion samples divided by number of total sample

Weighted average is a metric calculation that considers the contribution of each class or category in proportion to its prevalence in the dataset. It assigns more weight to classes with larger sample sizes, which is beneficial when dealing with imbalanced datasets. Weighted averages are commonly used in classification tasks. However, in cases of class-imbalanced data, it may not be appropriate. In contrast, the macro average is a metric calculation that independently computes the metric for each class and then calculates the unweighted average (simple arithmetic mean) of those class-specific metrics. It treats all classes equally, regardless of their prevalence in the dataset, providing a balanced assessment of model performance across all classes. Macro average is often used to evaluate the overall model performance when all classes are considered equally important.

We also used the relative performance improvement rate as an indicator of how much the model's performance improved compared to other conventional methods (Eq. 4).

$$\text{relative performance improvement rate}(\%) = 100 \times \frac{(\text{Performance}_{\text{new}} - \text{Performance}_{\text{previous}})}{\text{Performance}_{\text{previous}}}$$

Results

Table 2 shows the accuracy comparison across three different models, while Table 3 presents their other performance metrics.

The first model (WB), which was only trained up to Stage 1, used the conventional weight balancing technique to address the class imbalance. The overall accuracy of Stage 1 was 0.9278. The precision and F_1 score of OPMD (minority class) were 0.6875 and 0.7333, respectively.

Oversampling (OB) is another well-known technique to address class imbalance. The second model was trained using the oversampling method instead of the weight balancing method, while following the same

Table 2 Accuracy comparison among weight balancing technique (WB), oversampling technique (OS), and mosaic-based soft labeling (MBS)

Accuracy		
WB	**OS**	**MBSO**
0.9278	0.9111	**0.9444**

procedure as the first model. Cancer and OPMD images were sampled using replacement to match the number of the majority class. The overall accuracy of oversampling was 0.9111. The oversampling method showed poorer performance on each metric when compared to the weight balancing method.

The final model (MBO), which was trained using the newly added mosaic-based soft labeling dataset (Stage 2), achieved an accuracy rate of 0.9444. This outperformed conventional oversampling (0.9111) and weight balancing methods (0.9278). Our approach demonstrated comparable or slightly lower performance in various categories compared to conventional methods. However, it significantly improved performance in the OPMD category, which is characterized by a notably small sample size. We saw a 21.2% increase in relative precision improvement and a 4.9% increase in relative F_1 score improvement rate when compared to conventional weight balancing method (precision: 0.6875 → 0.8333; F_1: 0.7333 → 0.7692). This suggests that our method trains the classifier model more effectively when handling class-imbalanced training data, compared to traditional approaches.

Discussion

Medical data are frequently imbalanced, which can negatively affect classification performance because the model may not properly capture the minority class [43]. To address this issue, various methods have been introduced, including oversampling and weight balancing. Our proposed method, when tested on an imbalanced tongue lesion dataset, yielded an accuracy rate of 0.9444. This represents a modest improvement over the accuracy rates of conventional oversampling (0.9111) and weight balancing methods (0.9278). Although our proposed method occasionally exhibits lower performance compared to the Weight Balancing (WB) technique, it typically shows improvements across most metrics. This is particularly evident in the F_1 score, which accounts for both precision and recall, where our method consistently demonstrates a noticeable improvement. For instance, in the context of the OPMD category's F_1 score, the WB technique achieved 0.7333, and the Oversampling technique reached 0.5600. Our proposed method, on the other hand, attained a score of 0.7692.

Our proposed mosaic-based soft labeling may demonstrate the ability to achieve similar performance to using a class-balanced dataset even when dealing with a class-imbalanced dataset. Jubair et al. [11] and Heo et al. [12] performed binary classification on imbalanced datasets with the existing methods, such as oversampling and weight-balancing. Jubair et al. used 716 images, which consisted of 236 cancerous images and 480 benign images. Heo et al. used 5,576 images, including 3,635 non-cancer images and 1,941 cancer images. Their models achieved an accuracy of 85.0% and 84.7%, respectively. Although the data is different and so not directly comparable, we tackled a more challenging three-class classification using an imbalanced dataset and achieved an accuracy of 94.44%. Notably, our method was able to achieve comparable performance on the imbalanced dataset as the previous study obtained on the balanced dataset with OPMD [15]. Sharma et al. [15] achieved an accuracy of 76% in three-class classification using a conventional CNN with a class-balanced dataset. Their study utilized 329 images, comprising 106 normal, 102 OPMD, and 121 cancer images. Although their dataset was balanced, the F_1 score of OPMD was 0.74. In contrast, in our study, we achieved an F_1 score of 0.7692 for OPMD, despite using a class-imbalanced dataset.

We have proposed a new technique that utilizes mosaic-based soft labeling which provides the combined benefits of soft labeling, oversampling, and traditional

Table 3 Classification performance with weight balancing technique (WB), oversampling technique (OS), and mosaic-based soft labeling (MBS)

	precision			**recall**			**F_1 score**		
	WB	**OS**	**MBS**	**WB**	**OS**	**MBS**	**WB**	**OS**	**MBS**
cancer	0.8293	0.8250	**0.8500**	**0.9189**	0.8919	**0.9189**	0.8718	0.8571	**0.8831**
OPMD	0.6875	0.6364	**0.8333**	**0.7857**	0.5000	0.7143	0.7333	0.5600	**0.7692**
noncancerous lesion	**0.9919**	0.9612	0.9844	0.9457	0.9612	**0.9767**	0.9683	0.9612	**0.9805**
weighted average	0.9348	0.9080	**0.9450**	0.9278	0.9111	**0.9444**	0.9302	0.9086	**0.9441**
macro average	0.8362	0.8075	**0.8892**	**0.8835**	0.7844	0.8700	0.8578	0.7928	**0.8776**

image manipulation-based augmentation. Two-stage CL was used to apply mosaic-based soft labeling into deep learning-based tongue cancer classification.

The soft labeling technique replaces the one-hot ground truth with smoothed labeling, known to enhance generalization performance. Szegedy et al. demonstrated a 0.2% reduction in error for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 by incorporating soft labeling into training [44]. In our proposed method, each component of the label represents the area occupied by each class's corresponding regions, intrinsically having the effect of soft labeling.

Oversampling and undersampling are the resampling techniques [45–47] that are commonly used to address imbalanced classifications [48, 49]. Oversampling increases the minority class instances, whereas undersampling reduces the majority class instances. Our proposed method has an intrinsic oversampling effect owing to random sampling with replacement and allocation of larger areas to the minority class during dataset generation. This allows users to adjust the oversampling ratio by setting the number of synthesized images and the area of each patch in the synthesized image.

In our experiments, the models trained only with mosaic-based soft labeling images performed worse, apparently because the cancer- or OPMD-independent patterns present in the synthesized images had a negative impact during training. Therefore, conventional image transformations, such as rotation and translation, are necessary to ensure that the network does not focus on useless patterns that are independent of the region of interest, such as a grid pattern, which is a cross-shaped line between images in a synthesized image.

The dataset employed in this study comprised only 1,810 images of tongues and was highly imbalanced. Specifically, the OPMD class accounts for only 7.8% of the data. Generally, OPMD patients are more prevalent than cancer patients, so cancer images are expected to be less abundant. However, in this study, we collected imaging data from patients who visited the outpatient clinic of a professor specializing in cancer surgery (JHL). Due to the higher frequency of requests for patients requiring cancer surgery, it seemed that there are fewer imaging instances for OPMD compared to cancer. This can result in a poor performance in practical clinical situations. Using the proposed method, along with additional OPMD data collection, we expect that there will be a performance improvement compared to existing data augmentation techniques. We also anticipate that further performance gains can be achieved using the proposed data augmentation method to generate additional images when combined with other techniques to address class imbalance. Based on our experimental results, in the case of training with class-imbalanced datasets, the oversampling technique tends to be more effective compared to the undersampling technique. Additionally, incorporating the weight-balancing technique may lead to some improvements in performance. For those seeking a more balanced and stable outcome, our mosaic-based soft labeling method could prove beneficial, though the degree of improvement might vary. Moreover, if additional data collection is undertaken for classes with fewer samples, better performance can be expected. As highlighted by Halevy et al. [50], collaborative efforts in collecting more clinical tongue images could offer a fundamental solution to the challenges of data scarcity and class imbalance. However, the exact impact of such efforts on performance would require further exploration.

Although this study presents a new method for enhancing tongue cancer classification and provides some practical to remedy class imbalanced problem, it has certain limitations. The first limitation of our study is that it relies solely on tongue images. Oral cancer may also appear in other regions. Additionally, our study did not account for lesion location in the training and analysis of the model. Second, the lower performance in OPMD may be influenced by the limited sample size, but it could also be attributed to the diverse subcategories within OPMD. Because it is not merely a matter of an imbalanced dataset, additional research is required to understand the effect of diverse subcategories within OPMD both on deep learning model training and performance. Thirdly, in this study, only images acquired by one professor's camera were utilized. However, for the detection of images in more universal scenarios, it may be necessary to have images from various environments. As OPMD patients receive care not only from professors in oral and maxillofacial surgery but also from professors in oral medicine, we plan to collaborate with these professors in the future to gather imaging data collaboratively. Fourthly, if a sufficiently large and representative dataset is available, there may not be a significant performance difference between the conventional approach and our proposed method in the ideal situation. Previous studies have reported that the choice of training technique or algorithm has minimal impact on performance when the dataset is representative [50, 51]. In contrast, when the dataset significantly lacks representativeness, training deep learning models effectively becomes a challenge, even when employing our mosaic-based soft labeling technique. This requires further study to identify the range of dataset representativeness where mosaic-based soft labeling exceeds the performance of traditional methods. An initial step in such research could involve quantifying the relationship between the size of the dataset and the effectiveness of mosaic-based soft labeling.

Conclusion

In this study, we propose a novel data augmentation technique called 'mosaic-based soft labeling' to approach optimal performance on an imbalanced tongue cancer dataset. We introduced a CL strategy to generate synthetic mosaic images and improve the classification performance of tongue cancer. In the first stage, we trained a tongue cancer classification model using an imbalanced dataset. Utilizing information from the trained network, we synthesized and labeled mosaic images. The model in the second stage was then trained by incorporating synthetic mosaic images. The proposed approach improved the classification performance of tongue lesions compared to previous methods, such as oversampling and weight balancing. However, it fell somewhat short of our initial expectations. Nevertheless, it establishes a foundation for future research on effectively learning from imbalanced data, a common challenge in many diagnostic applications.

Abbreviations

OPMD	Oral potentially malignant disorders
RICAP	Random Image Cropping and Patching
CNN	Convolutional neural network
CL	Curriculum learning
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
NCL	Noncancerous lesions
ROI	Region Of Interest

Acknowledgements

This research was supported by grants from the Technology Innovation Program (20006105) funded by the Ministry of Trade, Industry & Energy, Republic of Korea, and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant number: HI20C2114).

Authors' contributions

SJ LEE and HJ Oh contributed equally as primary authors, and JH Lee and HK Kim contributed equally as corresponding authors. Conceptualization, HK Kim and JH Lee; methodology, SJ LEE and HJ Oh; validation, YD Son, JH Kim, and IJ Kwon; data curation, B Kim; figure preparation, SJ LEE, HJ Oh, and HK Kim; review and editing, SJ LEE, HJ Oh, JH Lee, HK Kim, YD Son, JH Kim, IJ Kwon, and B Kim; funding acquisition, HK Kim. All the authors have read and agreed to the published version of this manuscript.

Funding

This research was supported by grants from the Technology Innovation Program (20006105), funded by the Ministry of Trade, Industry & Energy, Republic of Korea, and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (Grant number: HI20C2114).

Availability of data and materials

Owing to privacy concerns, the datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The datasets and all procedures were reviewed and approved by the Institutional Review Board of Seoul National University Dental Hospital (approval number: ERI22034). All data were anonymized prior to collection. Because this was a retrospective study, it was practically impossible to obtain consent from the study participants, and the risk to them was extremely low. Therefore, a waiver of documentation of informed consent was used. The Institutional Review Board of Seoul National University Dental Hospital approved the waiver for informed consent.

Consent for publication

No information that could lead to the identification of the study participants is included in the manuscript. Consent for publication is not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Biomedical Engineering, College of IT Convergence, Gachon University, Seongnam, Republic of Korea. [2]Oral Oncology Clinic, National Cancer Center, Goyang, Republic of Korea. [3]Neuroscience Research Institute, Gachon Advanced Institute for Health Science and Technology, Gachon University, Incheon, Republic of Korea. [4]Department of Psychiatry, Gachon University College of Medicine, Gil Medical Center, Incheon, Republic of Korea. [5]Department of Oral and Maxillofacial Surgery, Seoul National University Dental Hospital, Seoul, Republic of Korea. [6]Dental Life Science Research Institute, Seoul National University Dental Hospital, Seoul, Republic of Korea. [7]Dental Research Institute, Seoul National University, Seoul, Republic of Korea.

Received: 24 August 2023 Accepted: 15 January 2024
Published online: 01 February 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin. 2021;71(3):209–49.
2. Moore SR, Johnson NW, Pierce AM, Wilson DF. The epidemiology of tongue cancer: a review of global incidence. Oral Dis. 2000;6(2):75–84.
3. Warnakulasuriya S, Kujan O, Aguirre-Urizar JM, Bagan JV, González-Moles M, Kerr AR, Lodi G, Mello FW, Monteiro L, Ogden GR, et al. Oral potentially malignant disorders: a consensus report from an international seminar on nomenclature and classification, convened by the WHO Collaborating Centre for Oral Cancer. Oral Dis. 2021;27(8):1862–80.
4. Rivera C. Essentials of oral cancer. Int J Clin Exp Pathol. 2015;8(9):11884–94.
5. Ojeda D, Huber MA, Kerr AR. Oral potentially malignant disorders and oral cavity cancer. Dermatol Clin. 2020;38(4):507–21.
6. Rajaraman P, Anderson BO, Basu P, Belinson JL, Cruz AD, Dhillon PK, Gupta P, Jawahar TS, Joshi N, Kailash U, et al. Recommendations for screening and early detection of common cancers in India. Lancet Oncol. 2015;16(7):e352-361.
7. van der Waal I, de Bree R, Brakenhoff R, Coebergh JW. Early diagnosis in primary oral cancer: is it possible? Med Oral Patol Oral Cir Bucal. 2011;16(3):e300-305.
8. Shamim MZM, Syed S, Shiblee M, Usman M, Ali SJ, Hussein HS, Farrag M. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. Comput J. 2020;65(1):91–104.
9. Ilhan B, Lin K, Guneri P, Wilder-Smith P. Improving oral cancer outcomes with imaging and artificial intelligence. J Dent Res. 2020;99(3):241–8.
10. Lin H, Chen H, Weng L, Shao J, Lin J. Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. J Biomed Opt. 2021;26(8):086007.

11. Jubair F, Al-karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection of oral cancer. Oral Dis. 2022;28(4):1123–30.
12. Heo J, Lim JH, Lee HR, Jang JY, Shin YS, Kim D, Lim JY, Park YM, Koh YW, Ahn S-H. Deep learning model for tongue cancer diagnosis using endoscopic images. Sci Rep. 2022;12(1):6281.
13. Fu Q, Chen Y, Li Z, Jing Q, Hu C, Liu H, Bao J, Hong Y, Shi T, Li K. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. EClinicalMedicine. 2020;27:100558.
14. Lavanya J, Kavya G, Prasamya N. Oral cancer diagnosis using deep learning for early detection. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS): 2022. New York, USA: IEEE; 2022. p. 1260–1268.
15. Sharma D, Kudva V, Patil V, Kudva A, Bhat RS. A convolutional neural network based deep learning algorithm for identification of oral precancerous and cancerous lesion and differentiation from normal mucosa: a retrospective study. Engineered Science. 2022;18:278–87.
16. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.
17. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14: 2016. Berlin, Germany: Springer; 2016. p. 630–645.
18. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning: 2019. PMLR; 2019. p. 6105–6114. https://proceedings.mlr.press.
19. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis. 2002;6:429–49.
20. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1):27.
21. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106:249–59.
22 Wang S, Yao X. Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybern Part B Cybern. 2012;42(4):1119–30.
23. Krawczyk B, Galar M, Jeleń Ł, Herrera F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl Soft Comput. 2016;38:714–26.
24. Song B, Li S, Sunny S, Gurushanth K, Mendonca P, Mukhia N, Patrick S, Gurudath S, Raghavan S, Tsusennaro I, et al. Classification of imbalanced oral cancer image data from high-risk population. J Biomed Opt. 2021;26(10):105001.
25. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
26. DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:170804552. 2017.
27. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision: 2019. 2019. p. 6023–32.
28. Takahashi R, Matsubara T, Uehara K. Data augmentation using random image cropping and patching for deep CNNs. IEEE Trans Circuits Syst Video Technol. 2019;30(9):2917–31.
29. Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:200410934. 2020.
30. Kaur P, Khehra BS, Mavi EBS. Data augmentation for object detection: a review. In: 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS): 9–11 Aug. 2021 2021. 2021. p. 537–43.
31. Bengio Y, Louradour J, Collobert R, Weston J. Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning: 2009. 2009. p. 41–8.
32. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision: 2017. 2017. p. 618–26.
33. Müller R, Kornblith S, Hinton GE. When does label smoothing help? Adv Neural Inf Process Syst. 2019;32:4694–703.
34. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Adv Neural Inf Process Syst. 2014;27:3320–8.
35. Kandel I, Castelli M. How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. Appl Sci. 2020;10(10):3359.
36. Karimi D, Warfield SK, Gholipour A. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. Artif Intell Med. 2021;116:102078.
37. Ravishankar H, Sudhakar P, Venkataramani R, Thiruvenkadam S, Annangi P, Babu N, Vaidya V. Understanding the mechanisms of deep transfer learning for medical images. arXiv preprint arXiv:170406040. 2017.
38. Deng J, Dong W, Socher R, Li LJ, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition: 20–25 June 2009 2009. 2009. p. 248–55.
39. Ali K, Shaikh ZA, Khan AA, Laghari AA. Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer. Neuroscience Informatics. 2022;2(4):100034.
40. King G, Zeng L. Logistic regression in rare events data. Polit Anal. 2001;9(2):137–63.
41. Géron A. Hands-on machine learning with scikit-learn, keras and tensorflow: concepts, tools and techniques to build intelligent systems. 3rd Edition. Sebastopol, CA: O'Reilly; 2022.
42. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv (CSUR). 2016;49(2):1–50.
43. Anand R, Mehrotra K, Mohan C, Ranka S. An improved algorithm for neural network classification of imbalanced training sets. Neural Netw IEEE Trans. 1993;4:962–9.
44. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2016. 2016. p. 2818–26.
45. Ling CX, Li C. Data mining for direct marketing: Problems and solutions. In: Kdd: 1998. 1998. p. 73–9.
46. Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863–905.
47. He H, Bai Y, Garcia E, Li SA. Adaptive synthetic sampling approach for imbalanced learning. IEEE international joint conference on neural networks. In: IEEE World Congress On Computational Intelligence: 2008. 2008.
48. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl. 2017;73:220–39.
49. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell. 2009;23(04):687–719.
50. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst. 2009;24(2):8–12.
51. Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Toulouse: Association for Computational Linguistics; 2001. p. 26–33.

Publisher's Note