# APS360 Project Final Report: Deep Learning for Oral Disease Classification from Real-Life Dental Images

**Victoria Piroian**
Student# 1006882090
victoria.piroian@mail.utoronto.ca

**Jonah Ernest**
Student# 1007065275
jonah.ernest@mail.utoronto.ca

**Sneha Balaji**
Student# 1007999212
sneha.balaji@mail.utoronto.ca

**Kylie Abela**
Student# 1006913565
kylie.abela@mail.utoronto.ca

## 1 INTRODUCTION

This project aims to classify various common oral diseases from real life dental images using deep learning. Oral diseases strongly affect quality of life with over one in four Canadians, 26% (as seen in Government of Canada (2024)), dealing with oral pain or avoiding foods because of mouth problems. Furthermore, these oral problems disproportionately affect the general public with 27% of individuals with a family income of less than $70,000 experiencing regular pain (as seen in Government of Canada (2024)) compared to only 16% of individuals with a family income of $90,000 or more. Thus, the primary motivation behind this project is to help inform the public about oral disease and to make identifying oral concerns more accessible to the underprivileged populations.

Specifically, this project aims to be used by underprivileged populations as a first access point to addressing their dental concerns. An example use case is if someone is having gum pain, they would upload a photo of their gums to the tool and based on the classification of their concern they can bring up more directed, informed concerns to their dentist or care provider. In order to achieve this, the model would need to correctly classify, within an acceptable accuracy range, oral diseases such as caries, calculus, gingivitis, tooth discolouration and ulcers using a CNN. The inputs to the model will be real-life images of teeth and the output will be a classification into one of the oral disease categories or healthy, Figure 1 illustrates the flow of information through the system. An important part of this project is using real life dental images, making it possible for anyone with access to a camera to input their images. This component of our project presents an interesting challenge as real-life images can be taken from different camera angles, in different lighting and include different parts of the mouth. Thus, the approach of using deep learning was chosen as it has proven (see Subbiah & S (2020)) to be adept at learning complex features from non-uniform scenes.

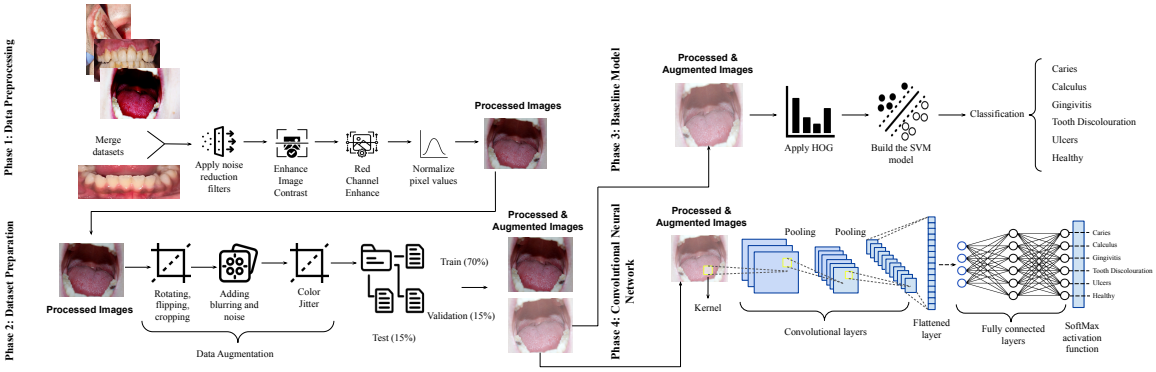



Figure 1: Resulting images after noise reduction

## 2 BACKGROUND AND RELATED WORK

In a project that focused on diagnosing oral and maxillofacial diseases using deep learning, a team of researchers worked on creating a novel neural network with global attention to classify odontogenic lesions from radiographs (Kang et al., 2024). They also proposed a data augmentation technique to create new abnormal cases for effective model training. The result of this study was a model that outperformed current models, increasing classification performance up to 42.4% in recall and 44.2% in F1 scores. This study is relevant to our current project as it proves that novel deep learning approaches can be effective in this problem space. Similarly, a study on deep learning models for classification of dental diseases aimed to create a model to detect and classify the four most common teeth problems: cavities, root canals, dental crowns, and broken-down root canals (Almalki et al., 2022). This study used panoramic X-ray images and achieved a model accuracy of 99.33% and performed better than the currently existing models. This study is relevant to our project as it demonstrates that it is possible to classify, with a high degree of accuracy, similar oral diseases as those in this project. Researchers who looked at mouth and oral disease classification in another study proposed using InceptionResNetV2 to classify diseases such as gingivostomatitis (Gum), canker sores (CaS), cold sores (CoS), oral lichen planus (OLP), oral thrush (OT), mouth can-

cer (MC), and oral cancer (OC) (Rashid et al., 2023). The proposed model achieved an accuracy of 99.51% and is relevant to this project as it demonstrates that deep learning approaches can be adept at classifying both teeth and mouth related diseases. In another study, a partitioned deep CNN with layers for labelling and classification is used to detect and classify oral cancer where researchers focused on computer-assisted medical image classification (Jeyaraj & Nadar, 2019). This novel approach obtained a classification accuracy of 91.4% for benign tumors and an accuracy of 94.5% for cancerous tumors. This particular study is relevant as it shows the viability of a CNN specifically to classify oral diseases, indicating that using a CNN for this project will also be viable. Lastly, a study at Shenzhen Stomatological Hospital Xiong et al. (2024), aimed to create a deep learning model, ToothNet, to detect dental caries and fissure sealants in intraoral photos. The model was able to achieve an AUC (area under the curve) score of 0.925 for caries and 0.902 for sealant detection. This particular study is pertinent for this project as it shows that deep learning approaches can be used effectively to classify oral diseases on real-life images, reaffirming the approach taken in this project.

## 3 Data Processing

This section details the steps taken to collect, clean, and process the dataset, ensuring it is suitable for training and evaluation for the baseline and CNN models.

### 3.1 Dataset Overview

The dataset containing all the images was downloaded from two publicly available sources: "Oral Diseases" dataset (see Sajid (2024)) on Kaggle with images of diseased teeth and "Teeth or Dental image dataset" for the healthy images (Chaudhary et al., 2024). The images from the Oral Diseases dataset were originally divided into 6 classes: caries, calculus, gingivitis, tooth discoloration, ulcers and hypodontia. However, as discussed in the project proposal and progress report, the team decided to not include the hypodontia class as this condition is typically caused by other root causes that the model is predicting, such as carries (Clinic, 2024). In addition, the team randomly sampled 200 healthy teeth images from the additional "Teeth or Dental" dataset to be labelled as healthy for the model to also predict. These images were sampled equally from the 8 different classes in this dataset, each of which represented the angles the images were taken from: maxillary/ upper front, right, left and occlusal views and mandibular/ lower front, right, left and occlusal views. An initial inspection of the 3,158 images revealed the imbalance nature of the dataset, where the majority belonged to the "healthy" and "gingivitis" classes, posing challenges to the model's potential ability to generalize well.

### 3.2 Data Processing Techniques

The various different brightness, orientation, contrast, colour representation and noise level conditions have motivated the use of various image processing techniques as described below. The team has improved upon the data processing strategy used for the progress report, with a new focus placed on reducing image noise, enhancing relevant features targeting underrepresented classes, and ensuring consistency across samples.

The team began processing the data by removing augmented data from the "Oral Diseases" dataset as this dataset also included augmented images. The images from both datasets were then reshaped to be 128x128 RGB to ensure uniform resolution using the transforms.Resize method. The team then experimented with a few major data processing operations including noise reduction, contrast enhancement, red channel enhancement, and standardization. First, noise reduction was performed using fastNlMeansDenoisingColored, a cv2 function, to reduce high-frequency noise without blurring important dental structures. The results are shown in Appendix A, along will the other data processing transformations. Another operation the team tried was contrast mapping, which also used the cv2 functions such as cvtColor, apply and merge for this technique. Next, a red channel enhancement step was introduced for this final model to improve gum visibility. By adjusting contrast in the red channel, this technique emphasized inflammation-related features that are associated with gingivitis. This addition was motivated by earlier misclassification patterns observed in the progress report. The final data processing technique was per-channel standardization to normalize

pixel intensities across all images. The images were scaled to zero mean and unit variance, then a scaling factor was applied to reduce the contrast as the original standardization yielded images with contrast that was too high for reliable detection.

The team also optimized the processing order since creating the model in the progress report to avoid interference between transformations. Denoising was applied first to remove noise, followed by the global contrast enhancement and red channel adjustment. Standardization was then applied last to normalize the processed images before using them as input for the model. An example of all processing steps are shown in Figure 2 below.


Figure 2: Data processing steps on a calculus sample

The class distribution of images after data processing can be seen in Table 1 below.

Table 1: Dataset Overview of Oral Conditions

| **Oral Condition** | **Description** | **Number of Samples** |
|---|---|---|
| Caries | Tooth decay, cavities, or carious lesions. | 219 |
| Calculus | Dental calculus or tartar buildup on teeth. | 1296 |
| Gingivitis | Inflamed or infected gums. | 2349 |
| Tooth Discoloration | Tooth discolouration or staining. | 183 |
| Ulcers | Oral ulcers or canker sores. | 265 |
| Healthy | No indication of an oral condition. | 200 |

### 3.3 Data Augmentation and Splitting

To accommodate the unbalanced nature of the dataset, the team has decided to utilize data augmentation techniques to increase the size of the training dataset, increase minority class representations, and to create variability across the images. The team used different techniques within the BalancedDataset class including rotating the images randomly up to 30 degrees, flipping images horizontally, random cropping, blurring, and adding noise. These augmentations were all applied using the transforms.RandomApply, transforms.RandomResizedCrop, transforms.RandomHorizontalFlip and transforms.RandomRotation methods. To split the images, the team used proportional stratified sampling without replacement, where each subgroup is based on the images classes. This is so that each class can be proportionally represented in the split datasets. The train test split function with the argument stratify=labels was used for this. In addition, a split of 70%, 15% and 15% was used for training, validating and testing.

## 4 Baseline Model

An overview of a traditional machine learning model was implemented to act as a baseline to assess classification performance before transitioning to a deep learning-based approach.

The team used a Support Vector Machine (SVM) as the benchmark for evaluation of the deep learning model. SVM was used for classifying the processed images of dental diseases that are extracted using a Histogram of Oriented Gradients (HOG). HOG was used for feature extraction to capture the structural and texture-based features of the images before classification. Despite the processing, the images still have varying angles, lighting, and colorations, and HOG was able to extract meaningful patterns from a single channel (grayscale representation) of the images to compute the normalized orientation histograms and combine them into a final feature vector for classification. The extracted HOG features were then fed into a SVM model, which is well-suited for high dimensional data.

As the dataset for oral diseases is difficult to classify, the team experimented with different SVM kernels and performed hyperparameter tuning. Using GridSearchCV the team optimized the regularization strength (C), the kernel type (RBF, Poly), and the gamma parameter for the kernels. This systematic searching of the best combination of hyperparameters resulted in more accurate classifications of the dental disease images.

The model achieved a macro-average F1-score of 0.53 and accuracy of 67% on the test set, with the best hyperparameters being regularization strength of 10, a polynomial kernel, and the gamma parameter as scale (see Figure 3 below for the classification report). While this performance is relatively poor, it is still much better than random classification (which would be 16.7% accuracy), and so it provides a good baseline for comparison with the primary model.

Test Accuracy: 0.6706056129985230

Test Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.43 | 0.53 | 194 |
| 1 | 0.42 | 0.33 | 0.37 | 33 |
| 2 | 0.69 | 0.88 | 0.77 | 353 |
| 3 | 0.74 | 0.47 | 0.57 | 30 |
| 4 | 0.29 | 0.15 | 0.20 | 27 |
| 5 | 0.68 | 0.75 | 0.71 | 40 |
| accuracy | | | 0.67 | 677 |
| macro avg | 0.58 | 0.50 | 0.53 | 677 |
| weighted avg | 0.66 | 0.67 | 0.65 | 677 |

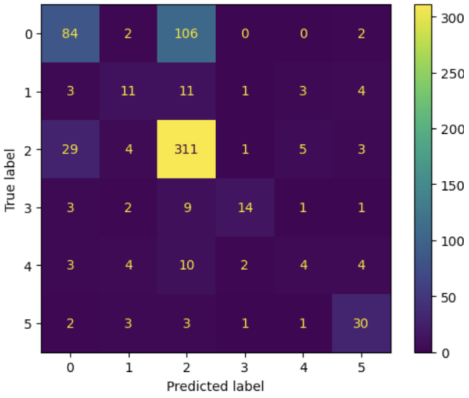



Figure 3: Baseline model classification report and confusion matrix

The confusion matrix in Figure 3 highlights the model's strengths and weaknesses in classifying the six dental conditions. While the SVM model performed well in classifying the ulcers and caries (with an F1-score of 0.71 and 0.77), it struggled with the calculus and tooth discoloration classes which were frequently misclassified: calculus as gingivitis, and tooth discoloration as caries. Tooth discoloration being misclassified as caries indicates that color-based distinctions were not effectively captured by HOG features. These issues suggest that HOG features alone are insufficient to capture subtle differences in oral disease images, particularly in cases where textures and structures overlap across different conditions.

## 5 Primary Model

The CNN discussed below performed better than the baseline model by learning features straight from the processed image data, capturing more complex structures, and eliminating the need for manual feature extraction.

### 5.1 CNN Architecture

The team refined the model architecture throughout the project to achieve the best performing model. These efforts included adjusting the number of convolutional layers and filters, experimenting with and without dropout (and the percent of neurons dropped), adding and removing batch normalization, and altering the number of neurons in the fully connected network. The team identified that the best architecture was achieved with a deeper network that had more convolutional filters in the final layer and incorporated batch normalization. A deeper network enabled the model to detect fine details in the dental images. Given the similar characteristics between several of the diseases (i.e. gum inflammation with both calculus and gingivitis), adding complexity to the model was critical for accurate classification. Other architectures with more convolutional layers but less filters and no batch normalization were too deep and overfit to the data, likely capturing more noise than meaningful characteristics of the oral conditions. In contrast, the shallower networks significantly underfit the data and failed to detect distinguishing features of the oral conditions.

As per the initial architectural plan, the best performing CNN has five convolutional layers and two fully connected layers. The number of filters in each convolutional layer progressively increases as the network deepens and each plays a role in extracting visual features, including texture, color, and shape, to help differentiate between the various oral diseases. However, to reduce computational cost, manage spatial complexity, and eliminate redundant information while retaining important features, max pooling was implemented between the convolutional layers with a kernel size of two and stride length of two. Further, to improve training stability and time to convergence, batch normalization has been applied to all of the convolutional layers prior to applying the ReLU activation function which is critical to introduce non-linearity into the model. After pooling occurs on the fifth convolutional layer, the feature map is flattened to a single vector for classification. The first fully connected layer maps the features to 256 neurons. Next, the ReLU activation function is applied to the first fully connected layer before implementing 50% dropout to prevent overfitting and allow the network to generalize better to unseen cases. The second fully connected layer maps the remaining neurons to the six classes. A detailed visualization of the CNN's complexity can be seen in Figure 4 below. The cross-entropy loss function was used to measure the difference between predicted and actual labels. Given that the cross-entropy loss function was used, the softmax function is applied implicitly to generate probability distributions over the given oral disease classes.

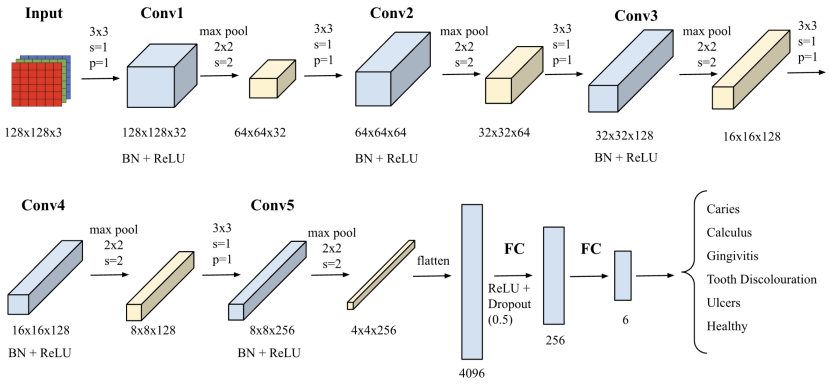



Figure 4: Finalized CNN architecture

In order to continuously improve the model, the team conducted hyperparameter tuning using random search and trained the model with various combinations of learning rates, number of epochs, batch sizes, and optimizers. The highest performing trained model was obtained using a learning rate of 0.0008, a batch size of 32, using the Adam optimizer, and was validated across 15 epochs. This model was selected based on validation loss (as opposed to error) since it incorporates a level of confidence into the prediction performance.

### 5.2 Quantitative Results

After obtaining the best performing model architecture and selecting the hyperparameters that achieved the lowest validation loss, the team analyzed the quantitative performance of the model. During training, the error and loss curves across the 15 epochs on both the training and validation sets were recorded and can be seen in Figure 5 below. The twelfth epoch had the lowest validation loss and scored a validation error of 0.27 and a validation loss of 0.71.

After training the model, the team measured the performance of the model on the overall test error and loss, as well as other evaluation metrics including accuracy, precision, recall, and F1-score at the class level to measure model performance. Similar to the baseline model, the team has placed emphasis on the macro average of the F1-score so that all classes are treated equally despite the class imbalance. The current model achieved a macro average F1-score of 0.65. The comprehensive test classification report for the CNN can be seen in Figure 6 below.

From these results, it can be noted that the model is very successful at classifying gingivitis (class 2), given it had the highest recall score of 0.93, meaning it was the least likely class to be misclassified as a different condition (fewest false negatives). The healthy class (class 3) had the best performance

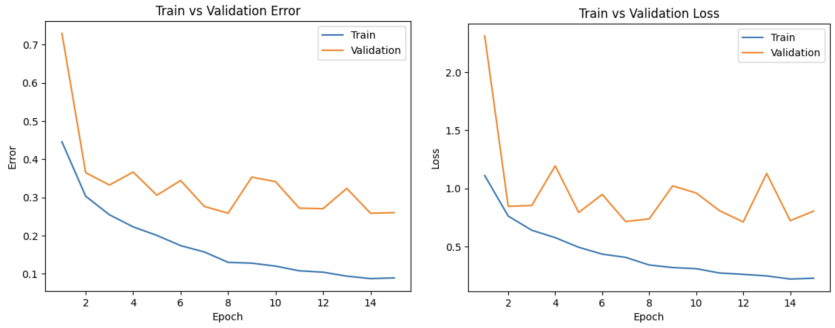
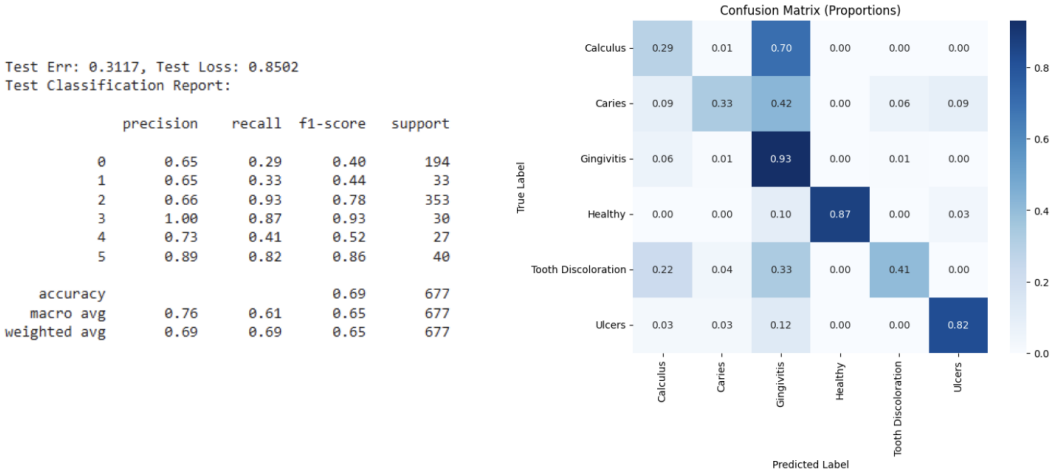



Figure 5: Train vs. validation error and loss for the primary model



Test Err: 0.3117, Test Loss: 0.8502
Test Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.65 | 0.29 | 0.40 | 194 |
| 1 | 0.65 | 0.33 | 0.44 | 33 |
| 2 | 0.66 | 0.93 | 0.78 | 353 |
| 3 | 1.00 | 0.87 | 0.93 | 30 |
| 4 | 0.73 | 0.41 | 0.52 | 27 |
| 5 | 0.89 | 0.82 | 0.86 | 40 |
| accuracy | | | 0.69 | 677 |
| macro avg | 0.76 | 0.61 | 0.65 | 677 |
| weighted avg | 0.69 | 0.69 | 0.65 | 677 |



Figure 6: Primary model classification report and confusion matrix

overall scoring the highest F1 which can be attributed to its perfect precision, meaning every time the model predicted a healthy case, it was correct (no false positives). On the other hand, calculus (class 0), caries (class 1), and tooth discoloration (class 4) performed much worse, illustrated by much lower recall scores that are reflected in poor overall F1 performance. This means that several of their images were misclassified as other conditions, likely due to overlapping visual features that made it difficult for the model to distinguish the disease.

### 5.3 Qualitative Results

In addition to the quantitative evaluation, the team used a combination of confusion matrices and viewing individual samples from the test set predictions to collect qualitative results. The normalized confusion matrix for the final model can be seen in Figure 6 above. The team focused on classes with lower recall and F1 scores, particularly calculus, caries, and tooth discolouration.

Compared to the progress report model, this showed cleared differentiation between minority classes like ulcers and healthy teeth. Visually identifying correctly classified examples, has shown that these classes typically have more distinct visual cues, such as the visibility of ulcers on the tongue, or unstained teeth. The model was able to effectively learn these features using the processed images as a result of the improved denoising, contrast enhancement, and colour normalization techniques. However, the model also demonstrated similar (though slightly improved) persistent challenges. Misclassifications primarily were among the classes with visually similar features including calculus, caries, and tooth discolouration, which were frequently misclassified as gingivitis. This aligns with the confusion matrix findings and suggests that despite the hybrid data processing pipeline and

the deeper CNN architecture, the model still struggles to distinguish between subtle differences in gum texture, discolouration, and plaque buildup within these classes. Additionally, the model often overpredicted gingivitis, which was the dominant class in the dataset, and is reflective of the higher recall and lower precision. This indicates that this bias was not fully prevented through class balancing using augmentation.

Overall, the qualitative results show that the improvements made to the data processing and architecture have enabled the model to generalize better for the visually distinct classes. However, despite seeing a slight increase in true positives for calculus, caries, and gingivitis, the model still had trouble in differentiating between these classes as a result of their similar appearances and locations.

## 6 Testing on Unseen Data

This section outlines the methodology and results of testing the CNN on entirely new data sources and evaluating its ability to generalize to realistic, real-world data.

### 6.1 Data Collection

In order to ensure accuracy when testing model performance, it is important to use never seen before data within the testing set. The team began by searching for datasets on Kaggle containing all five dental diseases this project focuses on, but was unsuccessful. Many of the available datasets on the site from similar projects repurposed the same set of images already used in model training, limiting the team's ability to obtain data unseen by the model. The search was then expanded to other online platforms, academic articles, and research papers, some of which referenced public datasets. The team found several smaller datasets on Roboflow that contained dental images from sources different from those the training data was obtained from. There were many different projects that, separately, contained data sets of all five dental diseases the team required, as well as healthy teeth. As a result, the unseen test data was sourced from these different datasets and combined to include images from all classes required for the model. One source provided a dataset for calculus images Trial (2024), another for gingivitis and caries (Vishal, 2025), one for ulcers and tooth discolouration Temini (2025)and lastly another for healthy (Sung, 2024). Although there was no single dataset found that fully matched the project requirements, merging these sources allowed the team to construct a more comprehensive and diverse test set.

Once all the images were retrieved, the team carefully examined them to ensure all the images in each class were unique and did not contain any images that were already augmented or preprocessed. If duplicates or preprocessed images were found, they removed to maintain the integrity of the testing set and the results when the model is run on this dataset. In addition to images sourced from these new datasets, the team included images from several academic articles (see Beaumont (2017), Cai (2024), Clark (2019)) that provided examples of the dental diseases the project focuses on. Team members also sourced a handful of testing images from family and friends, taking the pictures themselves and adding them to the appropriate classes depending on any dental diseases the individual was professionally diagnosed with, if any. In total, the team compiled a testing data set with approximately 140 - 180 images per class, providing a diverse and well distributed testing set. This approach ensured that the model was evaluated on truly unseen data, making the results a reliable measure of its performance in real-world scenarios.

### 6.2 Model Performance

Once the testing data had been collected and the pre-processed images removed, the team ran the CNN on the unseen data to understand how well it would perform on the new, unbiased material. A summary of all the scores for the unseen data can be found in Figure 7 below, in a general classification report. The model achieved an overall test error of around 0.49 and a test loss of around 2.64. In comparison to the validation error and loss from section 5.2, which were around 0.27 and 0.71 respectively, the test error and loss increased suggesting that the model did not generalize as well to the new data. This reduced performance is to be expected as models tend to perform worse on unseen data. The classification report indicates that ulcers (class 5) achieved the highest F1-score, 0.71, with a precision of 0.76 and recall of 0.66, indicating that the model was able to reliably classify this dental disease. This aligns with the results the team obtained in the training

results, where ulcers was also one of the more accurate classes. This is due the unique and easily identifiable visual characteristics of the disease. Similarly, gingivitis (class 2) had a high recall, 0.84, signifying that the model was able to correctly identify the majority of gingivitis cases. Its precision on the other hand dropped from 0.66 in training to 0.38, creating an increase in false positives. This suggests that the model was too liberal in labeling images as gingivitis. In contrast, the healthy class (class 3) experienced a significant decline in its recall, dropping from 0.93 to 0.18 suggesting that the model heavily misclassified healthy teeth as diseased when evaluated on new data. The macro average F1-score 0.49 is lower than the 0.65 achieved during training, further demonstrating the challenge the model had with generalizing to unseen data.

Test Err: 0.4897, Test Loss: 2.6407
Test Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.48 | 0.54 | 0.51 | 144 |
| 1 | 0.49 | 0.46 | 0.47 | 148 |
| 2 | 0.38 | 0.84 | 0.52 | 155 |
| 3 | 0.81 | 0.18 | 0.30 | 138 |
| 4 | 0.58 | 0.35 | 0.43 | 172 |
| 5 | 0.76 | 0.66 | 0.71 | 164 |
| accuracy | | | 0.51 | 921 |
| macro avg | 0.58 | 0.51 | 0.49 | 921 |
| weighted avg | 0.58 | 0.51 | 0.50 | 921 |

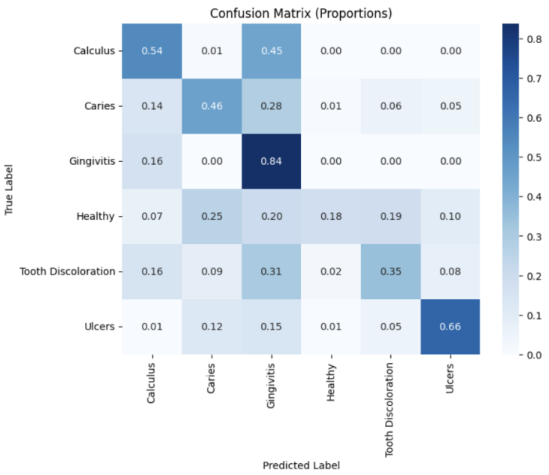



Figure 7: Classification report and confusion matrix of unseen data using the primary model

The confusion matrix in Figure 7 provides additional insights into the model's classification patterns, and its tendency to misclassify certain dental diseases. As in the training phase, gingivitis (class 2) remained the dominant predicted class, with images from calculus (class 0) and caries (class 1), both who experienced high misclassification rates, being labeled as gingivitis. It can also be seen in the confusion matrix that tooth discoloration (class 4) was only correctly classified 35% of the time similar to its struggles with low recall in training. Healthy teeth (class 3), which had perfect precision during training and experienced a severe drop in recall when testing on unseen data, suggests that the model had difficulty identifying healthy teeth when provided with new examples, instead assigning them dental disease labels perhaps due to the lack differentiation between normal teeth and disease like gingivitis, which do not always stand out visually. While the model improved in distinguishing ulcers and gingivitis, it still exhibited bias toward over-predicting gingivitis, and struggled with visually similar conditions like calculus, caries, tooth discoloration, and sometimes healthy teeth.

## 7 Discussion

The model's performance on the unseen test data highlights both its strengths and its limitations, with some of the obtained results suggesting the existence of biases in how the model learned from the training data. One example of this is the high recall but low precision seen in the testing data results for gingivitis (class 2). This could be a result of the model being exposed to a greater variety of gingivitis cases in the training set, making it more confident when predicting this condition, and even leading it to miss classify other diseases as gingivitis. If the gingivitis cases in the test set had been visually distinct from those in the training set, the model would have generalized much worse to the unseen data causing the precision score for this class to drop. In the results for the testing data the team also identified that there was a significant decline in the recall for healthy teeth (class 3). This result can be attributed to the model becoming too reliant on features associated with diseases, possibly because the training data contained many more disease images than healthy ones. Additionally, if the testing set had a different distribution of healthy teeth images than the training set, or if lighting conditions for images varied more than in the training set, this could also have

contributed to the sharp decline in recall for this class. This led the team to conclude that the model focused a great deal on the patterns in the training data, and the features specific to the images of each class, rather than learning more generalizable features to be able to distinguish between teeth with dental diseases and healthy teeth.

Differences in class distributions or image characteristics between these two sets could have contributed to low generalization on the part of the model. For example if the test set contained more challenging images, like images of diseases under different lighting conditions, lower resolution, various rotations, or atypical presentations of the diseases, the model would have a hard time applying white learned during training as it was not taught to classify these more difficult and unique images. This can be seen in the increase in both test loss and test error which shows that the model did not generalize well to the testing data.The high rates of misclassification of calculus (class 0) and caries (class 1) as gingivitis led the team to conclude that the model may have learned some shortcuts during training and relied on superficial visual similarities rather than relating images to classes based on deep feature representations. To counteract this, the training process could be refined by incorporating a more diverse and representative set of images in the training dataset. Techniques like domain adaptation or adversarial training would allow the model to learn deeper features and not just rely on visual similarities between images. Domain adaptation involves training the model to recognize differences between datasets by adjusting what it learns about each class to better align with the training and testing data distributions, resulting in less bias towards particular datasets (see Farahani (2021)), while adversarial training deliberately introduces small amounts of noise into training images, making the model more resistant to minor variations in the data (Tao Bai, 2021). Additionally the model's tendency to overpredict certain classes indicates overfitting, suggesting that it memorized the training data, leading to it failing to extract the important features that would allow it to adapt to unseen data. These shortfalls highlight the importance of carefully curating training and testing datasets to make sure that models are able to learn, and subsequently generalized to new data, by understanding meaningful patterns rather than relying on biases in the training sets. The limited scope of datasets the team had to work with, as well as the nature of dental diseases themselves posed challenges and difficulties throughout the project. Further details on project difficulties can be found in Appendix B.

## 8 Ethical Considerations

In the application of diagnosing health conditions, there are various ethical factors to consider in terms of the data that is being used to train the model, as well as risks and limitations within the results. It is important that the data being used to train the model is diverse and all demographic populations are equally represented to avoid dataset bias. If the training images are not diverse, the model runs the risk of performing poorly on underrepresented populations and if used to formulate a diagnosis, this could translate into healthcare quality discrepancies. To mitigate this, data should be sourced from across different age ranges, race, socio-economic levels and other factors to ensure an even distribution of data for the model to learn from. This can be done through searching online for pre-existing datasets or manually collecting dental images using the correct ethical practices. Furthermore, it is an ethical responsibility of the team to ensure that that data being used is available for public use. Medical imaging is personal and private information that should be provided with consent to use in advance of using it in the training of a model, and should be anonymous to avoid re-identification. The oral diseases dataset being used by the team is publicly available, anonymous, diverse, authentic, and is sourced from hospitals and reputable dental websites (see Sajid (2024).

In terms of other considerations, another use of the system that could give rise to ethical issues is misdiagnosis. The diseases being considered by the model are serious and have impacts on the user's mental and physical health and should be treated with the utmost responsibility as a result. To mitigate this risk, users and impacted patients of the model must be informed of the risk of being misdiagnosed and should not overly rely on the model when taking a course of action. For real-world deployment, the model should come with a disclaimer about its limitations and the advice to always contact a certified medical professional to verify any serious results. The results should not replace a professional diagnosis and should strictly act as an assistive tool. In the case of false negative predictions, missed diagnoses may delay treatment. On the other hand, unnecessary anxiety and procedures could result from taking action on false positives.

## REFERENCES

Yassir Edrees Almalki, Amsa Imam Din, Muhammad Ramzan, Muhammad Irfan, Khalid Mahmood Aamir, Abdullah Almalki, Saud Alotaibi, Ghada Alaglan, Hassan A Alshamrani, and Saifur Rahman. Deep learning models for classification of dental diseases using orthopantomography x-ray opg images, Sep 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9572157/.

Chesterman J. Kellett M. et al. Beaumont, J. Gingival overgrowth: Part 1: aetiology and clinical diagnosis, Jan 2017. URL https://www.nature.com/articles/sj.bdj.2017.71#citeas.

T. Bunmi. The burden of diagnostic error in dentistry: A study on periodontal disease misclassification, Jul 2024. URL https://www.sciencedirect.com/science/article/pii/S0300571224003907#fig0001.

Cheng Y. Liu Y. Song Y. Zhang N. Cai, D. Oral screening of dental calculus, gingivitis and dental caries through segmentation on intraoral photographic images using deep learning, Oct 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11515110/.

Shweta Dixit Chaudhary, Priyanka Paygude, and Preetam Shah. Teeth or dental image dataset, Apr 2024. URL https://data.mendeley.com/datasets/6zsnhrds9t/1.

Cohen D. Fitzpatrick S. Clark, A. Ulcerated lesions of the oral mucosa: Clinical and histologic review, Mar 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6405793/.

Cleveland Clinic. Tooth loss, Aug 2024. URL https://my.clevelandclinic.org/health/diseases/tooth-loss.

Voghoei S. Rasheed K. Arabnia H.R. Farahani, A. A brief review of domain adaptation, Oct 2021. URL https://link.springer.com/chapter/10.1007/978-3-030-71704-9_65.

Statistics Canada Government of Canada. Self-reported oral health problems in the canadian population living in the provinces, november 2023 to march 2024, Oct 2024. URL https://www150.statcan.gc.ca/n1/daily-quotidien/241023/dq241023b-eng.htm.

Pandia Rajan Jeyaraj and Edward Rajan Samuel Nadar. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm, Jan 2019. URL https://pubmed.ncbi.nlm.nih.gov/30603908/.

Junegyu Kang, Van Nhat Thang Le, Dae-Woo Lee, and Sungchan Kim. Diagnosing oral and maxillofacial diseases using deep learning, Jan 2024. URL https://www.nature.com/articles/s41598-024-52929-0.

Javed Rashid, Bilal Shabbir Qaisar, Muhammad Faheem, Arslan Akram, Riaz ul Amin, and Muhammad Hamid. Mouth and oral disease classification using inceptionresnetv2 method - multimedia tools and applications, Sep 2023. URL https://link.springer.com/article/10.1007/s11042-023-16776-x.

Salman Sajid. Oral diseases, Apr 2024. URL https://www.kaggle.com/datasets/salmansajid05/oral-diseases/data.

Uma Subbiah and Padmavathi S. Analysis of deep learning architecture for non-uniformly illuminated images, Jun 2020. URL https://ieeexplore.ieee.org/document/9112434.

Sung. Cavity check computer vision project, Apr 2024. URL https://universe.roboflow.com/sunglassdetection/cavity-check.

Jun Zhao Bihan Wen Qian Wang Tao Bai, Jinqi Luo. Recent advances in adversarial training for adversarial robustness, Apr 2021. URL https://arxiv.org/pdf/2102.01356.

Temini. Tooth diseases computer vision project, Apr 2025. URL https://universe.roboflow.com/teminiproj/tooth-diseases-0ocji.

Trial. Image recognition of dental computer vision project, Jun 2024. URL https://universe.roboflow.com/trial-07fpi/image-recognition-of-dental.

Vishal. Oa computer vision project, Apr 2025. URL https://universe.roboflow.com/vishal-lqlqm/oa-thu6d.

Yanshan Xiong, Hongyuan Zhang, Shiyong Zhou, Minhua Lu, Jiahui Huang, Qiangtai Huang, Bingsheng Huang, and Jiangfeng Ding. Simultaneous detection of dental caries and fissure sealant in intraoral photos by deep learning: A pilot study - bmc oral health, May 2024. URL https://bmcoralhealth.biomedcentral.com/articles/10.1186/s12903-024-04254-1.

## A Data Processing Examples


Figure 8: Resulting images after noise reduction


Figure 9: Resulting images after contrast mapping

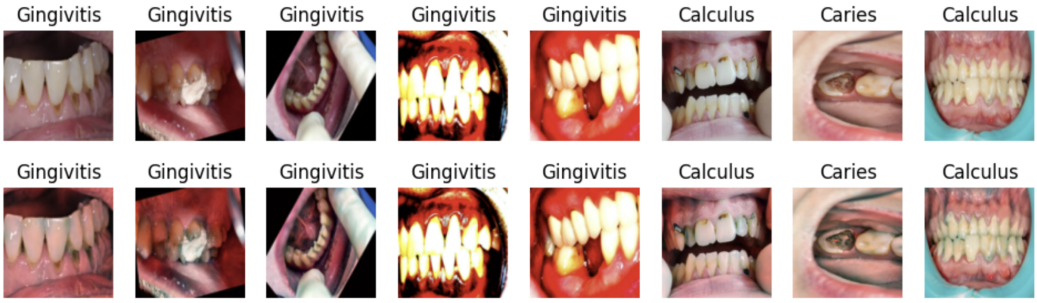



Figure 10: Resulting images after red-channel enhancement


Figure 11: Resulting images after standardization

## B Project Difficulty and Quality

Overall, there were many technical challenges associated with this project that required intense research, fine tuning and data processing methods far beyond the scope of this course. One of the biggest difficulties in this project was the data itself, since the dataset was real-life images the scenes often varied in brightness, angle, objects in the scene (such as the lips and tongue) and viewpoint. This extreme non-uniformity in the images presented a challenge in the data processing step as the team tried to make the data more uniform for the model to better predict and extract high level features from. To address this, the team had to experiment and combine multiple data processing techniques such as normalization, contrast mapping and noise reduction filters. On top of data processing, data augmentation was also used to address the class imbalance in the dataset. In addition, methods beyond the scope of the course were also used such as red channel enhancement and color jitter to further process the images and ensure better model prediction.

Furthermore, the nature of dental diseases themselves presented another major level of difficulty to the project. Dental diseases often present with overlapping symptoms given the specific area of the body the diseases are limited to. For example, swollen gums can be a sign of gingivitis or caries depending on the other present symptoms. As a result, a complex understanding of the features and their relation to each other in a dental image are required to make an accurate diagnosis. This adds a layer of difficulty to the project as it presents the need for the model to not only identify complex features from a non-uniform scene but also the relationships between the present features. To combat this, the team performed multiple tedious rounds of fine tuning the CNN model for numbers of layers, dropout implementation and other factors to create a complex model capable of understanding the dental image diagnosis. As a result, given the multiple layers of difficulty presented by both the nature of the dental diseases and the dataset, the final model performed considerably well. Overall, dental disease classification proves to be a tough task for even the trained professionals with a study in periodontal diseases finding that 32% of observed periodontal diseases were misclassified (see Bunmi (2024)) by trained, practicing dentists. Thus, given the limited dataset, the scope of the courses and the overall timeline, the performance of the model is quite satisfactory.