RMIT UNIVERSITY

Image Classification of Cancer (Project 1)

Assignment 2

COSC2793 Computational Machine Learning

Group 11

Wing Hang Chan - S3939713

Introduction

Colorectal cancer ranked as the third most common cancer worldwide, as reported by the World Cancer Research Fund International [1]. The incidence rate of new colorectal cancer cases in relation to all cancer cases is one-tenth. Computed tomographic (CT) colonography is a viable method for detecting colorectal cancer [2]. With advancements in technology, the utilization of convolutional layers and downsampling layers in neural networks, first introduced by Kunihiko Fukushima in 1980 [3], has gained significant popularity. Convolutional neural networks (CNNs) applied to medical images have emerged as valuable tools for training doctors and conducting medical research. In this report, a CNN serves as the base and oversampling model for classifying cell types and identifying cancerous features among more than 9,000 tiny images. Various techniques to enhance the model's performance are analyzed and discussed.

Background

The study consists of two tasks: multi-class classification of cell types and binary classification of cancerous cells. The dataset comprises 20,280 images, all of which have identical size and color channels. Korsuk [4] identifies more than four classes, with the three primary types being epithelial, fibroblast, and inflammatory cells. Additional cell types, such as adipocytes, endothelial cells, and dead cells, are categorized as "others".

A fibroblast [5] is a cell that plays a role in the formation of connective tissue and can be associated with cancer [6]. Inflammatory cells, on the other hand, contribute to the healing of injured tissues [7]. Epithelial cells [8] function as protective barriers on both the interior and exterior surfaces of the body. According to Canadian Cancer Society [9], any cell in the body has the potential to become cancerous.

To address these tasks, convolutional neural networks (CNNs) with different models are employed to classify cell types and determine whether a cell is cancerous.

Exploratory Data Analysis

The dataset is divided into two parts: the main dataset, which consists of 9,896 records with a complete set of attributes, and the extra dataset, which includes 10,384 records but lacks the cell type name attribute. The dataset exhibits an imbalance in both cell types and cancerous labels (Figure 1 & Figure 2). Epithelial cells represent the majority class, while “Others” cell types are the minority class. The distribution of the cancerous label is also uneven. Notably, all cancer cells belong to the epithelial cell type, and there are no other cell types classified as cancerous. These findings present a contrast with the information provided by Canadian Cancer Society.

Visual differentiation of cell types, especially other cells, can be challenging due to their diverse patterns (Figure 3). Since the task at hand involves image classification, the input consists of images, and the prediction corresponds to a label. Attributes such as the instance ID and patient information are not considered in the model.

Methodology

Baseline Model Design

The baseline model design (Figure 4) for both multiclass and binary classification tasks is similar, as they share the same input. The model consists of five 2D convolutional layers organized into three blocks, followed by two fully connected layers with 512 neurons, and an output layer with either one or four outputs. In the first block, a 3x3 size convolutional layer with 32 kernels is employed to extract low-level features. This layer is then connected to a 3x3 max pooling layer for down-sampling. The output of the first block is a collection of 9x9x32 feature maps. The second block consists of two 3x3 convolutional layers with 64 kernels each, and a 3x3 max pooling layer. These layers extract middle-level features from the low-level ones. The third block involves two 3x3 convolutional layers with 128 kernels each, along with a max pooling layer. These layers derive high-level features.

Once the features have been extracted, they are flattened and connected to two dense layers with 512 perceptrons each. All the preceding feature extraction and feedforward layers utilize the rectified linear unit (ReLU) as the activation function.

The final layer and output layer differ between the two models. For the cell type classification task, there are four possible labels, and a softmax activation function is employed to obtain probabilities for each class. In contrast, for the cancer classification task, there is only one unit with a sigmoid activation function, providing a probability for determining whether the cell is cancerous.

Analysis

Considering the relatively small number of input images, the dataset is divided into a 70:15:15 ratio for training, validation, and testing, respectively. This results in approximately 7,000 training samples, which are further divided into batches of 32 for processing efficiency. It has been discussed that using large batch sizes can negatively impact accuracy and generalization performance [10]. Ayush [11] suggests that a batch size of 32 yields the lowest error rate for the given dataset.

CNNs are known to perform well on balanced datasets, but they can struggle with imbalanced datasets [12]. To address this, class weights are calculated using the image data generator, which employs one-hot encoding for labels. It is important to note that these labels may differ from the labels in the main dataset. The use of class weights adjusts the loss function accordingly [13].

For optimization, the stochastic gradient descent method with the Adam optimizer is used. Adam offers the advantages of consuming less memory and faster calculations due to its adaptive moment estimation approach [14].

Convolutional layers are employed to extract features by applying different kernels. The choice of a 3x3 kernel size is based on the need for a central pixel to calculate the output. Padding is added to account for the tiny image size, and considering the margin pixels can provide valuable information for detecting cancer. The pooling layer size is set to 3x3, which ensures that the image size of 27x27 is evenly divided without missing any pixels.

The rectified linear unit (ReLU) activation function is chosen as it effectively propagates the gradient. Although ReLU can suffer from the "dying ReLU" problem [15], which causes some neurons to become inactive, this issue can be ignored in image classification since images do not have negative pixel values. Any negative values arising from the convolutional kernels will be effectively canceled out.

To prevent overfitting, early stopping is implemented by monitoring the categorical or binary accuracy. If the accuracy does not improve over the last 10 epochs, training is stopped to avoid further overfitting.

Evaluation

Cell Type Classification

The multiclass classification model is showing signs of overfitting, despite the implementation of early stopping. The validation accuracy remains around 70-75%, while the training accuracy continues to increase to over 80% (Figure 5). The categorical accuracy and weighted F1 score on the testing set are both 73.74.75% and 73.56% respectively.

Examining the confusion matrix (Figure 7), it is evident that the model performs well in identifying epithelial cells which is the heavy weighted class but struggles with classifying “others” cell types which is the least weighted class. Specifically, the "others" cell type is often misclassified as inflammatory. Out of a total of 1485 images in the testing set, the model correctly predicts 1095 images.

Overall, these results indicate that the model's performance is decent but could be improved, especially in correctly identifying the "other" cell type.

Cancerous Classification

The binary classification model for cancer identification is also showing signs of overfitting. It achieves a high accuracy (Figure 6) and F1 score of 89.75% (Figure 8). Out of the total predictions, 1798 non-cancerous cells are correctly identified as true negatives, and 931 cancerous cells are correctly identified as true positives. However, there are 173 false positives and 140 false negatives.

From these results, it can be inferred that both classifications are potentially affected by the imbalance in the dataset. Furthermore, both models exhibit indications of overfitting. To address these issues, an additional model is built to overcome these challenges.

Oversampling Model Design and Analysis

To address the issue of overfitting, several techniques are employed in the baseline model, including regularization, dropout, and data augmentation, in addition to early stopping. L2 regularization is utilized since there are no outliers in the image data, and feature selection is not necessary. The regularization term is penalized with a coefficient of 0.0001. Dropout is implemented after the two dense layers, randomly dropping out 20% of the units to reduce the complexity of the network.

Data augmentation is employed to artificially enhance the dataset by applying transformations to each data point. Zoom and rotation transformations are chosen, ensuring that the center of the image remains intact. This is important as the cell of interest is typically located at the center, making it a prominent feature for recognition. Brightness and channel shift transformations are also utilized to alter the color tone of the images. These transformations can be particularly helpful in classifying dark or unclear images.

To address the class imbalance issue, oversampling is performed. This approach is preferred over downsampling since the dataset is not large, and oversampling, coupled with data augmentation, can lead to better generalization. The model generates 50,000 additional images for each class, resulting in a total of 200,000 images for cell type classification.

Given the oversampling technique, class weights are no longer needed. The equal proportion of data across classes ensures a smoother gradient signal during training. While there may still be a slight imbalance in the data within each batch, the stochastic gradient descent (SGD) optimization algorithm can effectively handle this scenario.

Evaluation

Cell Type Classification

The issue of overfitting has been effectively addressed in the improved model compared to the baseline model (Figure 10). The validation accuracy is close to the training accuracy, indicating better generalization. However, there might be a slight underfitting issue due to data augmentation. The model generates new images during each epoch of the training process, which helps in enhancing the generalization of the learned features. It is important to note that data augmentation is not applied to the validation and testing processes.

Regularization plays a crucial role in preventing large losses in the model by applying a penalty. Although it may not be easy to observe the impact directly in a neural network model with numerous weights and features, it aids in controlling the complexity and promoting better generalization. Dropout is another technique employed in the model, where a certain percentage of units are randomly dropped out, helping to reduce complexity and prevent overfitting.

Overall, this improved model demonstrates better training performance compared to the original one by incorporating these three techniques: regularization, dropout, and data augmentation. It is worth mentioning that the loss and accuracy between the two models cannot be directly compared since they use different optimizers. However, we can evaluate the predictive performance through the confusion matrix. The F1 score has improved to 81.40% compared to the baseline model's 73.56%. The number of correct predictions for the "Other" cell type has significantly increased from 81 to 143. Although there are still some incorrect predictions for the "Other" cell type, the situation has shown improvement.

Cancerous Classification

The classification model for cancerous cells is like the cell type classification model, as they both use a similar architecture. The F1 score for cancerous classification has improved from 89.75% to 91.41%, which is a slight improvement. However, the baseline model already achieved a good performance, so a 2% increase is considered a significant improvement.

Discussion

Considering the imbalance in the dataset, both the techniques of class weight and oversampling have had significant effects. However, it is important to note that choosing the right optimizer when applying class weight is crucial, as selecting the wrong optimizer can yield poor results [13].

Oversampling with data augmentation has proven to be beneficial, as it generates new images based on the original dataset. However, it is essential to carefully choose the augmentation method and determine the appropriate range of augmentation. In this particular case, coloring has shown positive effects, while shifting is not recommended as it can lead to the loss of the cell's center, which is a critical feature for recognition.

Although oversampling with augmentation helps improve the model's generalization, it does not create entirely new features. If there are unique features present in the test set that are not present in the training data, the model may still struggle to learn and recognize them. Additionally, the "Other" cell type poses a challenge to the model due to its small size and the presence of multiple other cell types within it. This makes it difficult for the model to generalize the features of all these sub-types within the "Other" cell category. The imbalance within the "Other" cell type further complicates the dataset. A potential approach to address this issue is to classify the three major cell types and set a threshold for identifying the "Other" cells. However, finding a suitable threshold can be challenging.

Usage of Extra Dataset

While oversampling with data augmentation is a good approach to address the issue of imbalance, it can still result in an exhausted dataset. In such cases, there are several ways to make use of the extra dataset. One suggested approach by Wouter [16] is unsupervised learning, where a machine learns to classify images without relying on labels. This involves a combination of self-supervised learning (SSL), clustering, and self-labeling, known as Semantic Clustering by Adopting Nearest neighbors (SCAN).

SSL, a form of representational learning [17], involves comparing an image with its transformed version using a neural network to extract features. This helps in capturing important features while discarding low-level details. The next step is to apply k-nearest neighbor clustering to group similar images together. Finally, self-labeling is performed, which uses more confident neighbors to train the classifier.

SCAN has achieved a classification score of 81.9% for 50 classes from ImageNet, slightly lower than the results obtained through supervised learning (86.5%). Furthermore, SCAN may encounter challenges with over-clustering. In the case of the "Other" cell category, which represents various other cells mentioned in the background, SCAN may struggle to accurately group them together. This can result in lower accuracy for sub-types within the "Other" category that have fewer images.

Unsupervised learning focuses on grouping images without creating explicit labels. However, it does not provide labels for the images. To address this limitation, Zhu [18] proposed semi-supervised classification, which combines labeled data with unlabeled clustered data to build better classifiers. Practical example [19] shows that this approach works better than supervised baseline model and generalize better.

Conclusion

CNN models typically have a large number of parameters that require careful tuning to achieve optimal performance. To aid in this process, there are tuners available that can assist beginners in finding suitable parameter values. Furthermore, research papers, such as Jorai mentioned [20], propose using reinforcement learning for hyperparameter tuning. In addition, there are papers that compare the results of different CNN architectures, such as AlexNet, CiFAR-VFF, and WRN, using the same dataset. Basha [21] reports an accuracy of 84.94% for their model, while the other architectures achieved accuracies of 82.15%, 84.25%, and 81.55%, respectively. The reported accuracy of 81.40% in this study is comparable to the results of other architectures. However, it is important to emphasize that there is no singular "most suitable" model that applies universally. The choice of model depends on the specific task and the characteristics of the data being used.

Appendices

[1] World Cancer Research Fund International (2022, March. 23), “Worldwide cancer data”, [Online]. Available: https://www.wcrf.org/cancer-trends/worldwide-cancer-data/

[2] National Cancer Institute (2021, August. 2), “Screening Tests to Detect Colorectal Cancer and Polyps”, [Online]. Available: https://www.cancer.gov/types/colorectal/screening-fact-sheet

[3] Fukushima, Kunihiko (1980). "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position" (PDF). Biological Cybernetics.

[4] Korsuk Sirinukunwattana, (2016, May. 5), “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images”, IEEE Transactions on Medical Imaging, Vol. 35, No. 5

[5] Ellen Sidransky, M.D. (2023, May. 11), “Talking Glossary of Genomic and Genetic Terms - Fibroblast”, NIH, [Online]. Available: https://www.genome.gov/genetics-glossary/Fibroblast

[6] Cirri, Paolo; Chiarugi, Paola (2011, May. 12). "Cancer associated fibroblasts: the dark side of the coin". American Journal of Cancer Research. 1 (4): 482–497.

[7] Thomas C. King, “Elsevier's Integrated Pathology”, 1st ed, Elsevier Inc. 2007

[8] National Library of Medicine (2022, August. 3), “Epithelial Cells in Urine”, [Online]. Available: https://medlineplus.gov/lab-tests/epithelial-cells-in-urine

[9] Canadian Cancer Society, “How cancer starts, grows and spreads”, [Online]. Available: https://cancer.ca/en/cancer-information/what-is-cancer/how-cancer-starts-grows-and-spreads

[10] Nitish Shirish Keskar (2016, September. 15), “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima”, [Online]. Available: https://arxiv.org/abs/1609.04836

[11] Ayush Thakur (2022, March. 24), “What's the Optimal Batch Size to Train a Neural Network?”, [Online]. Available: https://wandb.ai/ayush-thakur/dl-question-bank/reports/What-s-the-Optimal-Batch-Size-to-Train-a-Neural-Network---VmlldzoyMDkyNDU

[12] Iren Valova (2020), “Optimization of Convolutional Neural Networks for Imbalanced Set Classification”, 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

[13] TensorFlow (2022, December. 15), “Tutorials - Classification on imbalanced data” [Online]. Available: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data#class_weights

[14] Diederik P. Kingma (2014, December. 22), “Adam: A Method for Stochastic Optimization”, 3rd International Conf. for Learning Representations, San Diego, 2015, [Online]. Available: https://arxiv.org/abs/1412.6980

[15] Lu Lu (2019, March. 15), “Dying ReLU and Initialization: Theory and Numerical Examples”, [Online]. Available: https://arxiv.org/abs/1903.06733

[16] Wouter Van Gansbeke (2020, May. 25), “SCAN: Learning to Classify Images without Labels”

[17] Carl Doersch (2015), “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015, pp.1422-1430

[18] Xiaojin Zhu (2005, September.), “Semi-Supervised Learning Literature Survey”, [Online]. Available: https://minds.wisconsin.edu/handle/1793/60444

[19] Andras Beres (2021, April. 24), “Semi-supervised image classification using contrastive pretraining with SimCLR”, Keras, [Online]. Available: https://keras.io/examples/vision/semisupervised_simclr/

[20] Jorai Rijsdijk (2021, July. 09), “Reinforcement Learning for Hyperparameter Tuning in Deep Learning-based Side-channel Analysis”, [Online]. Available: https://tches.iacr.org/index.php/TCHES/article/view/8989

[21] S.H. Shabbeer Basha (2019, October. 6), “Impact of fully connected layers on performance of convolutional neural networks for image classification”,[Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219313803

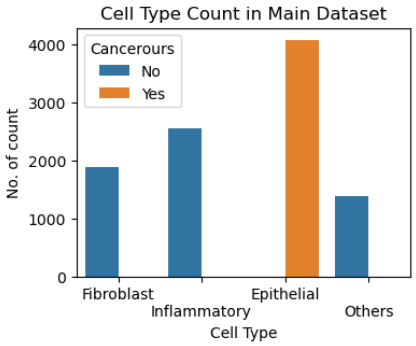

Figure 1 Cell Type Count in Main Dataset

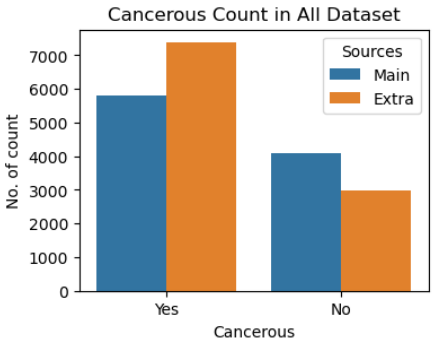

Figure 2 Cancerous Count in All Dataset

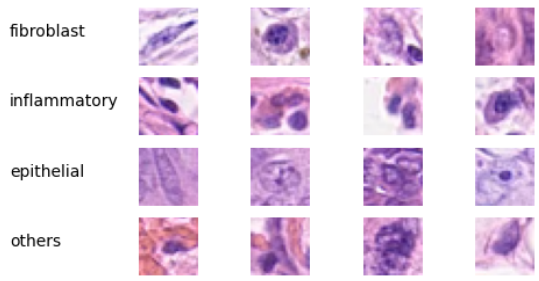

Figure 3 Cell Images by cell type

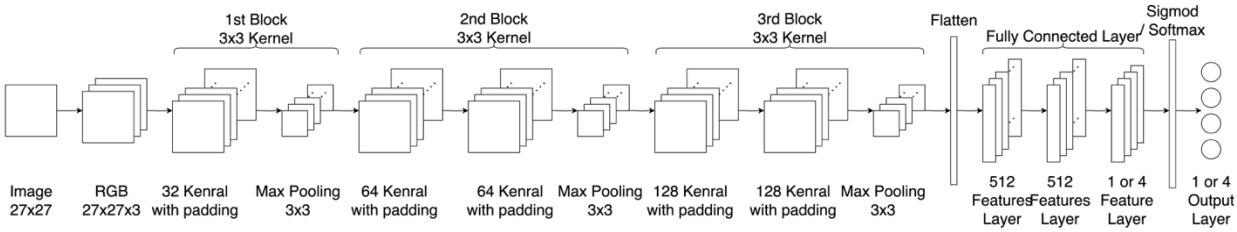

Figure 4 Baseline Model Architecture

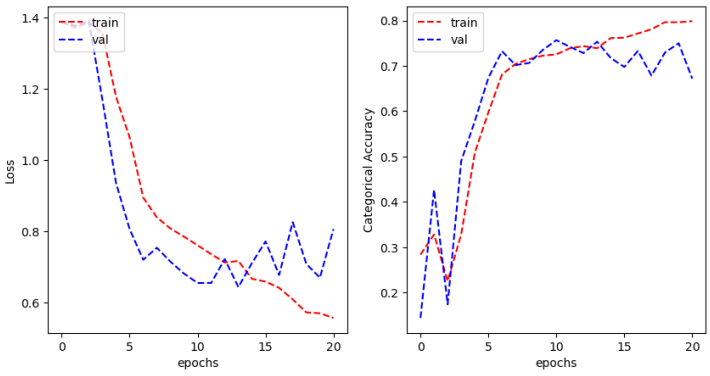

Figure 5 Cell Type Baseline Model Train. and Val. Acc.

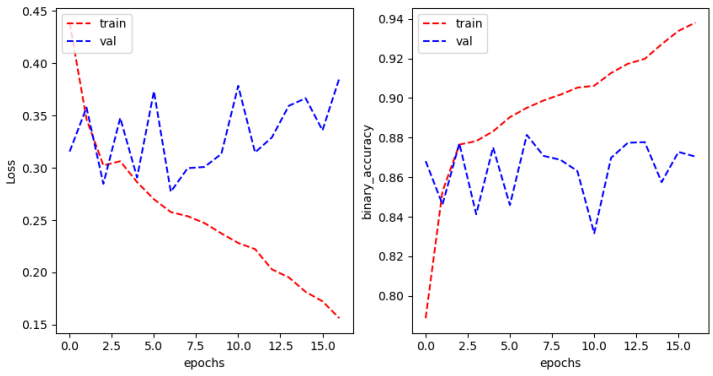

Figure 6 Cancerous Baseline Model Train. and Val. Acc.

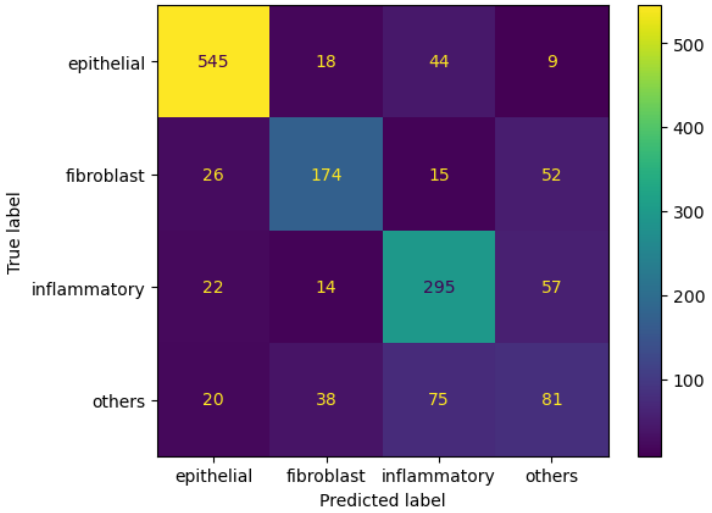

Figure 7 Cell Type Baseline Model Confusion Matrix

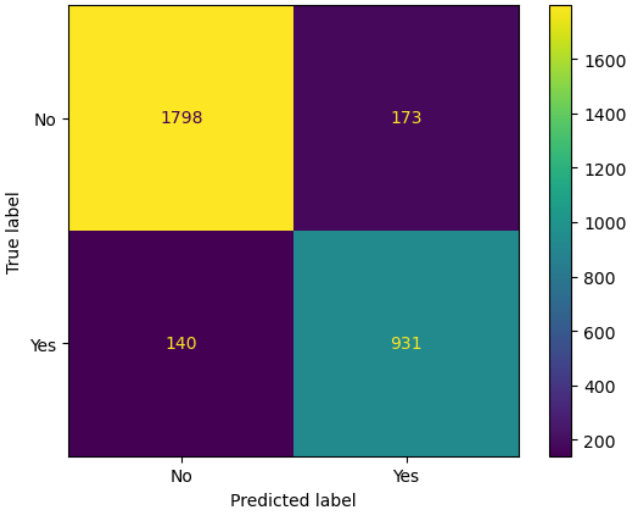

Figure 8 Cancerous Baseline Model Confusion Matrix

Figure 9 Data Augmentation Images

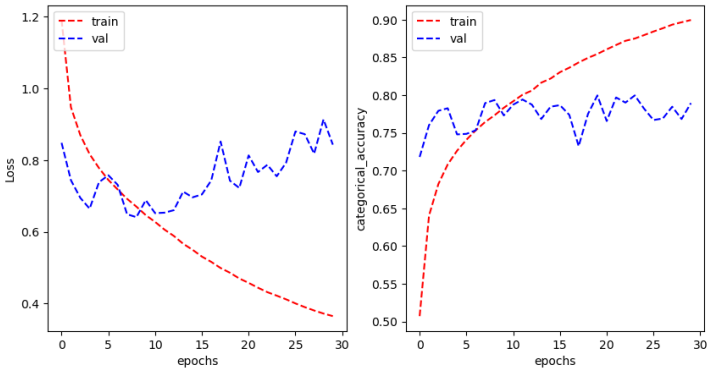

Figure 10 Cell Type Over Sampling Model Train. and Val. Acc.

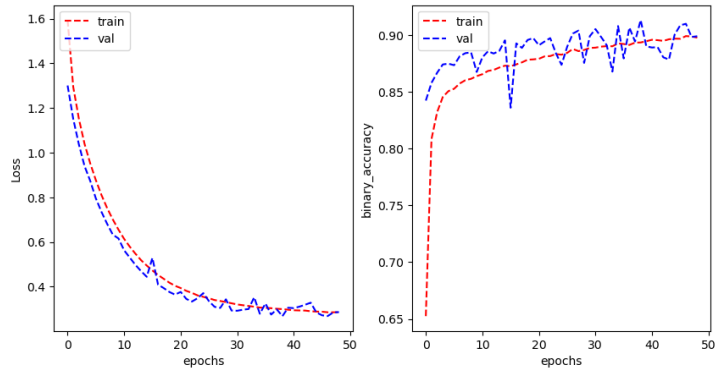

Figure 11 Cancerous Over Sampling Model Train. and Val. Acc.

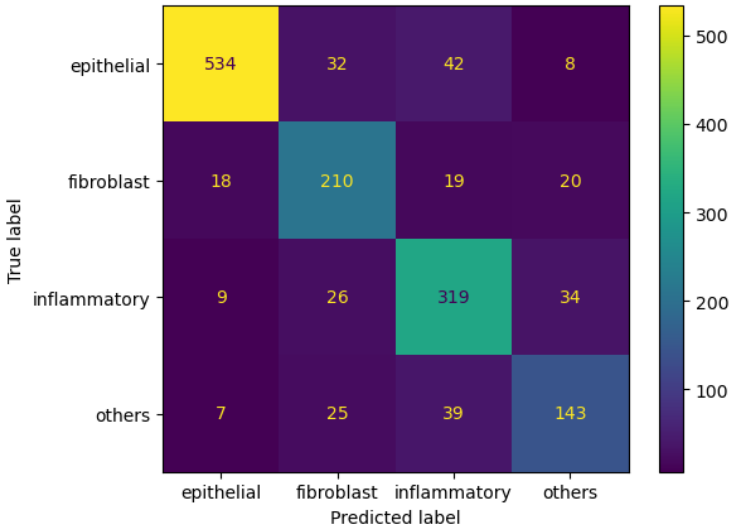

Figure 12 Cell Type Oversampling Model Confusion Matrix

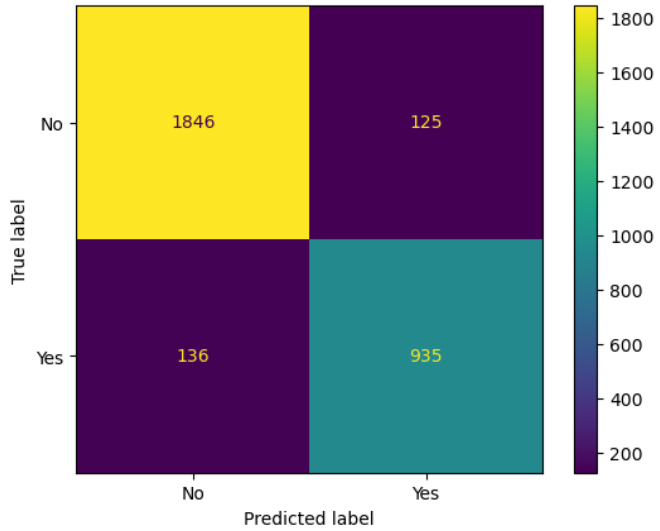

Figure 13 Cancerous Oversampling Model Confusion Matrix