Classifying Knee X-rays

Identifying the severity of knee osteoarthritis in x-ray images

Introduction

According to the National Institute of Health, osteoarthritis (OA) is a joint disease in which the tissues in the joint break down over time. It is the most common type of arthritis and is more common in older people. Knee osteoarthritis is one of the most common types of OA. The CDC states that OA affects over 32.5 million adults in the United States. As the population grows and ages, this number is projected to increase. There is no cure for OA, and thus those afflicted must learn strategies to manage their pain and reduce disability. Early identification of OA in patients may help them change their lifestyle in order to slow the progression of OA.

Knee radiographs, or X-rays, are examined by radiologists and quantified using the 5-point, semiquantitative grading scale known as the Kellgren-Lawrence (KL) scale. This classification was proposed by Kellgren and Lawrence in 1957 and later accepted by the World Health Organization (WHO) in 1961 as the radiological definition of OA for the purpose of epidemiological studies[1]. The grade descriptions are as follows, wherein osteoarthritis is deemed present at grade 2 although of minimal severity:

- Grade 0 (None): No pathological features, i.e. a healthy knee image
- Grade 1 (Doubtful): Doubtful joint narrowing with possible osteophytic lipping
- Grade 2 (Minimal): Definite presence of osteophytes and possible joint space narrowing
- Grade 3 (Moderate): Multiple osteophytes, definite joint space narrowing, with mild sclerosis.
- Grade 4 (Severe): Large osteophytes, significant joint narrowing, and severe sclerosis.

With this backdrop, can machine learning aid in the identification of knee osteoarthritis in knee x-ray images? We will use a dataset from Kaggle: Knee Osteoarthritis Dataset with Severity Grading . This dataset was organized from OAI and provided by Mendeley Data (Chen, Pingjun (2018), "Knee Osteoarthritis Severity Grading Dataset", Mendeley Data, V1, doi: 10.17632/56rmx5bjcr.1).

Problem Statement

We want to build a machine learning classification model that can analyze a knee radiograph and assign it a grade on the KL scale. Using a neural network with transfer learning from an image classifier, can we develop a model that can accurately grade the severity of a knee X-ray?

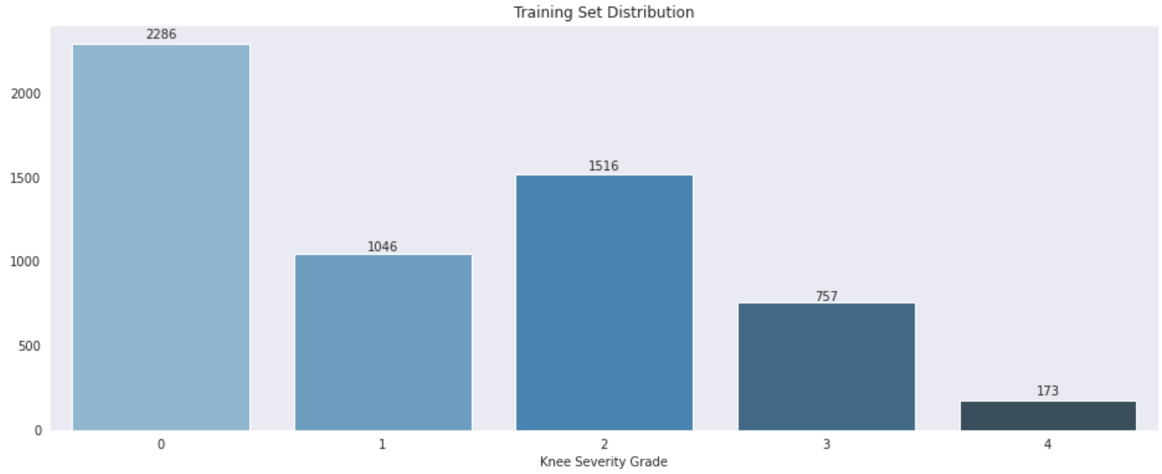
Ultimately, our goal is to aid patients with knee pain and help them understand whether their knee X-ray image shows OA, and if so, how severe that OA is to help them determine their next steps. Thus, if the classification model using 5 classes has insufficient accuracy, we can simplify our classes from 5 to 3 as follows:

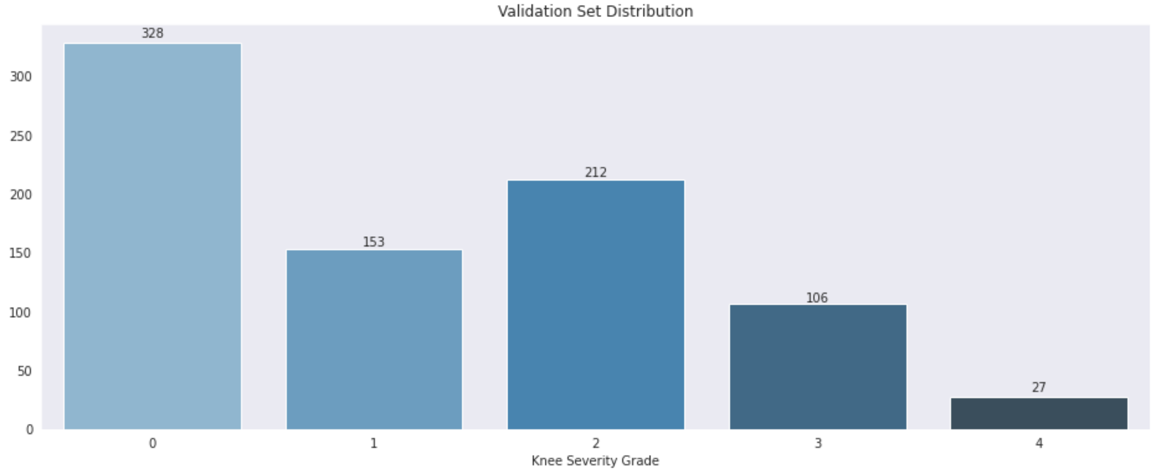
- **A** : Grades **0** and **1** : a reasonably healthy knee, patient with knee pain may wish to explore other causes
- **B** : Grade **2** : patient has minimal OA and should consider lifestyle changes, and possibly medication
- **C** : Grades **3** and **4** : knee replacement and other procedures and/or surgeries may be an option

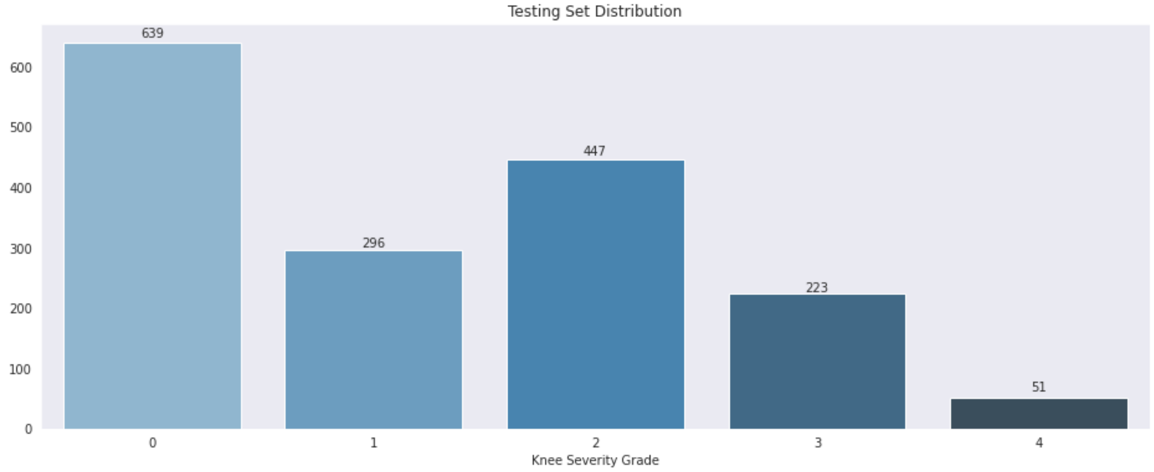
Using the simplified classes, we can create another image classification model and review its accuracy.

Exploratory Data Analysis

We first looked at the overall distribution of the X-ray images in the training set, validation set, and test set. We can see that we have an imbalanced dataset: Grade 0 consists of approximately 40% of the entire dataset, Grade 1 consists of ~18%, Grade 2 of ~26%, Grade 3 of ~13%, and Grade 4 of only ~3%. This imbalance will impact our model, so we must pass along class weights when training the model.

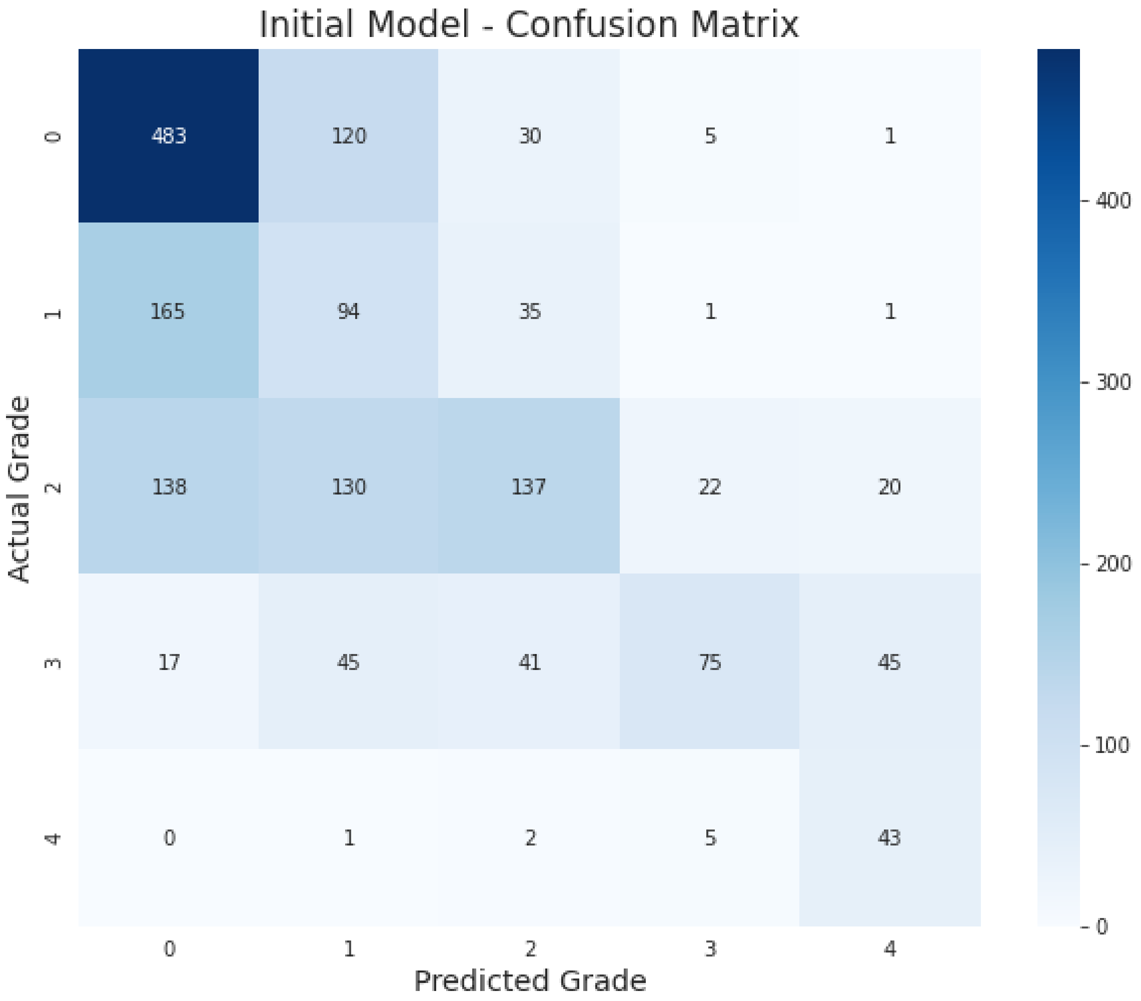


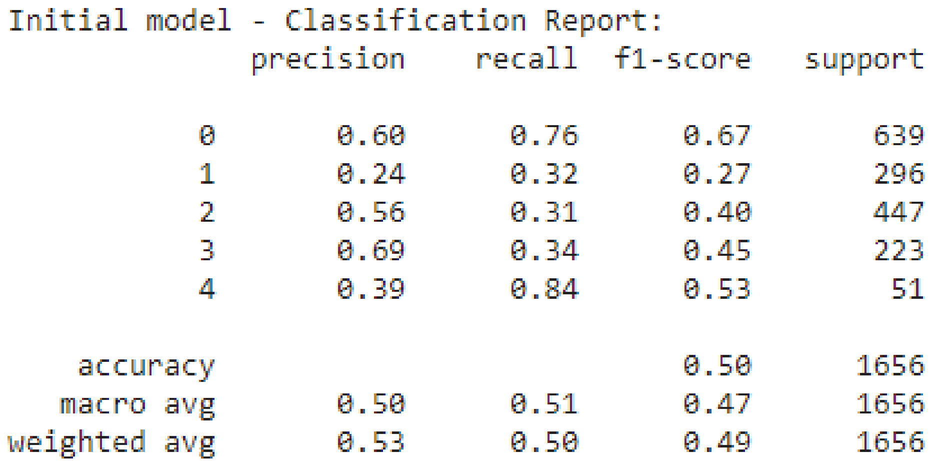




Modeling – 5 grades on the KL-scale

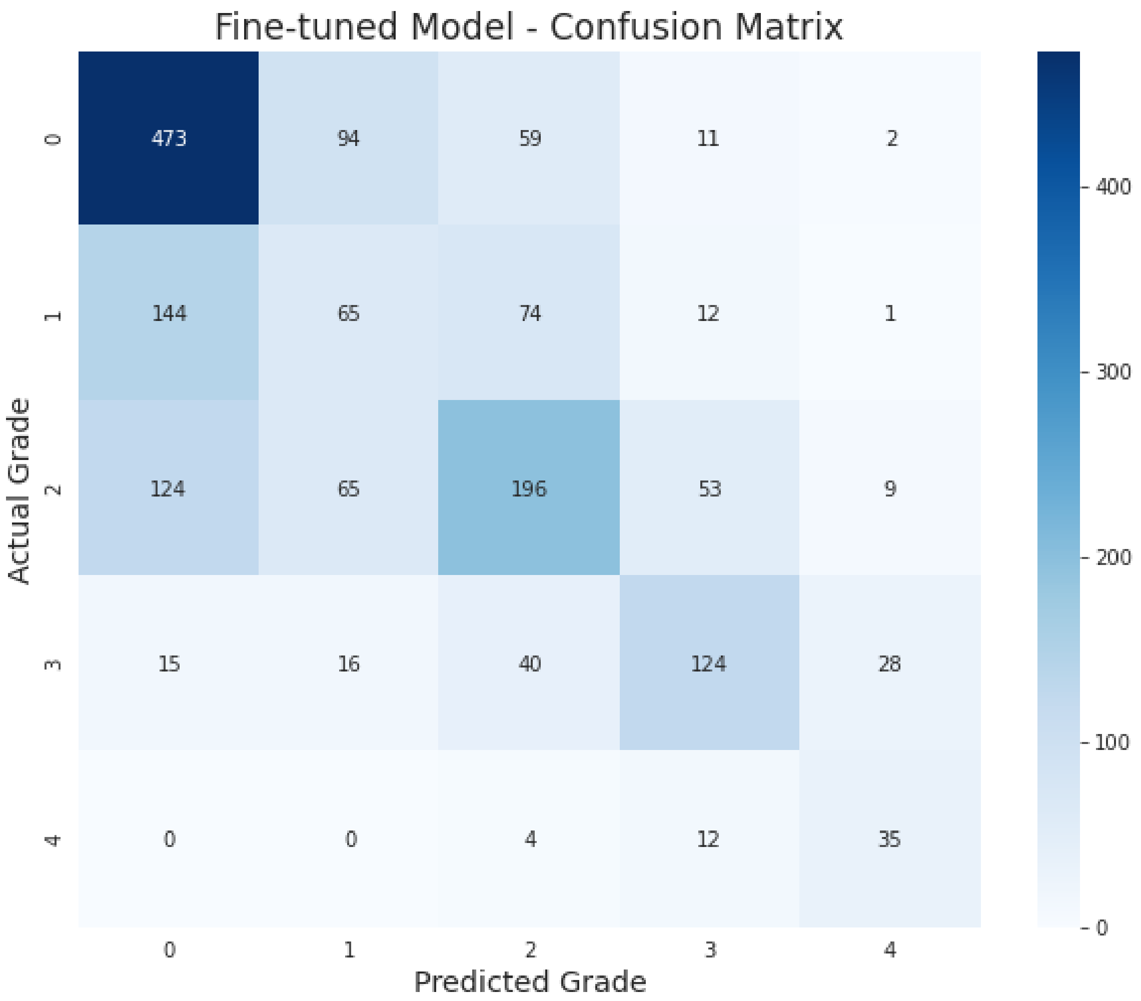
We used transfer learning to initialize our base model. We tried both Xception and DenseNet201 and determined that DenseNet201 had better results. As such, we created a base model using DenseNet201 with weights pre-trained on ImageNet and added a few of our own layers and classifier on top. We froze the base model's layers and did an initial run of 15 epochs to initialize weights for our classifier. This initial model gave us an accuracy of **50.24 %** on the test set.



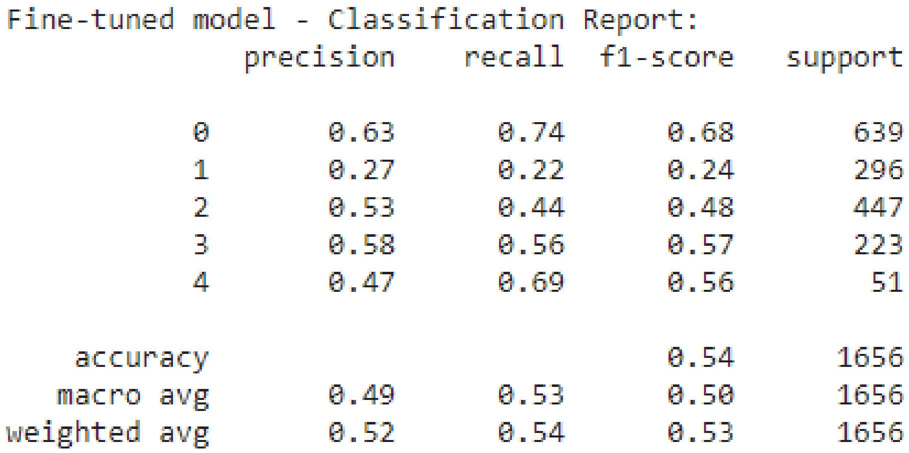


```
Initial model - Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.76      0.67       639
           1       0.24      0.32      0.27       296
           2       0.56      0.31      0.40       447
           3       0.69      0.34      0.45       223
           4       0.39      0.84      0.53        51

    accuracy                           0.50      1656
   macro avg       0.50      0.51      0.47      1656
weighted avg       0.53      0.50      0.49      1656
```

Then we fine-tuned the model by unfreezing the last 8 layers of the base model and using a learning rate scheduler to lower the learning rate as we ran 50 epochs. This improved our accuracy to **53.93 %** on the test set.


Fine-tuned Model - Confusion Matrix

Actual \ Predicted	0	1	2	3	4
0	473	94	59	11	2
1	144	65	74	12	1
2	124	65	196	53	9
3	15	16	40	124	28
4	0	0	4	12	35

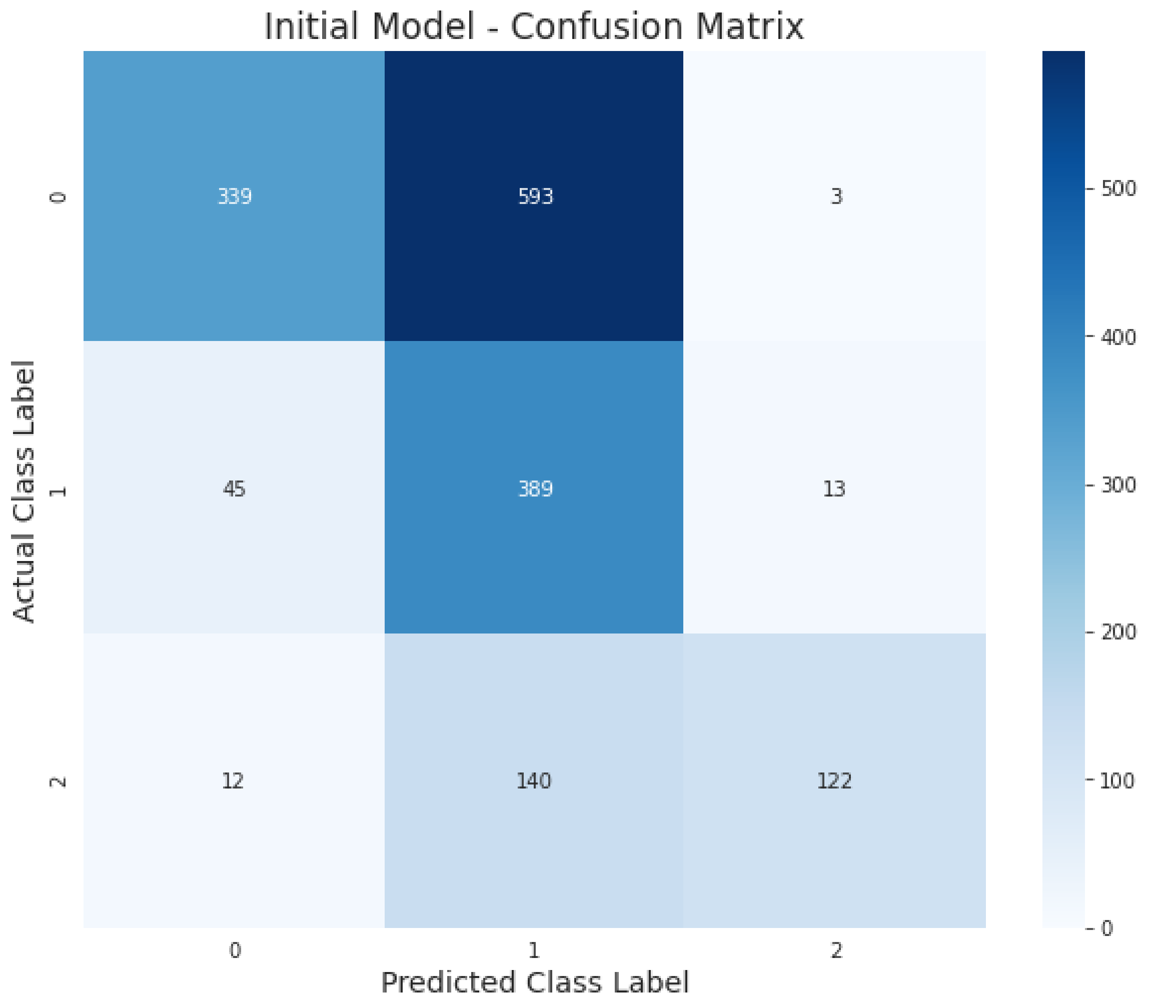

```
Fine-tuned model - Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.74      0.68       639
           1       0.27      0.22      0.24       296
           2       0.53      0.44      0.48       447
           3       0.58      0.56      0.57       223
           4       0.47      0.69      0.56        51

    accuracy                           0.54      1656
   macro avg       0.49      0.53      0.50      1656
weighted avg       0.52      0.54      0.53      1656
```

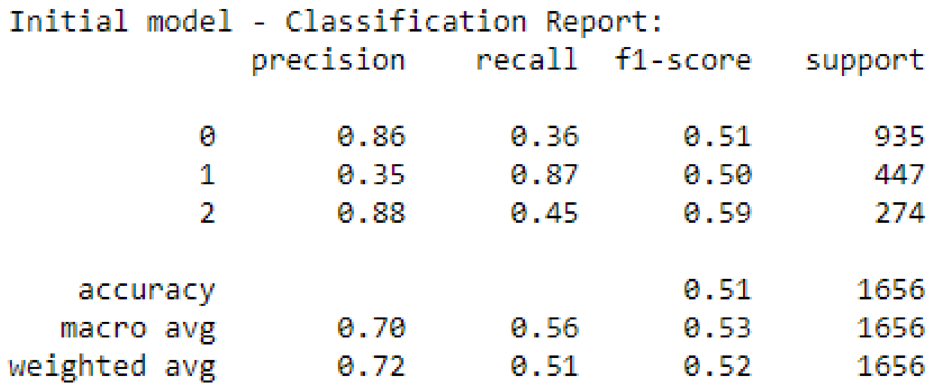
The metrics reveal that we do a passable job at identifying grades 0 and 4, a mediocre job at identifying grades 2 and 3, and an awful job at identifying grade 1.

Although we tried to improve upon this model, all versions of this model produced an accuracy on the test set that topped out at around ~50%. Thus, we shifted to simplifying our data by reducing the 5 grades into 3 classes.

Modeling – simplified into 3 classes

Using our simplified classes (condensing the grades as follows: A = grades 0 and 1, B = grade 2, C = grades 3 and 4), we continue to use DenseNet201 as our base model. The initial model, trained on 15 epochs, produced an accuracy of **51.33 %** on the test set.

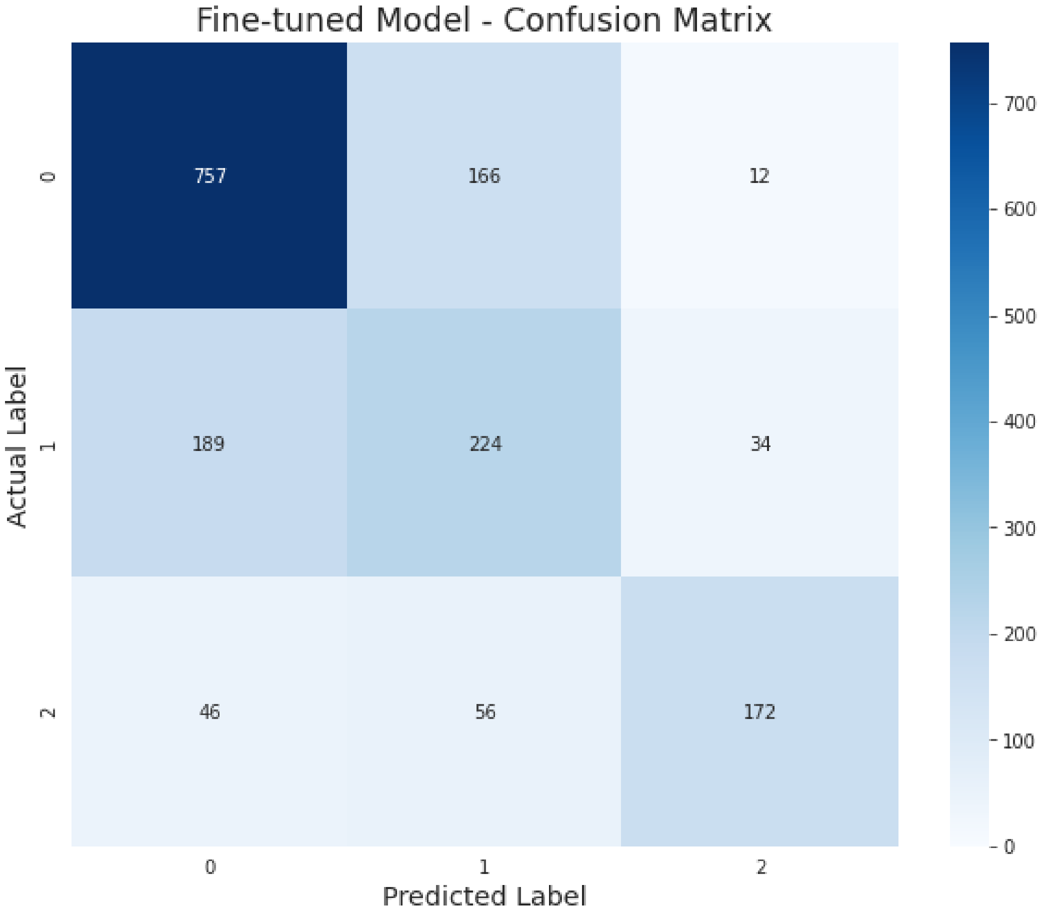


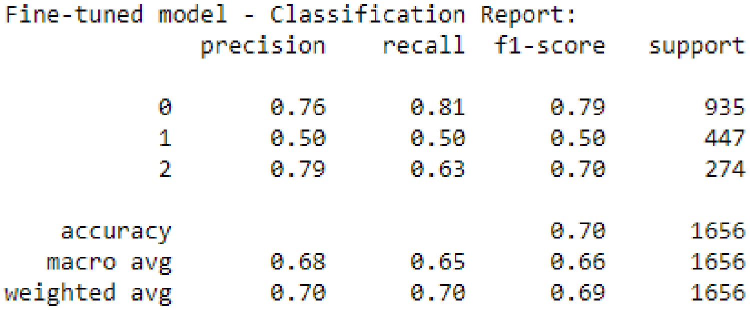


```
Initial model - Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.36      0.51       935
           1       0.35      0.87      0.50       447
           2       0.88      0.45      0.59       274

    accuracy                           0.51      1656
   macro avg       0.70      0.56      0.53      1656
weighted avg       0.72      0.51      0.52      1656
```

Note that classes A, B, and C are referred to in the confusion matrix and classification report as 0, 1, and 2 (0 = A, 1 = B, 2 = C).

We fine-tuned the model by unfreezing the last 8 layers of the base model. We added a learning rate scheduler to decay the learning rate every 30 epochs. We also increased the number of epochs to 100, and added an early stopping callback to end the model run if no improvement was shown. This fine-tuning training ran for 73 epochs and produced an accuracy of **69.63 %** on the test set.





```
Fine-tuned model - Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.81      0.79       935
           1       0.50      0.50      0.50       447
           2       0.79      0.63      0.70       274

    accuracy                           0.70      1656
   macro avg       0.68      0.65      0.66      1656
weighted avg       0.70      0.70      0.69      1656
```

Note that classes A, B, and C are referred to in the confusion matrix and classification report as 0, 1, and 2 (0 = A, 1 = B, 2 = C).

These metrics are much improved from the previous 5-class model. We do a good job identifying class A (as we defined earlier, a reasonably healthy knee) as well as class C (knee OA is sufficiently advanced as for the patient to explore knee replacement or alternate procedures and surgeries). We still only do a mediocre job on identifying class B (minimal knee OA), and from the confusion matrix, we seem to err on the side of “healthy”.

Ideas for further research

1. While the KL scale for grading knee OA severity is one of the most widely used, it is not the only evaluator of knee OA. Furthermore, it can be subjective (grades are dependent on the radiologist). We should explore other methods of identifying knee OA from knee radiographs and use those as additional features for the model.
2. Along those lines, we can seek out additional data. There might be additional knee X-rays out there, although they may not be as clean or accurately labeled.
3. In the hopes of aiding in early identification, we will want to continue fine-tuning to improve classification of the grade 2 / class B radiographs, which would indicate minimal knee OA for the patient.