Classifying Knee X-Rays Into Different Categories of Knee Osteoarthritis

Servet Efe Tekin[1,*]

[1] *Uskudar American Academy, Istanbul, Turkey*

ABSTRACT

Due to varying severity in patients with knee osteoarthritis, appropriate staging of patients by studying clinical features of X-ray imaging is critical for proper diagnosis and management. Staging of patients is performed by manual observation of X-ray images by a radiologist. Machine learning models have been trained to diagnose knee osteoarthritis from X-ray images, however most of these models will perform a binary classification of patients which does not reflect disease severity. In this study, we train a model to categorize patients into one of three classes based on the Kellgren-Lawrence classification of knee osteoarthritis severity. With a dataset of 4,267 knee X-ray images we were able to achieve a validation F1 score performance ranging from 0.67 to 0.79. Our strategy for achieving high performance from a limited training set size included proper augmentation of X-ray images, early stopping, and application of weight decay to the cost function.

1. INTRODUCTION

The lifetime treatment cost for an patient with knee osteoarthritis (OA) is estimated to range from 12,400 to 16,000 dollars in the USA.[1] Diagnosis of knee OA can be performed by manual observation of a knee X-ray, in a healthy joint a gap will be present between the femur and tibia which narrows with progressing OA. Knee OA is assessed with a grading scale which ranges from no signs to severe OA as assessed by observation of the knee joint in X-ray images. Narrow artificial intelligence (AI) has been shown to achieve superior performance than humans in automating specific tasks, in particular image recognition tasks [2]. Given the information for knee OA grading is contained entirely within the X-ray images of the knee joint, grading of knee OA is an excellent candidate for a narrow AI problem. In this research paper, we describe our study to generate a convolutional neural network (CNN) for grading of knee OA severity. We tested different cost functions, gradient descents, and CNNs and find the optimum combination. We addressed difficulties relating to class imbalance and overtraining with appropriate merging of classes, weight decay, and image augmentation. Furthermore, we attempted to utilize a distinct dataset of poorly labelled knee X-ray images using a transfer learning and semi-supervised consistency learning approach.

1.1 Knee Osteoarthritis

The largest synovial joint that the human body has is the knee. Thus, there is stress and high use of this joint which eventually degrades, in most of the cases, into painful diseases such as OA. Classification of OA is done most frequently with the Kellgren-Lawrence system. OA can be a degenerative disease of the cartilage as well as a disease due to trauma, biomedical reactions, and mechanical forces. Cartilaginous tissue isn't only part involved in OA due to its lack of vasculature and innervation. Therefore, the first stages of OA generates pain due to the changes in non-cartilaginous components of the knee. The further stages of the disease can cause osteophyte formation, weak muscles, remodeling of bones. [3]

1.2 Supervised Machine Learning

Supervised machine learning relies on an iterative algorithm which updates parameters of the model to improve predictive performance on a specific task. In relation to diagnosis of knee osteoarthritis the algorithm will iteratively update parameters in the neural network in order to converge on a model which can accurately diagnose grade of severity of knee osteoarthritis from X-ray images. Therefore, supervised machine learning requires pre-labelled training data which is used in the training process, however labeled data is often limited, therefore, it is difficult to use supervised learning.

1.3 Knee Osteoarthritis X-ray dataset

The dataset used in this research is divided into 4 different files as test, train, val, autotest. For the purpose of this research, the test and train data will be used which is a total

of 4,267 knee x-ray images with 3,332 for training and 935 for testing. The images have already been cropped for the area of interest. The image size is 224x224 pixels. [4].

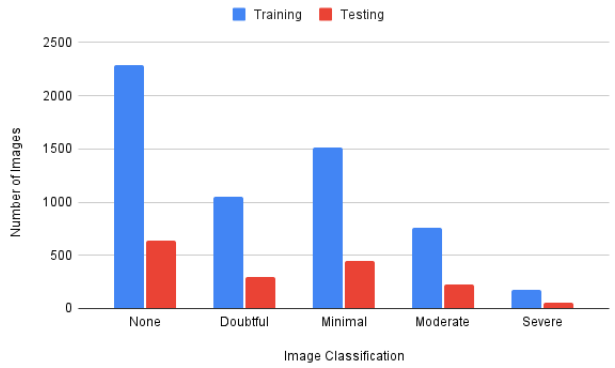

Fig. 1. The distribution of label data before merging

Kellgren-Lawrance classification system for OA was aimed to be used

Table. 1. Kellgren-Lawrance

Class	Definition
0	Healthy knee image.
1	Doubtful joint narrowing, possible osteophytes
2	Definite osteophytes, possible joint space narrowing
3	Multiple osteophytes, definite joint space narrowing
4	Large osteophytes, significant joint narrowing

However, due to the lack of dataset that will be addressed further in method parts, a new classification was generated.

Table. 2. Classes of OA

Class	Definition
0+1	Evidence of joint space narrowing
2	Possible joint space narrowing
3+4	Definite joint space narrowing

2. METHODS

Machine learning enables automation of tasks through an iterative approach, without explicitly programming algorithms, this enables models to complete complex tasks which would otherwise be difficult to manually program. In particular, supervised machine learning is one type of Machine Learning that uses the method of giving samples of the labeled data to the computer in order for it to generate a model. Transfer learning, another kind of machine learning, uses a different dataset of similar images in order to generate a model prior to the actual dataset being uploaded. Later on, the pre-model is altered according to the new dataset. Furthermore, consistency learning uses a neural network to conduct both supervised and unsupervised machine learning. While the supervised learning measures the accuracy of the model, unsupervised learning measures only the consistency of the model. Consistency learning enables unlabelled datasets to be used to improve the performance (consistency) of a supervised model.

2.1 Sample Weighting

Supervised machine learning uses batches of data in each iteration that are randomly selected from the overall dataset. Random selection results in the probability of an image from each category being selected in a batch to be proportional to the frequency of that class): (number of images in that class)/(total number of images). This results in rare classes being often absent from the batch resulting poor predictive performance for that class.

Sample weighting modifies the probability of selection images from each class with a multiplier that makes selection of images from rare classes more likely: (total number of images)/ (number of images in that class). The result of sample weighting is that images from each class are equally likely to be selected in a batch with images from rare classes repeated to enable an epoch of training to complete.

2.2 F1 Score

The harmonic mean of precision (Eq. 1) and recall (Eq. 2) is used to get the F1 score (Eq. 3). Since the F1 score is the average of Precision and Recall, their contribution to the F1 score are equally weighted [5]. F1 scores will be used to determine the level of performance of a model for each of the classes, with TP, representing true positives, FP, representing false positives, and FN, representing false negatives.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

2.3 Weight Decay

Weight decay enables regularization by modifying the cost function to account for model complexity in addition to accuracy. L2 weight decay calculates complexity with the

squared sum of all the parameters in the model and multiplied by hyperparameter λ , a term which is added to the overall cost function (Eqn. 4). The λ hyperparameter allows modulation of the weight decay term influencing the degree of weight decay.

$$RegularizedCost = Cost + \lambda \sum_{i=1}^{n} \theta_i^2 \quad (4)$$

2.4 Convolutional neural networks

2.4.1 Resnet18

ResNet-18 model was originally proposed by He et al., 2016 [6], with an accuracy of 89% after five iterations of training on ImageNet data. ResNet-18 has a residual learning framework which enables a deeper network while keeping the number of trainable parameters lower than other network architectures of comparative size [6]. The layers in ResNet-18 are reformulated as functions that do not learn unreferenced functions, but learn residually with reference to the layer. These residual networks can gain accuracy from considerably increased depth and are easier to optimize [6].

2.4.2 Mobilenetv3small

MobileNetV3 model was originally proposed by Howard et al.,2019 [7], with an accuracy of 87.4 % after five iterations of training on ImageNet data. Mobile classification, detection and segmentation were aimed in the creation of MobileNetV3. MobileNetV3 is designed for mobile phone's CPUs whereas in our research a GPU is used [7].

2.4.3 Googlenet

GoogLeNet (Inception v1) model was originally proposed by Szegedy et al., 2014 [8], with an accuracy of 89.530 % after five iterations of training on ImageNet data. In the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014), GoogLeNet had great success on classification and detection. Enhanced utilization of the computing resources within the network is the hallmark of GoogLeNet. Increasing the depth and width of the network while keeping the computational budget constant, was both the difficulty and key to success. The architectural decisions were based on the the intuition of multi-scale processing to optimize quality.[8]

2.4.4 Efficientnetb0

EfficientNet B0 model was originally proposed by Tan et al., 2019 [9], with an accuracy of 93.5 % after five iterations of training on ImageNet data. EfficientNet B0 model scales proportional to the dataset size [9]. However, the dataset in this research is limited so it is insufficient for EfficientNet B0.

2.4.5 Regnety400mf

RegNetY400MF model was originally proposed by Radosavovic et al., 2020 [10], with an accurecy of 91.7 % after five iterations of training on ImageNet data. Regnet outperforms even Efficientnet model. However, the model again requires wide range of flop regimes [10] which is lacking in this research.

2.5 Transfer Learning

Transfer learning is a strategy to enable better initialization of model parameters. The model is initially trained on an off-target dataset with similar outcome and input information prior to training on the intended dataset. Later on, the pre-model is altered according to the new dataset.

2.6 Consistency Learning

Consistency learning uses a neural network to conduct both supervised and unsupervised machine learning. While the supervised learning measures the accuracy of the model, unsupervised learning measures only the consistency of the model. Consistency learning enables unlabelled datasets to be used to improve the performance (consistency) of a supervised model. When the labelled data is insufficient, semi-supervised learning is utilized [11]

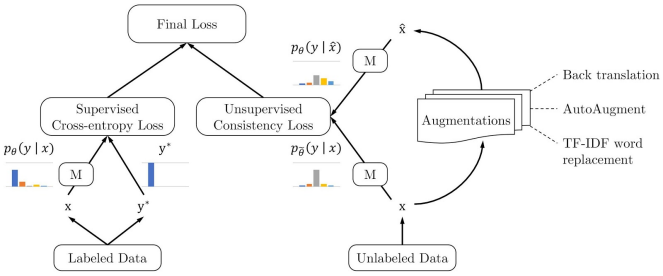

Fig. 2. A summary of UDA (Unsupervised Data Augmentation).

On the right, a consistency cost is computed between an example and its augmented version with unlabeled data. On the left, standard supervised cost is computed when labeled data is available.

2.7 Cost function

2.7.1 Ordinal Cost Function

Ordinal classification or ordinal regression is a kind of regression analysis, for the prediction of an ordinal variable such as a variable whose value can be expressed on any scale where the relative ordering of different values is the only thing that matters

2.7.2 Cross entropy loss

The performance of a classification model whose output is a probability value between 0 and 1 is measured using cross-entropy loss, also known as log loss. Cross-entropy loss increases as the anticipated probability depart from the true label. Therefore, forecasting a probability of.005 when the actual observation label is 1 is bad and leads to a high loss value. The log loss in a perfect model would be zero.

2.7.3 Multi-Margin Loss

Creates a criterion for optimizing a multi-class classification hinge loss (margin-based loss) between the input aa (a 2D mini-batch Tensor) and the output bb (a 1-dimensional tensor of target class indices).
HingeLoss: The hinge loss is a form of cost function that calculates the cost based on a margin or distance from the classification boundary. Even if new observations are correctly classified, they may be penalized if the margin from the decision border is insufficient. The hinge loss rises in a linear way.

2.8 Gradient descent

Different algorithms are used to obtain a model performance as high as possible.

2.8.1 Adam

The Adaptive Movement Estimation algorithm, or Adam for short, is a natural successor to techniques like AdaGrad and RMSProp that automatically adapts a learning rate for each input variable for the objective function and further smooths the search process by making updates to variables using an exponentially decreasing moving average of the gradient.

2.8.2 Adagrad

AdaGrad is a stochastic optimization technique that adjusts the learning rate according to the parameters. It changes parameters associated with frequently occurring features, more frequently than parameters associated with infrequently occurring features. Adagrad's update rule modifies the general learning rate for each parameter at each step depending on previous gradients.

2.8.3 Adamax

AdaMax is a modification to Adam's gradient descent algorithm that generalizes the approach to the infinite norm (max) and may result in more effective optimization on particular situations.

2.8.4 Nadam

The Nadam, or Nesterov-accelerated Adaptive Moment Estimation, is an extension of the Adam method that integrates Nesterov momentum and can improve the optimization technique's performance.

2.8.5 Adadelta

Instead of accumulating all prior gradients, Adadelta is a more resilient variant of Adagrad that adapts learning rates based on a changing window of gradient updates. Adadelta learns in this way even after multiple updates have been completed.

3. RESULTS

3.1 Weighting distributes the images in a batch equally causing the class with fewer images to have three-fold higher F1 scores while having no major effect on other classes.

Supervised machine learning relies on an iterative training approach. In each iteration images from all the categories are drawn randomly from the training set, with the possibility that a category is not selected in that particular batch. Therefore, when a class has fewer data than others, there is a chance of getting no images from that class in a batch, therefore the application of gradient descent in a batch where a class is missing will result in reduced predictive performance following gradient descent for the missing class. By using sample weighting (Section 2.1), the probability of every image appearing in a batch is equalized. As a result, the F1 scores for class 1 and class 4 almost tripled (Fig. 3). There was no major change in F1 scores of class 2 and 3. The average increase in F1 scores was 87%.

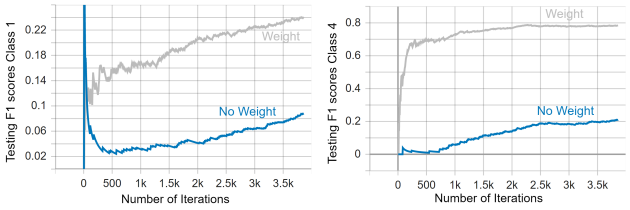

Fig. 3. The F1 score for class 1 and 4 vs iteration graph before and after weighting. The F1 score increased for both of the cases

However, by doing so F1 score of class 1, which was the dominant class before the weighting, decreased (Fig. 4).

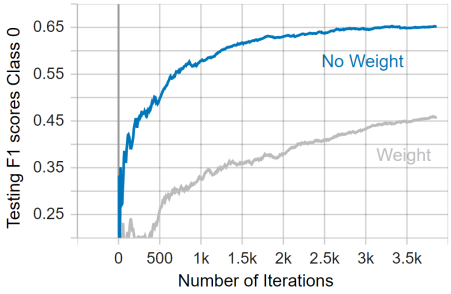

Fig. 4. The F1 score vs iteration graph for class 0

3.2 Merging balances the lack of data for categories 1 and 4 which results into a 92% overall increase in the model performance

Since the number of data in each class profoundly differs, weighting decreased some of the scores in return for equally distributed batches. To increase the F1 score of each class, the dataset is merged by combining class 0 with class 1 and class 3 with class 4 (Table.5) so that the weighting is not excessive.

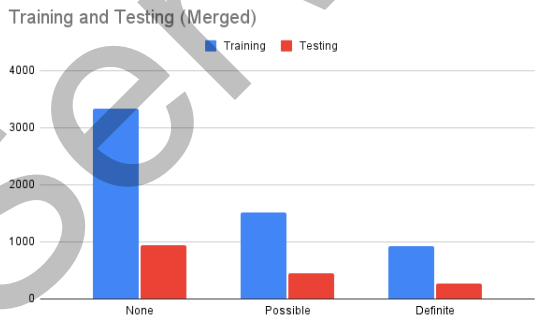

Fig. 5. The distribution of label data after merging

Since images in class 2, 3, and 4 were lacking, merging them enhanced the training process resulting in better overall F1 scores (Table.6).

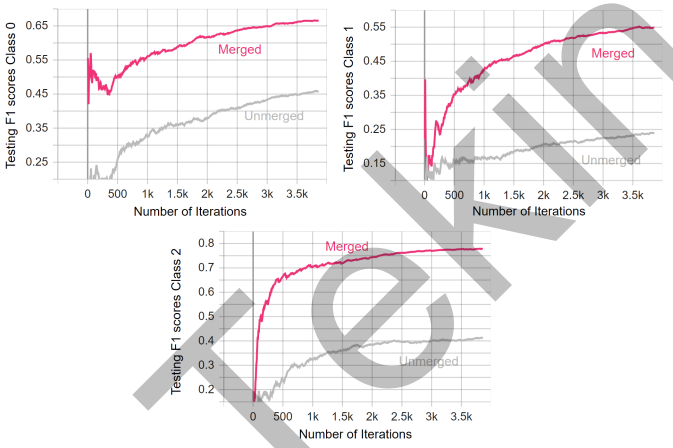

Fig. 6. The F1 score vs iteration graph for class 0, 1, and 2

With the combination of weighting and merging, the class imbalance is significantly eliminated.

3.3 Image augmentation (Shear and Horizontal Flip) increases generalization which reduces overfitting by 45% while keeping the model performance fixed.

Shear and horizontal flip are decided to be the most appropriate image augmentations (decided by test runs) for the purpose of reducing overfitting while minimizing its effect on the model performance. Horizontal flip rotates the image 180 degrees with respect to y-axis (Fig.9) and shear distorts the image in order to train the model for x-rays that are taken from another angle (Fig.8).

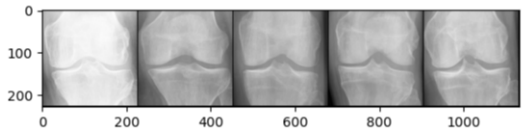

Fig. 7. Samples of knee x-ray images before any transformation

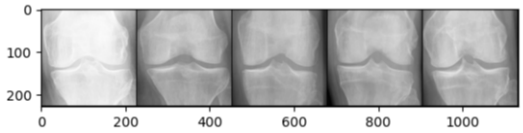

Fig. 8. Samples of knee x-ray images after shear alone

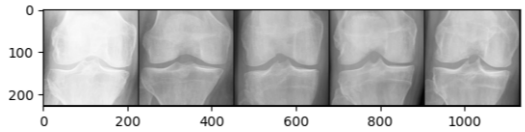

Fig. 9. Samples of knee x-ray images after horizontal flip alone

Image Augmentations generate variability in data simulating an increase in the number of data which creates a more efficient model. The variability in data also causes the model to be accustomed to distinct images. Thus, the model is aimed not to be limited only to the extend of the training dataset used. In other words, overtraining is attempted to be eliminated. Moreover, the difference in the test F1 score and the train F1 score is successfully lowered in all cases (Table.3).

The training F1 scores decreased after the utilizing shear and horizontal flip. The aim was always to generate F1 scores that are as high as possible. However, even though the training F1 scores seems substantially higher before the augmentations, it is not what is desired. The training scores getting profoundly larger than testing scores indicates overfitting (Table.3). Overfitting occurs when the model becomes too accustomed for the training data that it gets very high scores for the training data while lower scores for testing since the testing data are different than what the model was overly adjusted to.

3.4 Early Stopping decreases the difference between train F1 scores and test F1 scores by 12% which indicates prevention of overtraining in the model

Epoch number is the number of times the model is trained with the data. By cross-examining the Train-Test F1 scores versus number of iterations graph of 30 epochs and the overall test performance, the epoch number of 18 is decided to be the most suitable number of epochs that the model is trained since at the 18th epoch, there are low the Train-Test F1 scores meaning no overtraining apparent as well as high performance. (Fig.10).

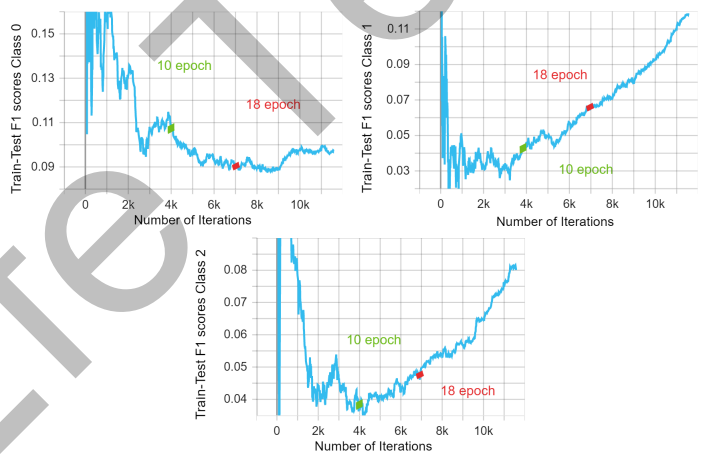

Fig. 10. The Train-Test F1 score vs iteration graph for class 0, 1, and 2 (A total of 30 epochs)

10 epochs is the original epoch number that we have been using. It is observed that at the 18th epoch all of the train-test differences are less than 0.1 without sacrificing any level of performance that we previously had in the 10th epoch and after 18th epochs, the model starts to get very focused on the training data.

Class	Before Augmentation			After Augmentation		
	Testing F1 Score	Training F1 Score	Difference	Testing F1 Score	Training F1 Score	Difference
0	0.66	0.80	-0.14	0.67	0.76	-0.084
1	0.54	0.73	-0.19	0.67	0.65	-0.091
2	0.79	0.91	-0.11	0.79	0.86	-0.064

Table. 3. The Training and testing F1 scores with their differences before and after the image augmentation (Horizontal Flip + Shear).

3.5 The optimum weight decay value for both highest performance and least overfitting was found to be 0.001

Weight decay value is the value of lambda in the cost equation (Fig.4). The value of weight decay determines the change in the complexity of the model (Section 2.3). If the model is too complex, meaning the model adjusted itself to fit perfectly to the training data, we will decrease the complexity so that it is more general. However, if we make it too general, then the model will not be successful at predicting the classes, meaning F1 scores will be less. By altering the value of weight decay, an appropriate value that both keeps the performance up and decreases overtraining is found to be 0.001. Higher values of weight decay made the model too general which decreased the model performance, and lower values did not reduce overfitting as much.

3.6 Transfer learning to leverage an external dataset of Knee X-ray images did not result in improvements in performance

The model is first trained by a dataset [12] of knee x-ray images. Later on, the pre-trained model is further trained on the intended dataset (Section 2.5). However, due to the fact that the dataset that is used for pre-training, was not as similar as required, the transfer learning did not result in a better outcome but only decreased the F1 scores in the specified epoch number for each class by at least 0.05 up to 0.1 (Fig.**??**). The dataset was different from the original dataset in terms of classification numbers (3 classes in the original dataset but only 2 classes in the second dataset) and image size (the images in the original dataset was 224x224 pixels whereas the images in the additional dataset was 1127x2660 pixels). Moreover, the dataset consisted of only 372 images while the original dataset possesses 3,925 images.

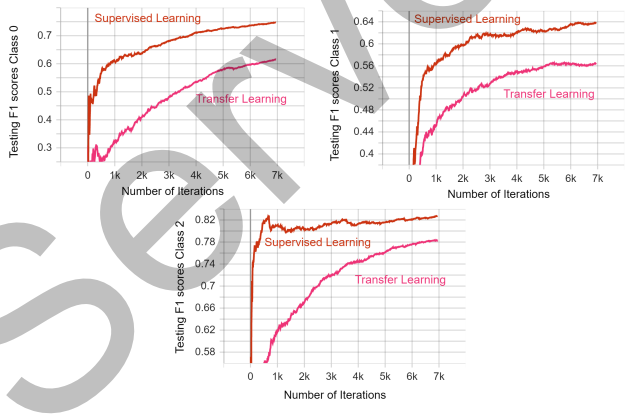

Fig. 11. The test score vs iteration graph for class 0, 1, and 2

3.7 Semi-supervised consistency learning with data augmentation results in decreases in model performance

Consistency learning is a strategy that uses both supervised and unsupervised learning to generate a model that is both accurate and consistent (Section 2.6). We used two datasets for this learning strategy. The original dataset [4] with the supervised for accuracy and the second dataset [12] with the unsupervised for consistency. The gradient descent SGD was used for the learning method. For the supervised training dataset, the following augmentations were applied in order to generalize the data:

```
Pad(4),
RandomCrop(32, fill=128),
RandomHorizontalFlip(),
Normalize((0.485, 0.456, 0.406), (0.229,
    0.224, 0.225)),
RandomErasing(scale=(0.1, 0.33)),
```

For the unsupervised dataset, only padding and random cropping was used with RandAugment. RandAugment is a data augmentation approach that is automated. There are two interpretable properties in the data augmentation search space: is the number of augmentation transformations to apply sequentially, and is the magnitude of all transformations. To keep image diversity while reducing parameter space, learnt policies and probabilities for applying each transformation are replaced with a parameter-free algorithm that selects a transformation with uniform probability every time [13]. The F1 score for class 0 decreased by 0.15, for class 1 by 0.4, and for class 2 by 0.7. This profound drop on the F1 scores are due to the extend that the datasets differ from each other (Section 3.6).

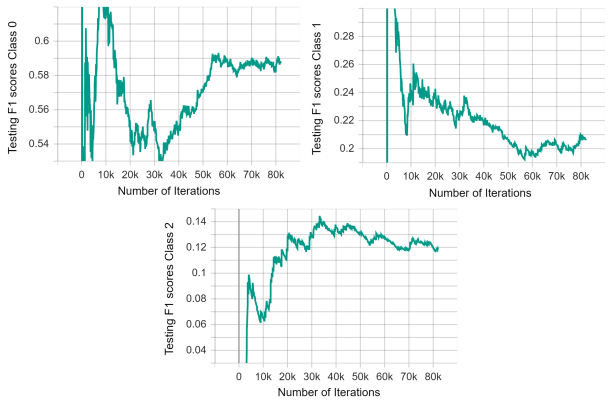

Fig. 12. The test F1 score vs iteration graph for class 0, 1, and 2

4. Conclusion

A major challenge in training a supervised model to predict knee osteoarthritis severity is the imbalance in sample size for each grading of severity. By definition there are fewer patients with more severe knee osteoarthritis, therefore it is more difficult to effectively train a model to accurately predict samples in the more severe classes. The dataset contained 2,925 images, of which only 224 images are in the most severe category 4, patients with large osteophytes and marked evidence narrowing of joint space with severe sclerosis and definite deformity of bone ends. By sample weighting, we overcame the unequal probabilities of classes to appear in a batch. After the weighting, F1 score of the most severe category 4, increased by 300% which demonstrates using sample weighting profoundly improved prediction. On the other hand, sample weighting decreased the F1 score of the most dominant category which is the healthy category 0 from 0.65 to 0.45. Therefore, we tried other options to make up for this decrease in certain categories and came up with merging some of the classes. The original dataset was not distributed equally among classes. Therefore, the globally used Kellgren-Lawrence grading system could not be used to classify the knee x-ray images. To overcome this inequality, some classes were merged which increases the performance, however, decreased the extend of the classification to only 3 categories from 5 categories. Even though sample weighting and merging enabled the classes to be equally distributed, the model should be trained with images that are as general as possible in order for the model to be not only focused on the limited dataset but to greater extents so that it can be utilized outside of the training dataset. Otherwise, the model would not be useful since it would only be successful in categorizing the images that it was trained with, which are already labeled. However, with only about 4,000 images, it is difficult to simulate a real life variability among images. Therefore, we used data augmentations, shear and horizontal flip, to generate artificial variability. As a result of this attempt, overfitting (getting extremely accustomed to the training data) was minimized by 45%. Another method that was used to decrease the overfitting was altering the epoch number which is how many times the model is trained with the training data. After a certain number of epochs, the model starts to overtrain. Therefore, we ran the code for 30 epochs and cross-examine the testing F1 scores versus iterations graphs and training-testing F1 scores versus iterations graphs for each category to find the optimum epoch number in which the testing F1 scores are the highest and the training-testing F1 scores were the lowest among all category. Furthermore, as the model trains the parameters increase in absolute value and thus the model becomes more complex. A weight decay parameter added to the cost function will result in optimization on both a simple model and also a model that has good predictive performance on the training set. The degree of weight decay is modulated with the λ hyperparameter, we found the optimum value to be 0.001. We also attempted both semi-supervised consistency learning and transfer learning in order to utilize a distinct dataset of additional poorly labelled knee X-ray images. This external dataset was not similar enough in terms of image size, classification number, and sample size to the intended training set. Therefore the performance of the consistency and transfer learning was poor. The decrease in F1 scores of each class while using the transfer and consistency compared to the traditional supervised learning provides sufficient evidence to conclude that the second dataset was not fit for the purpose. In future work, we intend to implement weighted sample loading to equalize proportion of outcome categories in each batch in order to better assess the performance of consistency and transfer learning as this method provided increased performance in the traditional supervised learning workflow.

REFERENCES

[1] K. L. Ong, M. Runa, E. Lau, and R. Altman, "Cost-of-illness of knee osteoarthritis: potential cost savings by not undergoing arthroplasty within the first 2 years," *ClinicoEconomics and Outcomes Research*, vol. Volume 11, pp. 245–255, Mar. 2019.

[2] R. Mahum, S. U. Rehman, T. Meraj, H. T. Rauf, A. Irtaza, A. M. El-Sherbeeny, and M. A. El-Meligy, "A novel hybrid approach based on deep CNN features to detect knee osteoarthritis," *Sensors*, vol. 21, p. 6189, Sept. 2021.

[3] J. C. Mora, R. Przkora, and Y. Cruz-Almeida, "Knee osteoarthritis: pathophysiology and current treatment modalities," *Journal of Pain Research*, vol. Volume 11, pp. 2189–2196, Oct. 2018.

[4] P. Chen, "Knee osteoarthritis severity grading dataset," *Mendeley Data*, 2018.

[5] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Lecture Notes in Computer Science*, pp. 345–359, Springer Berlin Heidelberg, 2005.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[7] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[9] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019.

[10] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," 2020.

[11] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019.

[12] Kaggle, "Osteoporosis knee x-ray dataset," 2021. Available at https://www.kaggle.com/datasets/stevepython/osteoporosis-knee-xray-dataset.

[13] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," 2019.