Resampling Techniques for Multi-label Imbalanced Classification of Medical X-Ray Data

Anusha Umesh

Master in Computer Science
University of Ottawa, Ottawa, ON K1N 6N5, Canada
aumes019@uottawa.ca

Abstract. Chest X-ray imaging technology is an important approach to imaging diagnosis in today's world, due to the developments in technology. Increasing interest in the use of deep learning techniques in this context is being driven by the also increased availability of labeled X-ray images. Assistance based on machine learning or, in particular, on deep learning methods can also be provided which reduces the burden of experts to some degree. Our goal is to expand the overall understanding of the various existing approaches as well as their use for chest X-ray classification by using the NIH Chest X-ray dataset and a structure that includes non-image data (patient ID, patient age, and gender) into the classification process. In this paper we focus on a state of the art deep learning model and compare two resampling strategies to deal with the problem of unbalanced multi-label data to predict multiple labels on images of chest x-rays from the NIH Chest X-ray dataset. From Resampling we can see that the evaluation parameters such as F1-score Precision and Accuracy have shown significant increase in many labels especially in Oversampling case.

Keywords: Chest X-ray · Lung disease · MobileNet · Resampling.

1 Introduction

The most frequent and economical medical imaging method is chest X-rays. A chest X-ray medical examination is feasible. However, in certain circumstances, it is harder to predict from these images than with computed tomography (CT) imaging [1]. Another difficulty is related to discovering clinically relevant findings due to the shortage of a large dataset that is publicly available with experts annotations.

In computer vision, deep learning has already proven its capacity to categorise pictures with high accuracy [18]. In deep learning, the field of medical image processing is a hot topic [15,3]. In the medical area, however, one key issue is the availability of huge datasets with unbalanced problems, which can lead to an undesirable performance in any model when focusing on the minority and interesting class cases. The term "unbalanced dataset" refers to a collection of data that is not evenly distributed.

Lung disorders pose a significant threat, particularly in emerging and low-middle-income nations, where millions of people live in poverty and are exposed to pollution. According to WHO estimates, about 4 million people die prematurely each year as a result of illnesses caused by home air pollution, such as asthma and pneumonia [13]. As a result, actions must be taken to minimise air pollution and carbon emissions. It's also critical to put in place effective diagnostic technologies that can help diagnose lung problems. A new coronavirus illness 2019 (COVID-19) [18] has been causing major lung damage and breathing issues since late December 2019. Furthermore, pneumonia, a kind of lung illness, can be caused by the COVID-19 causal virus or another viral or bacterial infection.

Many academics have looked at machine learning approaches for predicting diagnostic information from X-ray images of lung illnesses. In this paper we focus on the NIH chest X-ray picture dataset which is obtained from [22] and is completely open source. This research introduces a new method that is successfully used on the above-mentioned dataset to classify lung disease. A novel algorithm where adapted resampling techniques are employed to improve the performance when forecasting each lung disease from X-ray pictures is the research's key contribution.

This paper is organized as follows. Section 2 covers some relevant work on X-ray image categorization or resampling approaches for lung X-ray images. Section 3 presents a comprehensive examination of the implemented dataset. In Section 4 we discuss the study's methodology. Section 5 describes the experimental assessment. The results and associated discussion are provided in Section 6, while Section 7 concludes the paper.

2 Literature Review

2.1 Resampling for tackling the imbalance problem

When the distribution of the input variables is unbalanced, we say that we face an unbalanced class problem. To implement resampling techniques, introducing special purpose learning methods, or using post-processing strategies are the three main ways to manage an imbalanced environment [20].

Among the three types of solutions, resampling approaches are the most common way to solve imbalanced domain learning problems. Resampling is a data pre-processing technique for forcing the classifier to focus on the most essential examples by altering the original training data distribution. The two primary resampling techniques [20] are (i) undersampling, which involves removing examples from the majority class, and (ii) oversampling, which involves adding more examples from the minority class.

Oversampling is a data pre-processing technique that forces the classifier to focus on the most important examples by changing the original distribution of the training data [18]. The Synthetic Minority Oversampling Technique (SMOTE) algorithm [5] is an oversampling method for creating unique synthetic examples. SMOTE works by choosing examples in the feature space that are

nearest neighbours, drawing a line between them, and creating a new synthetic example at a random position along that line. Another Oversampling Technique that builds on the methodology of SMOTE is the adaptive synthetic sampling technique, ADASYN algorithm [10]. ADASYN accomplishes oversampling using a weighted distribution where examples are created for each observation belonging to the minority class. There are also resampling methods that carry out undersampling of the majority class. A simple, yet effective method that removes examples from majority class, with or without replacement is called Random undersampling [18]. Cluster centroids [23] is a technique that uses the cluster centroid of a K-means algorithm [12] to substitute a cluster of instances, with the number of clusters determined by the amount of undersampling.

2.2 Resampling Strategies in the medical context

Özçift, A. [17] strategy consists of two parts: (i) To identify important characteristics from the cardiac arrhythmia dataset, a correlation-based feature selection technique is utilised. (ii) To test the effectiveness of the proposed training approach, the RF machine learning algorithm is used to analyse the performance of selected functions with and without simple random sampling. The collection comprises 452 samples from 14 different kinds of arrhythmias, with sample sizes of fewer than 15 in eleven of them. The classifier's accuracy was determined to be 90.0 percent, which is a very outstanding diagnosis efficiency for cardiac arrhythmia. Valdovinos, et al [21] suggest a variation of resampling approaches that involves selecting examples while considering the original training set's class distributions. They effectively create classifier ensembles using bagging, AdaBoost, and Arc-x4. Bagging beats AdaBost and Arc-x4 in experiments using six actual data sets from the UCI Machine Learning Database Repository [16], three of which relate to distinct medical areas (Heart, Liver, and Pima). Fahad Alahmari [1] examined the effectiveness of many oversampling and undersampling techniques. To study the effects of data imbalance in autism spectrum disorder (ASD) applications, the author used three different resampling strategies: RUS, ROS, and SMOTE. Random Forest and Nave Bayes experiments generated actual results. When the considered dataset was re-sampled by ROS, empirical results obtained through experiments with Random Forest and Nave Bayes classification benefits showed advantages in terms of ROC, sensitivity, specificity, and precision evaluation values.

The conventional imbalance problem is not the same as the multilabel imbalance problem. The classification is mostly divided into four primary groups. I) Binary Classification: In Binary Classification, an instance is classified as either belonging to one of two classes. II) Multi-class classification occurs when the target variable has more than two unique values. III) Multi-label Classification: In this form of classification issue, the target variable has many categories, each of which has just two unique values. IV) Multidimensional Classification is a multi-class classification extension in which each attribute of the target variable is non-binary [5]. Each example in a multilabel imbalance problem is connected to several labels, rather than having single type of label like in binary

or multiclass classification tasks [24]. In most multi-label learning approaches, the number of positive training examples vs. each class label is considerably fewer than its negative equivalents, which might lead to performance loss [16].

Some solutions have been put forward to address the multilable imbalance problem. For instance, to decouple multilabel datasets, Zhou, Shuyue, et al. [24] presented a bidirectional resampling approach. They began by decreasing label overlap by establishing decoupling termination criteria. Then, by combining oversampling and undersampling, the loss of example information and overfitting were reduced. The cases with the least effect on minority labels were chosen to resample by assessing their minority labels. The algorithm's performance was tested on seven standard multi - label datasets, with a focus on datasets having a high probability of majority and minority labels. Charte, Francisco, et al. [4] demonstrated how the occurrence of unbalanced labels, a unique feature of multilabel datasets (MLDs). MLD are usually associated simultaneously to two or more labels, may have a significant influence on the behaviour of resampling methods. In order to achieve this aim, a metric called SCUMBLE is developed, and its utility is empirically verified. This measure's applicability has been proved empirically against six MLDs and two resampling methods. The correlation study revealed that the SCUMBLE metric may be used to predict whether resampling will be beneficial for a certain MLD or not.

2.3 Imbalance problem in the medical context

Due to recent improvements in the automation of medical systems, a significant quantity of health record (HR) data has been recorded. But, reviewing HR data is not always straightforward, especially when the number of people with a specific disease is low in contrast to the whole population. This situation is called the imbalanced data problem [20]. In this field, there are currently no automated techniques for monitoring and evaluation, the most frequent and economical medical imaging method is chest X-rays. A chest X-ray medical examination is feasible, and in certain circumstances, it is easier to predict than with computed tomography (CT) imaging [22]. For categorization based on lung segmentation Chen, Bingzhi, et al. [6] proposed the TSCN, a two-stream collaborative network which is used for multi-label Automated chest X-ray CXR image. They used U-Net to train a strong lung segmentor, which then is used to extract the lung field from the original CXR picture. Furthermore, for additional feature learning, they integrate data that gives context to images and the lung field using a two-stream feature fusion method. With two types of picture inputs and two-stream structure utilised simultaneously to perform adaptive two-stream feature subset selection and optimization, a unique self-adaptive weighted fusion technique is proposed. Extensive trials on the ChestX-ray14 dataset show that the suggested approach outperforms the current state-of-the-art baselines.

Filice, Ross W., et al. [7] annotated and adjudicated a huge collection of NIH chest radiographs that will be made available with the objective of stimulating innovation. They looked at the benefits of employing AI models to generate comments for review, because picture tagging is time-consuming and tedious.

The use of this machine learning annotation (MLA) approach seems to speed up our annotation process while sacrificing specificity. Antin, B., et al. [2] began by doing a quick analysis of the data using unsupervised approaches. They choose deep learning approaches for data exploration because they can typically infer complex characteristics through training. As a result, they believe that any filters based on unsupervised techniques may be captured by the neural network's complexity. Finally, it was determined that the problem lies with the network's learnt features rather than the final linear layer. They believe that by providing more cases of Pneumonia, the network will be able to learn more important characteristics.

COVID-19 can lead to serious pneumonia, and chest X-rays (CXR) and computed tomography (CT) scans are the most common imaging diagnostic procedures for pneumonia. Pereira et al. [18] presented a categorization scheme that took the following factors into account: i) a multi-class classification; ii) a hierarchical classification, because pneumonia has a hierarchy. The schema uses resampling methods such as oversampling and undersampling to rebalance the distribution of classes. They observed that a CNN model that has previously been trained extracts characteristics from categorization schema. In the RYDLS-20 hierarchical classification scenario, the proposed method gave an F1-Score of 0.89 for COVID-19 identification.

To process multilabel imbalance datasets, data and algorithms are used unlike the traditional unbalanced data processing methods. Deep learning techniques have recently become the most common technology for image classification detection, thanks to the growth of image classification detection. For example, Chen, Bingzhi, et al. [6] proposed U-Net to quickly retrieve a relative precise lung field as the input of the local branch of the proposed TSCN, to reduce the effect of noise regions in CXR image. Several studies have been conducted of using resampling to solve the multilabel data imbalance such as Zhou, Shuyue, et al. [4] and Charte, Francisco, et al. [4]. However, there is very few research in resampling multilabel imbalance data using the x-ray dataset and images which we try to solve in this paper. The proposed approach, which is based on the above-mentioned works, classifies the search into two ways. First, apply resampling then build a two-layer predictive model which will evaluate all the lung illnesses and group them together under the term "lung disease." It will discover the distinction between "lung disease" and "no finding" as a first step. Second, it will dig deeper into the different types to anticipate all the lung problems that exist.

3 Data Pre-Processing

3.1 Description

The dataset utilised to train and assess the proposed approach is NIH chest x-ray dataset. The ChestXray dataset, which includes both pictures and structured data, was used in this study. There are 112,120 pictures in the image dataset, including 30,000 patients. Multiple scans may be available for certain patients

as same patients have different follow-ups, which will be taken into account. All pictures were created at a resolution of 1024 × 1024 pixels [22] as shown in Figure 1.

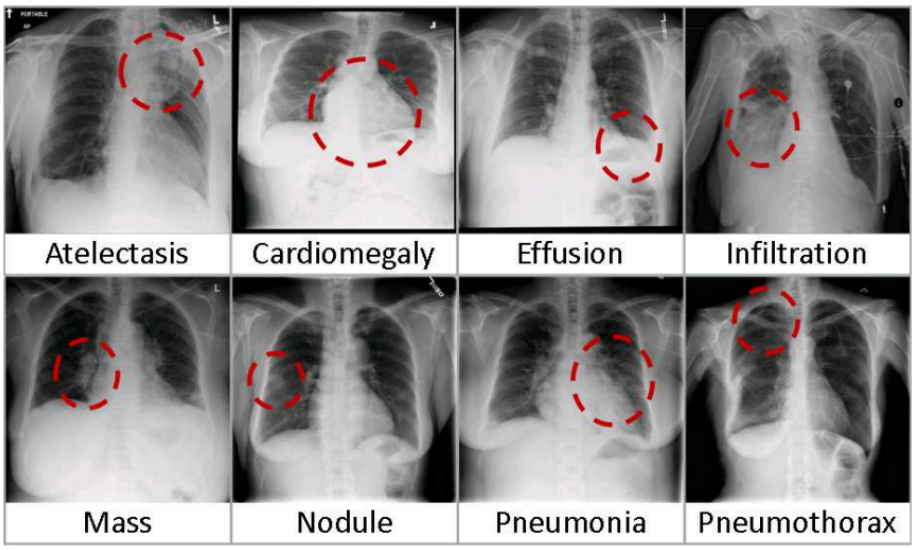

Fig. 1. Eight common thoracic diseases observed in chest X-rays

This dataset includes the following features:

- Image Index: File name
- Finding Labels: Disease type (Class label)
- Follow-up number
- Patient ID
- Patient Age
- Patient Gender
- View Position: X-ray orientation
- OriginalImageWidth
- OriginalImageHeight
- OriginalImagePixelSpacing-x
- OriginalImagePixelSpacing-y

This is a multi-label problem where the target class may be labeled as "No findings" or, alternatively, may have multiple findings which correspond to different lung diseases that may be present simultaneously. Overall, there are 15 classes: 14 different diseases, and one class for "No findings" which represents a healthy patient. Thus, each image can be classified as "No findings" or one or more disease classes:

1. "No findings"
2. 14 diseases:
   - Atelectasis
   - Cardiomegaly
   - Consolidation
   - Edema
   - Effusion
   - Emphysema
   - Fibrosis
   - Hernia
   - Infiltration
   - Mass
   - Nodule
   - Pleural-thickening
   - Pneumonia
   - Pneumothorax

The authors utilised Natural Language Processing to text-mine illness categories from the accompanying radiological data to produce these labels. The labels are anticipated to be $> 90\%$ accurate and adequate for learning without supervision. Although the original radiological reports are not available publicly [22].

3.2 Exploratory Data Analysis

In this section, a number of plots will provide some insights about the lung diseases data.

The bar graph in Figure 2 shows how many patients fall into the different categories of the labels from dataset [22]. On the x-axis, we have the 15 different label diseases and on the y-axis, the number of images associated to each label.

In this bar gragh shows the values for combination of labels. In dataset [22] we have 836 unique combinations of labels. Plotting it against the people associated with each label is shown in Figures 3 and 4.

The dataset [22] also contains information regarding the patients gender. the bar graph in Figure 5 shows the ratio between Male and Female on the dataset that are associated to certain diseases. Lastly, this figure also provides an overview of the "No Findings" label for each Male and Female category.

4 Our Proposed Solution

In this work, the algorithms are implemented using Google Colab Notebook, Tensorflow, and Keras. The implementation processes are described below.

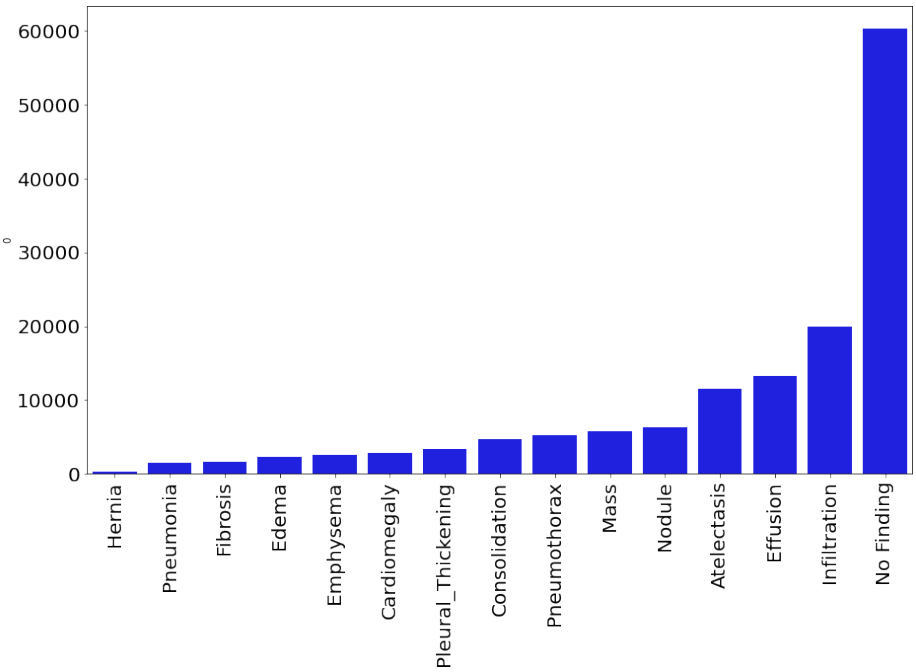

Fig. 2. Frequency values of unique labels in dataset.

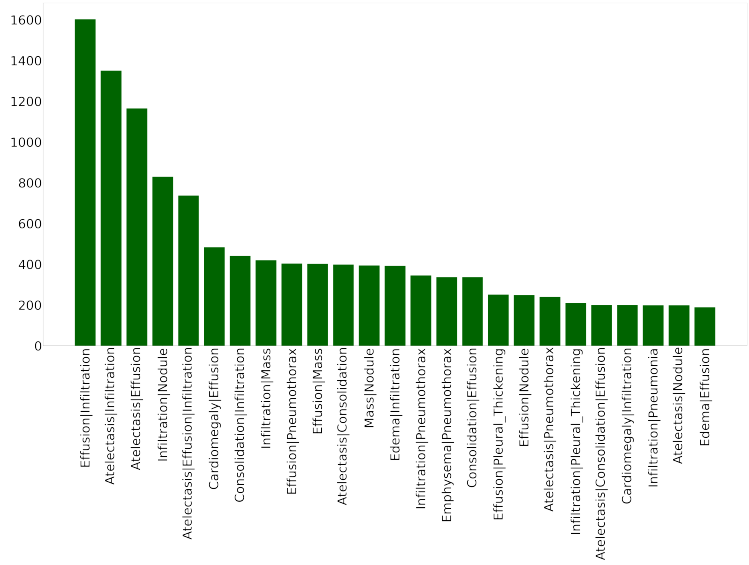

Fig. 3. First 25 combinations of labels in dataset.

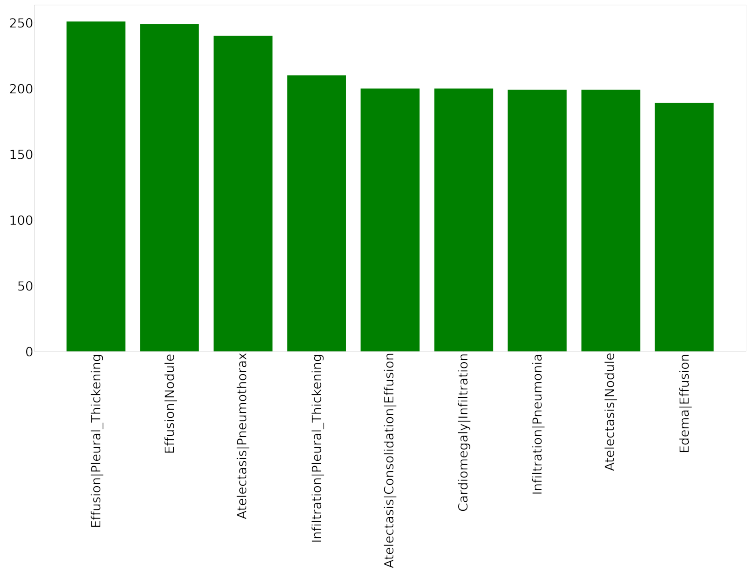

Fig. 4. Expanding values from 16th label from above 25 combinations of labels.

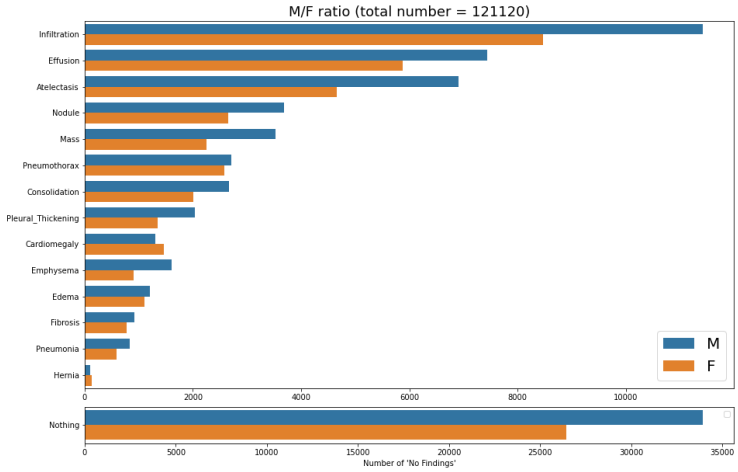

Fig. 5. Male and Female ratio of labels in dataset.

4.1 Data Partition

The train-test-split method is utilized to anticipate the performance of machine learning models as they are utilized to make predictions about the data that are not used to train the model. The method required the acquisition of the dataset and the separation into two subsets. The first subset is used to match the model and is called the training data. The subsequent subset is not utilized to train the model, rather the dataset input feature is given to the model, at that point the predictions are made and compared with the anticipated values. The second dataset is referred to as the test dataset.

We have divided the data into training and validation datasets a well. We split the data into subsets with 75% of the data as training data, and 25% as validation data.

4.2 Training the Images

The dataset consists of many X-ray images. To enhance images while training the model, using the ImageDataGenerator class in Keras. The ImageDataGenerator class in Keras provides three alternative methods for loading picture datasets into memory and generating batches of augmented data. The data augmentation function is provided by a Keras deep learning package, which performs augmentation automatically while training the model. Many image alteration procedures, such as flips, rotate at various angles, shifts, zooms, and others, are included in these transformations [9]. The function flow_from_directory() creates batches of enhanced data in a given directory. The images in the directory are organised into subdirectories based on their class. The following are the preprocessing processes that were employed in this study for images are:

- At first rescale all images for the purpose of reducing size leading to faster training stage.
- All the images are transformed to gray, and are mutually conducted for various models.
- Define some of the specific feature labels.

4.3 Implementation of MobileNet

MobileNets are built primarily from depthwise separable convolutions [11] which requires only one-eighth of the computation cost. MobileNet has got two hyper-parameters: width multiplier and resolution multiplier [19]. Although the base MobileNet architecture is already small and has low latency, many times a specific use case or application may require the model to be smaller and faster. MobileNets architecture has two Dense layers, one global max pooling layer and two dropout layers. So, the summation of total params is 3,753,601, Trainable params: 3,731,713 and Non-trainable params: 21,888 as shown in Figure 6.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 mobilenet_1.00_128 (Functio  (None, 4, 4, 1024)       3228288
 nal)

 global_average_pooling2d (G  (None, 1024)             0
 lobalAveragePooling2D)

 dropout (Dropout)           (None, 1024)              0

 dense (Dense)               (None, 512)               524800

 dropout_1 (Dropout)         (None, 512)               0

 dense_1 (Dense)             (None, 1)                 513

=================================================================
Total params: 3,753,601
Trainable params: 3,731,713
Non-trainable params: 21,888
```

Fig. 6. Model of MobileNet.

4.4 Resampling Algorithm

In the work of the imbalance problem of multilabel data, to solve the data imbalance, there are two ways of resampling methods used, Random Undersampling and Random Oversampling.

The pseudocode of Resampling-Algorithm is shown in Algorithm1.The algorithm is divided into two stages, i.e., Binary classification and Resampling techniques.The starting data preprocessing steps (1-6) are explained in subsections 4.1 4.2 4.3.
In the first stage, Binary classification is applied on each "Disease" to "No Disease" on data and it is passed on to MobileNet() classifier to calculate accuracy, precision, recall (Steps 7–18) in Algorithm 1.
In the second stage, oversampling and undersampling are applied and stored in dataRO and dataRU respectively, this data is passed to MobileNet() classifier to calculate accuracy, precision, recall (Steps 8–18) in Algorithm 1.
Finally, the results are combined to select best accuracy, F1-score for each label in step 19 in Algorithm 1.

5 Experimental Evaluation

5.1 Experimental Settings

In order to allow the reproducibility of our experiments, all the code used is freely provided to the interested researchers in the following link: `https://github.`

Algorithm 1: Resampling Algorithm

Input: NIH Dataset - Images, Labels
Output: Accuracy, Precision, Recall, F1-Score, Classification Report

```
Start;
Import dataset.csv file and Images directory;
for j in { "No findings", "Cardiomegaly","Consolidation","Edema ","Effusion
  ","Emphysema","Fibrosis","Hernia","Infiltration","Mass","Nodule","Pleural-
  thickening","Pneumonia ", "Pneumothorax" }
  do:
      Pre-process the data into binary classification using classes j and ¬ j;
      Split data into train, validation and test;
      Pre-process Image using ImageDataGenerator();
      Map Label <− Image using flow_from_dataframe();
      data <− Data of Binary classification;
      dataRu <− Random Under sample data to balance the classes;
      dataRO <− Random Oversampling data to balance the classes;
      for k in { data, dataRU, dataRO } do:
          Pass data, dataRU, dataRO to MobileNet() Classifier
          Run for 'k' number of epochs
          Finally, calculate:
              Accuracy
              Precision
              Recall
              F1-score
          Display the Classification Report
      Choose the best results of data, dataRU, dataRO;
End;
```

com/Anusha1396/Resampling-Imbalanced-X-Ray-Data. Single-label evaluations are generally different from multi-label evaluation metrics.In single-label classification simple metrics such as precision, recall, accuracy, etc, are used. There are a few ways to combine results across labels, specified by the average argument such as average_precision_score. Macro and micro-averaged F1-scores are frequently used to evaluate classification performance in multi-label categorization [8].

Accuracy

Accuracy is the ratio of the number of correctly predicted instances to the total number of predicted instances, regardless of whether the instances are positive or negative. The accuracy is calculated as follows in Equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Recall and Precision

The number of afflicted patients can be reduced with Recall and precision. These variables may be useful in predicting the onset of this lung illness.

In Equation 2, recall represents the proportion of patients properly predicted as unwell out of the total number of patients who are actually ill:

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

In Equation 3, precision refers to the percentage of patients who are properly predicted to be unwell out of the total number of patients that are anticipated to be ill:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Macro Averaged F1 score

For models with numerous classes/labels, macro-averaging is implemented; all classes contribute equally, regardless of how frequently they appear in the data set.

The F1-scores for each class/label are computed separately and then averaged in macro F1-averaging as shown in Equation 4:

$$F1_{macro} = \frac{1}{|C|} \sum_{k \in C} F1score_k = \frac{1}{|C|} \sum_{k \in C} \frac{2P_k R_k}{P_k + R_k} \tag{4}$$

We have three tags t1, t2 and t3. In case of macro averaged F1 score, we will calculate the precision, recall and the subsequent F1 scores for each of these labels (or tags). Then, we will sum of all the individual F1 scores for all the tags and divide it by the total number of samples (in this case 3 samples) [14]. When there is a very unbalanced distribution of tags, the macro averaged F1 score does not function well. It also doesn't account for the frequency with which tags appear. As a result, when it comes to multi-label classification and we have a highly unbalanced distribution of tags, the micro averaged F1 score should be our primary measure of choice because it considers tag frequency.

The micro F1-score

The micro F1-score (short for micro-averaged F1 score) is a metric for evaluating the quality of multi-label binary issues.It calculates the F1-score based on the sum of all class inputs [14] as shown in Equation 5:

$$MeanFScore = F1_{micro} = \frac{2Precision_{micro} Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{5}$$

The best micro F1-score is 1 (perfect micro-precision and micro-recall), while the worst is 0. Because each sample is given equal weight, micro-averaging underlines the common labels in the data set. For multi-label classification issues, this may be the optimal behaviour.

6 Results

Based on the accuracy of the approaches on the full dataset with binary classification without sampling, with under sampling and with over sampling can be compared as shown in Tables 1 and 2.

Accuracy and F1-Score Table

In table 1 we can see the Accuracy and F1-score of binary classification without sampling, with under sampling and with over sampling results.
In this table we can observe that most of the labels performed well in Oversampling than compared to Undersampling. But, label "Pneumothorax" results show that it performed significantly well in Binary classification than Undersampling and Oversampling. The label that outperformed in UnderSampling is "No Findings" in terms of Accuracy of 97% and label "Hernia" with F1-score of 0.54. These Classification Reports are shown below in Figure 7

Fig. 7. Label "Hernia" Classification Report

Label "Pneumonia" has shown the best accuracy in Oversampling of 98% and Label with best F1-score of Oversampling is by "Consolidation" with score 0.51. These Classification Reports are shown below in Figure 8

Precision and Recall Table

In table 2 we can see the Precision and Recall of binary classification without sampling, with under sampling and with over sampling results.

```
[42] print('test accuracy = ',mobilenet_GAP_model.evaluate(X_test,y_test, verbose=0)[1])

     test accuracy =  0.9888163805007935

[43] from sklearn.metrics import roc_auc_score
     print('roc score = ',roc_auc_score(y_test.astype(int), pred_Y))

     roc score =  0.5

     from sklearn.metrics import classification_report, confusion_matrix
     plt.matshow(confusion_matrix(y_test, pred_Y>0.02324778))
     print(classification_report(y_test, pred_Y>0.02324778, target_names = ['Healthy', 'Pneumonia']))

                   precision    recall  f1-score   support

          Healthy       0.99      1.00      0.99      2122

        Pneumonia       0.01      1.00      0.02        24

         accuracy                           0.01      2146
        macro avg       0.01      0.50      0.01      2146
     weighted avg       0.00      0.01      0.00      2146

     /usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1308: Undefine
       _warn_prf(average, modifier, msg_start, len(result))
     /usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1308: Undefine
       _warn_prf(average, modifier, msg_start, len(result))
     /usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1308: Undefine
       _warn_prf(average, modifier, msg_start, len(result))
```

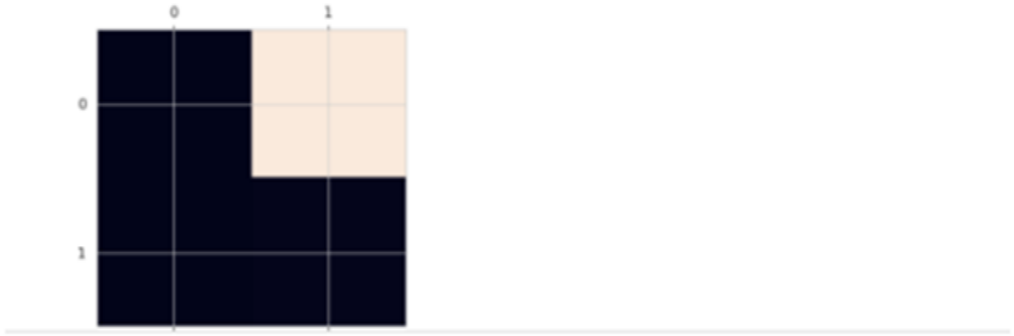

Fig. 8. Label "Pneumonia" Classification Report

In this table we can observe that most of the labels performed equally. But, label "Consolidation" Precision results show that it performed significantly well in Undersampling with score 0.47. These Classification Reports are shown below in Figure 9

Label "No Findings" has shown the best Precision in Oversampling of 0.54 These Classification Reports are shown below in Figure 10

Recall of labels have achieved score 1.0 in Binary, Oversampling and Undersampling as we can see in Table 2

Table 1. Overall Accuracy (Acc) and F1-Score (F1) Results.

Labels	**Binary_Classif.**		**Under_Samp.**		**Over_Samp.**		**Proposed Alg.**	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Atelectasis	89	0.16	67	0.06	67	0.32	67	0.32
Cardiomegaly	97	0.05	66	0.49	67	0.49	67	0.49
Consolidation	95	0.08	67	0.02	65	0.51	67	0.51
Edema	98	0.03	33	0.5	66	0.5	66	0.5
Effusion	88	0.21	67	0.48	67	0.49	67	0.49
Emphysema	98	0.04	66	0.5	66	0.45	66	0.5
Fibrosis	98	0.03	66	0.5	66	0.49	66	0.5
Hernia	99	1	62	0.54	62	0.37	62	0.54
Infiltration	82	0.26	66	0.48	66	0.26	66	0.48
Mass	95	0.02	65	0.51	65	0.07	65	0.51
Nodule	94	0.11	80	0.33	68	0.48	80	0.48
Pleural-thickening	97	0.8	65	0.5	66	0.26	66	0.26
Pneumonia	98	0.2	79	0.33	98	0.02	98	0.33
Pneumothorax	95	0.98	67	0.02	67	0.49	95	0.98
No Findings	50	0.61	97	0.06	54	0.7	97	0.7

7 Conclusion

In this work, a new resampling technique is proposed and applied to compare the multi-label dataset for detecting lung diseases from X-ray images. The code is available on GitHub link - Resampling-Imbalanced-X-Ray-Data. The new model is applied to NIH chest X-ray image dataset collected from [22]. Estimated to maximize evaluation scores such as the micro- and macro-averaged,and average labeling F1-scores are used. For the case of binary classification without resampling, the results show that F1-score was lower compared to the case of undersampling and oversampling. The best results for Acucracy and F1-Score for label "Consolidation". Experiments show that the resampling can effectively improve the classification performance of the classifier and useful to help the screening of patients in the emergency medical support services, that has been severely affected by lung diseases.

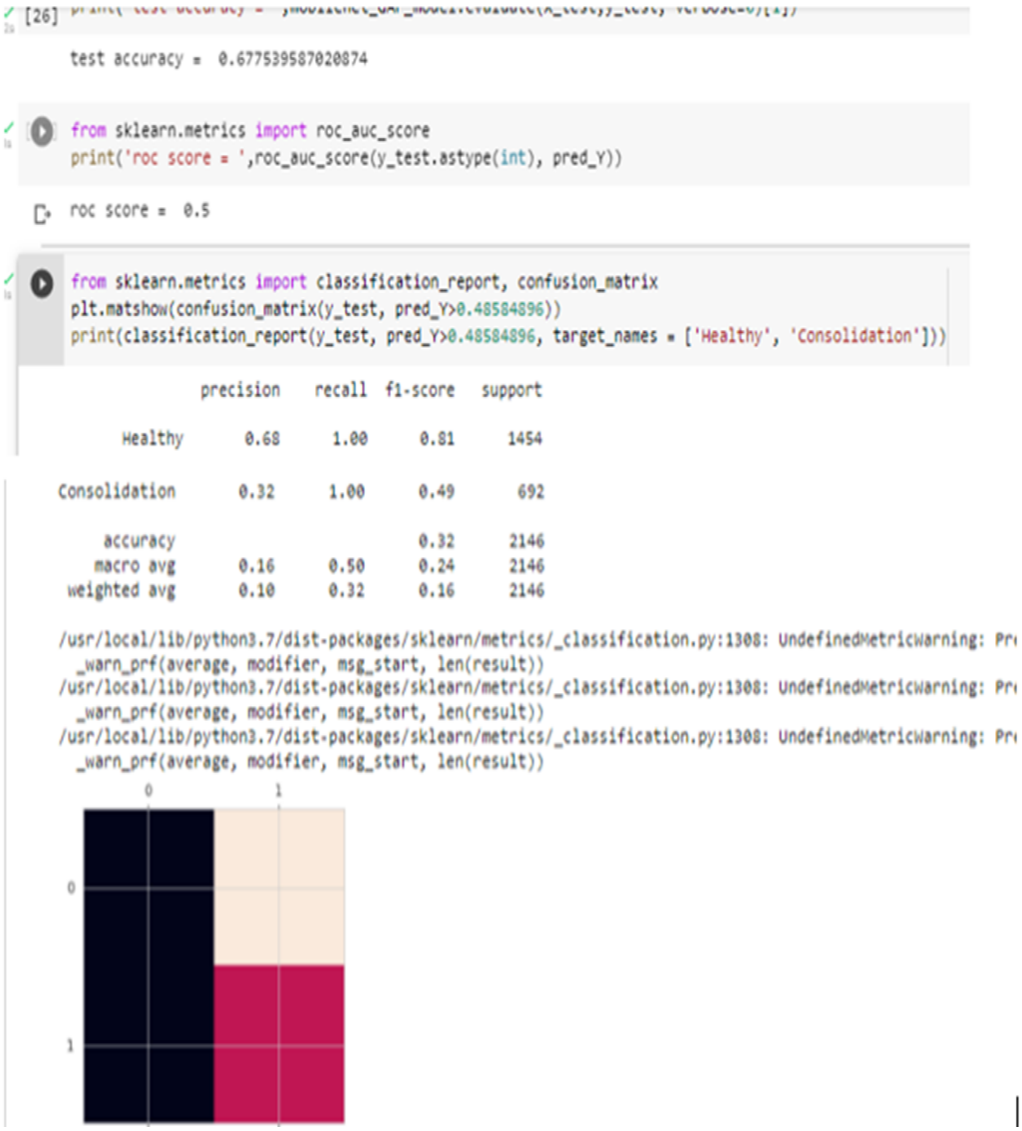

Fig. 9. Label "Consolidation" Classification Report

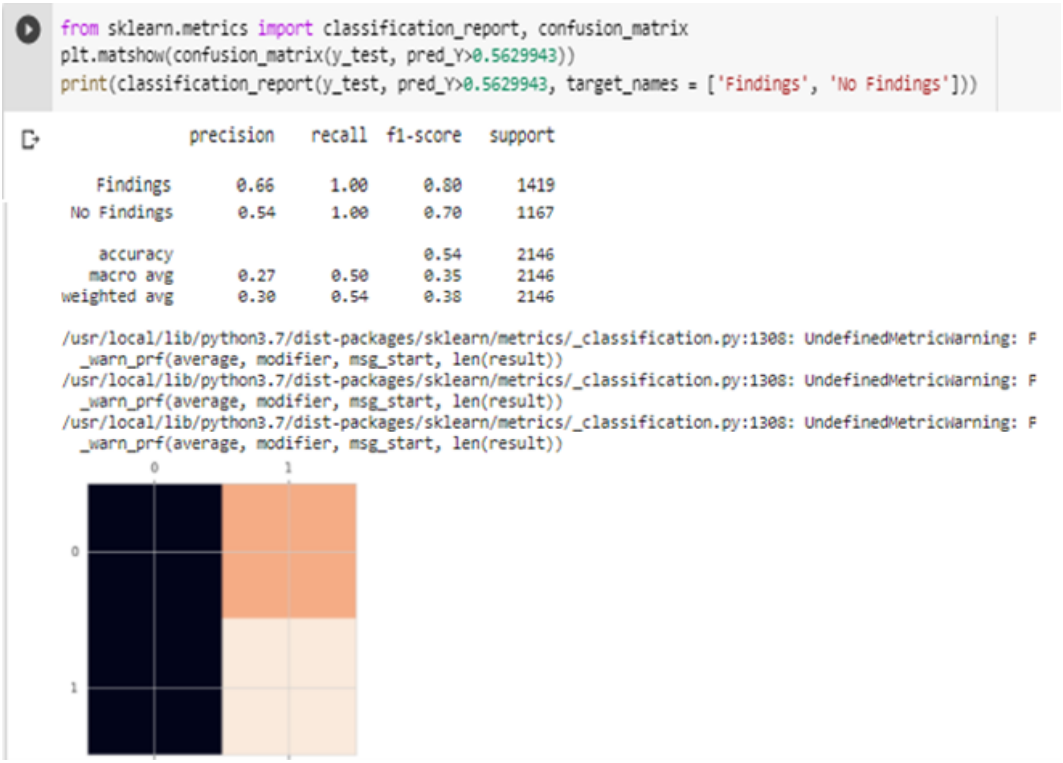

Fig. 10. Label "No Findings" Classification Report

Table 2. Overall precision (prec) and recall (rec) results.

Labels	**Binary_Classif.**		**Under_Samp.**		**Over_Samp.**		**Proposed Alg.**	
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
Atelectasis	0.12	0.25	0.23	0.03	0.3	0.35	0.3	0.35
Cardiomegaly	0.03	0.34	0.33	0.92	0.31	1	0.33	1
Consolidation	0.04	0.99	0.47	0.01	0.35	1	0.47	1
Edema	0.01	0.66	0.33	1	0.33	1	0.33	1
Effusion	0.12	1	0.32	0.93	0.32	1	0.32	1
Emphysema	0.02	1	0.34	0.96	0.34	0.67	0.34	0.96
Fibrosis	0.02	0.59	0.34	0.99	0.33	0.95	0.34	0.99
Hernia	1	1	0.37	0.99	0.37	0.36	0.37	0.99
Infiltration	0.18	0.47	0.33	0.88	0.34	0.2	0.33	0.88
Mass	0.14	0.01	0.35	1	0.38	0.04	0.38	1
Nodule	0.06	0.94	0.2	0.97	0.32	1	0.32	1
Pleural-thickening	0.09	1	0.34	0.98	0.34	0.2	0.34	0.98
Pneumonia	0.9	1	0.2	0.92	0.01	1	0.2	1
Pneumothorax	0.95	1	0.31	0.01	0.33	1	0.33	1
No Findings	0.45	0.93	0.03	1	0.54	1	0.54	1

In order to improve the accuracy of automated chest X-ray diagnostic systems, future research will involve the use of image data augmentation techniques such as colour space augmentations, kernel filters, feature space augmentation, and so on.The extraction of other feature sets may be also experimentally tested into our proposed classification schema. In the future, the suggested novel resampling approach might be used on X-ray pictures of suspected COVID-19 patients to determine whether or not they have COVID-19-related pneumonia.

References

1. Alahmari, F.: A comparison of resampling techniques for medical data using machine learning. Journal of Information & Knowledge Management **19**(01), 2040016 (2020)
2. Antin, B., Kravitz, J., Martayan, E.: Detecting pneumonia in chest x-rays with supervised learning. Semanticscholar. org (2017)
3. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9**(1), 1–10 (2019)
4. Charte, F., Rivera, A., del Jesus, M.J., Herrera, F.: Concurrence among imbalanced labels and its influence on multilabel resampling algorithms. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 110–121. Springer (2014)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Chen, B., Zhang, Z., Lin, J., Chen, Y., Lu, G.: Two-stream collaborative network for multi-label chest x-ray image classification with lung segmentation. Pattern Recognition Letters **135**, 221–227 (2020)
7. Filice, R.W., Stein, A., Wu, C.C., Arteaga, V.A., Borstelmann, S., Gaddikeri, R., Galperin-Aizenberg, M., Gill, R.R., Godoy, M.C., Hobbs, S.B., et al.: Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset. Journal of digital imaging **33**(2), 490–496 (2020)
8. Fujino, A., Isozaki, H., Suzuki, J.: Multi-label text categorization with model combination based on f1-score maximization. In: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II (2008)
9. Gulli, A., Pal, S.: Deep learning with Keras. Packt Publishing Ltd (2017)
10. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. IEEE (2008)
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
12. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery **2**(3), 283–304 (1998)
13. Jiang, X.Q., Mei, X.D., Feng, D.: Air pollution and chronic airway diseases: what should people know and do? Journal of thoracic disease **8**(1), E31 (2016)
14. Kar, S., Maharjan, S., López-Monroy, A.P., Solorio, T.: MPST: A corpus of movie plot synopses with tags. In: Proceedings of the Eleventh International Con-

ference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1274
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
16. Mohammed, R., Rawashdeh, J., Abdullah, M.: Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS). pp. 243–248. IEEE (2020)
17. Özçift, A.: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in biology and medicine **41**(5), 265–271 (2011)
18. Pereira, R.M., Bertolini, D., Teixeira, L.O., Silla Jr, C.N., Costa, Y.M.: Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. Computer Methods and Programs in Biomedicine **194**, 105532 (2020)
19. Sinha, D., El-Sharkawy, M.: Thin mobilenet: An enhanced mobilenet architecture. In: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). pp. 0280–0285. IEEE (2019)
20. Sun, Y., Wong, A.K., Kamel, M.S.: Classification of imbalanced data: A review. International journal of pattern recognition and artificial intelligence **23**(04), 687–719 (2009)
21. Valdovinos, R.M., Sánchez, J.S.: Class-dependant resampling for medical applications. In: Fourth International Conference on Machine Learning and Applications (ICMLA'05). pp. 6–pp. IEEE (2005)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
23. Zhang, Y.P., Zhang, L.N., Wang, Y.C.: Cluster-based majority under-sampling approaches for class imbalance learning. In: 2010 2nd IEEE International Conference on Information and Financial Engineering. pp. 400–404. IEEE (2010)
24. Zhou, S., Li, X., Dong, Y., Xu, H.: A decoupling and bidirectional resampling method for multilabel classification of imbalanced data with label concurrence. Scientific Programming **2020** (2020)