POLITECNICO
MILANO 1863

ML techniques for X-Ray Chest Image Classification

Author: Vicente Castro

Abstract

This report explores different ML techniques for the classification of chest X-ray images. In specific, it approaches the problem first from a traditional ML perspective, presenting common useful features and classifiers. Later, established Deep Learning techniques are discussed, along with modern techniques as transfer learning and fine-tuning. Finally, a comparison with a hand-crafted CNN is done trained network

Professor: Valentina Corino

Tutor: Guadalupe Garcia

Course: Applied AI for Biomedicine

September 7, 2023

Contents

1 Introduction

Artificial intelligence (AI) has the potential to greatly enhance the field of medical image recognition, offering numerous benefits to patients, healthcare providers, and researchers. Specially in the biomedical sector, the wide-spread usage of AI can be beneficial to, not only reach more people in remotes locations, but also to personalised and refined the prognosis and treatment of diseases.

This project presented an exploration of a AI techniques for the classification of chest X-ray images. This AI approach for the analysis of X-ray images is vastly studied, as the potential to standardise the procedure of analysis could improve both the velocity and the precision of disease detection.

In this report, a detailed description of the different approaches used in the path and present can be found. Furthermore, it offers a data-driven justification for the choices and assumptions encountered. The results obtained are presented together with valid interpretations and, finally, a comparison of performances is made among the different proposed models. The report concludes with a summary of the findings and recommendations for future research.

Apart from performance metrics, Explainable AI (XAI) techniques were explored to aid the understanding of the decision-making process of this black-box models. This is a specially relevant step in DL, and for medical application, as explainability offers not only re-assurance of the result, but can also communicate, in a more confident way, to untrained eyes.

2 Material and Methods

2.1 Dataset Description

The dataset comprises $15,470$ chest X-ray images obtained from various anonymous patients. These images are accompanied by a `.csv` file containing labels categorising the images into three distinct classes related to validated diseases: *Tuberculosis (T)*, *Pneumonia (P)* or *No-disease(N)*.

Furthermore, it is possible to extract patient information from the image filenames. This allows to assign a unique `patient_id` to each image and, in some instances, discern whether two or more images originate from the same patient. This process yields a total of $12,086$ unique patients, with $8,702$ of them having only one image and the remaining $3,384$ having two images each. It is relevant to note that patients with multiple images all share the same label.

When analysing the labels provided with the `.csv` file, the distribution on Figure 2 is discovered. This distribution remains consistent when considering patients instead of individual images. The pronounced **class imbalance** represents one of the primary challenges in our classification task, as it naturally introduces a bias towards the majority class (N) when training machine learning algorithms.

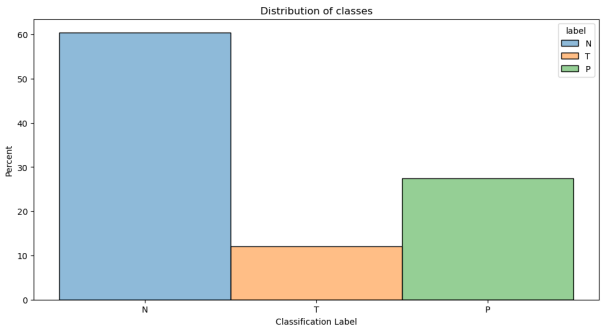


Figure 2: Distribution of samples from the three classes

All the images in the dataset are grayscale representations of the chest area of patients, typically spanning from the shoulders to the lower abdomen. Upon visual inspection it is possible to discover common perturbations affecting the dataset, specifically, three distinct types of perturbations were studied (as presented in Figure 3). Further discussions on these challenges and the proposed solutions for each perturbation can be found in the forthcoming "Image Processing" section of this report.

The first perturbation, the ***added noise***, involves the addition of noise into the images. This noise may originate from diverse distributions, and no single consistent pattern has been identified. The extent of noise varies as well, with some images exhibiting only minor perturbations, while others are nearly unrecognisable (see Figure 4).

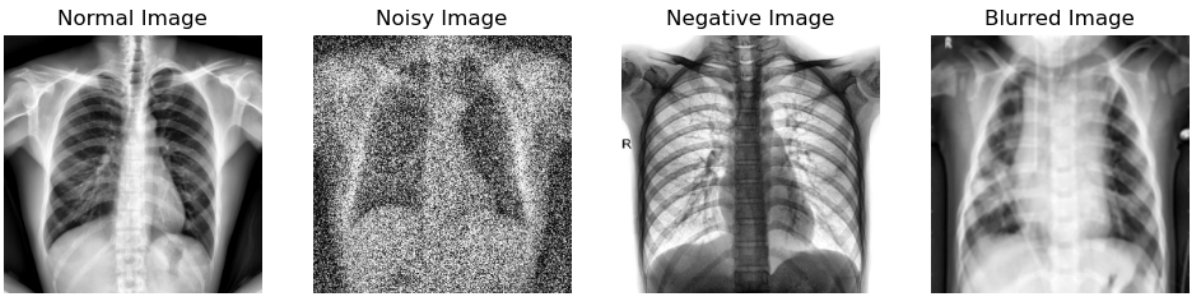


Figure 3: Common perturbation present in the dataset

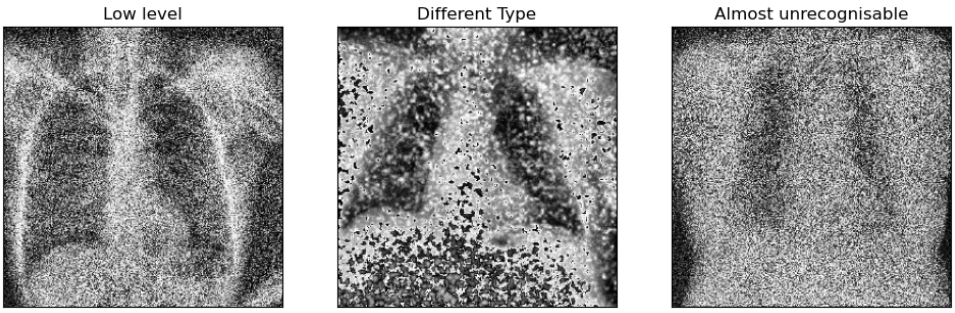


Figure 4: Different types of Noise

The second alteration observed is the ***image inversion***, where certain images are inverted with respect to the traditional X-ray colour scheme. The key challenge posed by this perturbation is the accurate identification of the issue, with the solution being the *bitwise* inversion of the affected image.

The third perturbation is the ***blurring*** presented in some of the images, also applied at different scaled. Different solutions for *sharpening* these images were explored and will be discussed in subsequent sections.

2.2 Image Preprocessing

This section described the preprocessing steps applied to the data in order to normalise it. Each of these presented technique answers to a identified data problem, already discussed or presented in this section. Furthermore, it will be later discussed how this techniques may be more relevant for the *Feature-based models*.

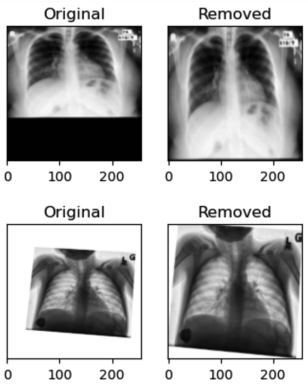
It is relevant to notice that all the hyper-parameters discussed in this section were calculated only from a *training split* taken from the data. The main parameter, the *threshold*, to decide whether an image was affected by any of the perturbation was calculated from the distributions in Figure 10b.

Framed images

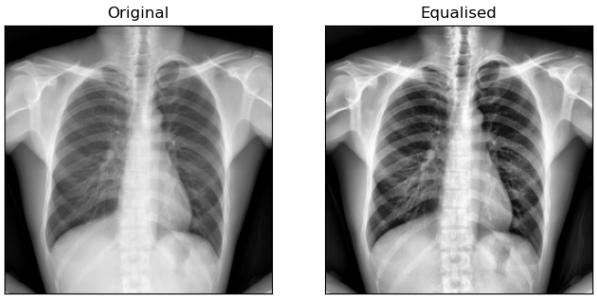
The first identified problem comes from a small subset of the dataset presenting "frames" or static bands at the sides (see Figure 5a). The algorithm to clean these images is rather simple: each row and column, starting from each side of the box is checked and removed if its *standard deviation* is less than a threshold. When a row/column does not meet this requirements, the algorithm stops.

Histogram Equalisation

Histogram Equalisation extends the pixel's intensity range from the original range to 0 to 255, traditionally improving the contrast of the image. This method was applied with a localised version called `CLAHE` (Contrast Limited Adaptive Histogram Equalisation). This alternative is more robust against loss of information due to over-brightness, as it divides the image into small blocks which are then equalised individually. The average result can be observed in Figure 5b.



(a) Framed images from the dataset



(b) Equalised version of the image

Noise reduction

Different detection methods for noisy images were analysed to detect in a robust way the noise presented on the image. Finally, the estimator used was based on the *standard deviation* of the pixel values at the corners and at the centre of the image. Gaussian blurring and post sharpening was applied when a certain threshold was met.

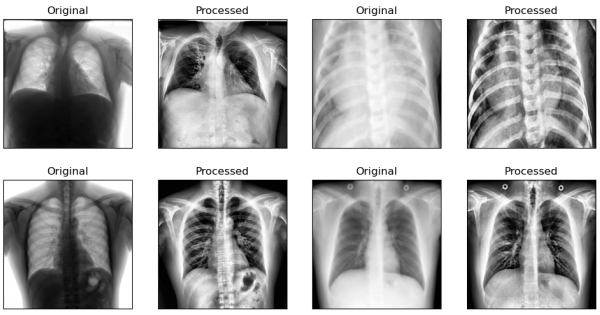
It is relevant to mention that, given the different distributions and scales of noise presented on the images, there was no method that completely succeeded on the removal of noise.

Negative Image Inversion

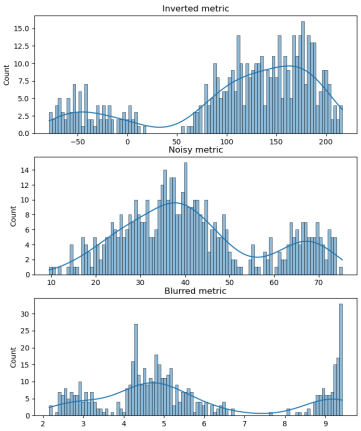
In order to detect negatives, the idea of different section of the image was also used. For this statistic, the mean of the pixel values at the center minus the top corners was used to make a decision. The logic behind this approach is based on what should normally be the position of white pixels on the image.

Image Sharpening

The sharpening of the images was done using a *Laplacian* sharpening. This method consist in computing the second derivative of the image pixels and subtract this values to the original image. The detection was also aided by the different crops, computing the *variance* of the pixels in the *Laplacian*.



(a) Results of the preprocessing pipeline



(b) Threshold for each of the proposed statistics

The final pipeline used for the preprocessing was driven by the performance of utilising in different order each of the discussed steps. Considering this, the taken approach was (1) Cut frames and equalise the image, (2) Detect noise and cancel it, (3) Detect inversion of the image and (4) Sharpen the image. The final results of this method can be seen in Figure 10a.

2.3 Traditional ML

In traditional machine learning, before deep learning era, one of the most relevant steps is the *Feature Engineering* of the problem. Often a process requiring deep understanding of the semantics of the data

and the task aiming to solve with it. The general pipeline of feature engineering can be divided in three major steps: (1) Feature Extraction, (2) Feature Selection / Reduction and (3) Model training. In the following paragraph this approach is described.

(1) Feature extraction

As images are spatial related objects, relevant features try to capture this kind of relations. Broadly, in the field of computer vision, three families of features can be defined, based on what on the image activates them.

For each of the following families, several features were studied and discard depending on the how functional they were for the dataset. Special focus must be made to the *texture* features, as is the traditional group with best results in this kind of applications.

1. ***Geometric features***, characterising the shape of the bodies on the image. Among the features analysed are:
   (a) The four Flusser Moments (*discarded*)
   (b) The seven *Hu Moments*, applied on the binary enclosed image.
   (c) Fourier Descriptors, taken from the Fourier transform of the image, 32 descriptors were kept.
2. ***Pixel Intensity***, characterising the distribution of the pixel values on the image.
   (a) Statistical Features: Mean, standard deviation, kurtosis, skewness
   (b) Gabor Features, calculated from applying different convolution of Gabor-filters with the image.
3. ***Textures***, probably the most widely used and rich family of features. They are based on complex local-structures associated with the gradients or pixel patterns.
   (a) Local Binary Patterns (LBP)
   (b) Histogram of Oriented Gradients (HOG) (*discarded*)
   (c) Haralick Features, measuring correlation values on the image.

Most of the hyper-parameters for this set of features were studied based on how much *variance* and *separability* they carried. This is a natural approach as, in the context of X-ray images, the dataset may seem too similar under features (the i.e: silhouettes of the chest are similar in most humans).

Furthermore, one of the most traditionally used features, *Histogram of Gradients*, proved not to be fitted for this dataset. This can be explained by the amount of noise present in the images and how this feature reacts to it. In Figure 7 both the original image and the calculated histograms can be seen, this provides a good visual example of how the feature is not able to compute meaningful values in presence of noise.

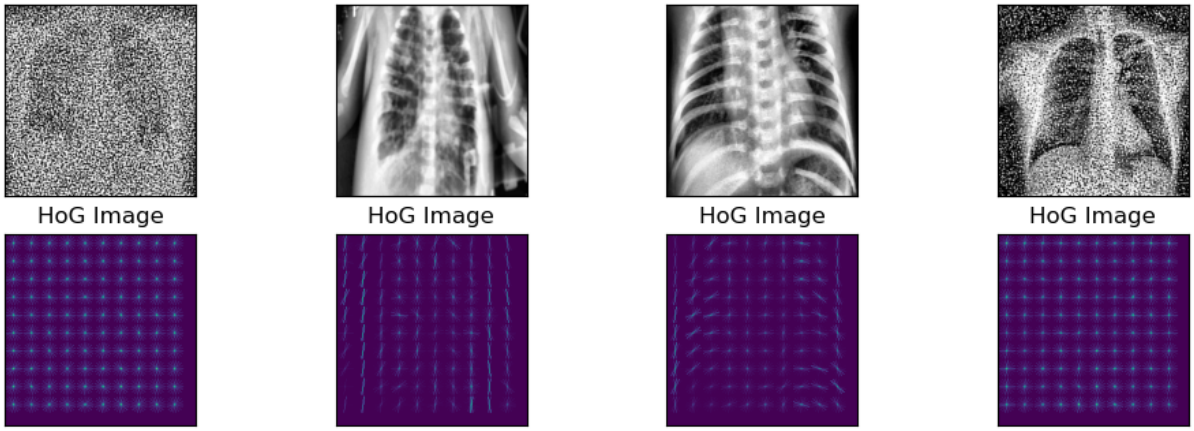


Figure 7: HoG calculations on sample images

The final number of features, under these assumptions is 247 per image.

(2) Feature selection

Feature selection procedures are based on the idea of obtaining the best set of minimal features for the dataset. Among different techniques to accomplish this a difference can be made between iterative processes (i.e: *Backward Feature Elimination*), based on removing/adding the best next candidate feature given a targeted classifier, and statistical approaches, where the reduction is made without a classifier and trying to keep most of the information carried by the features.

The analysis was centred around the last idea, how to keep the major part of the information, while removing less relevant features. For this, two ain approaches were used: **Variance Thresholding** (VT) and **Principal Component Analysis** (PCA).

VT drops features whose variance falls below a specified threshold, while PCA reduces the dimensionality of a large dataset by projecting it onto a smaller set of principle components that explain the variance in the data. Later the kept or projected features can be used to train the classifier. In both cases, the selection of the final number of features was based on maintaining a minimal amount of explained variance in the final features sets.

Finally, to maintain some posterior explainability on the features, the reduction technique was applied to the features as grouped by the discussed families. This procedure can be seen in Figure 8, where reducing the geometrical features from 132 to 96 is the optimal reduction according to the expected explained variance.

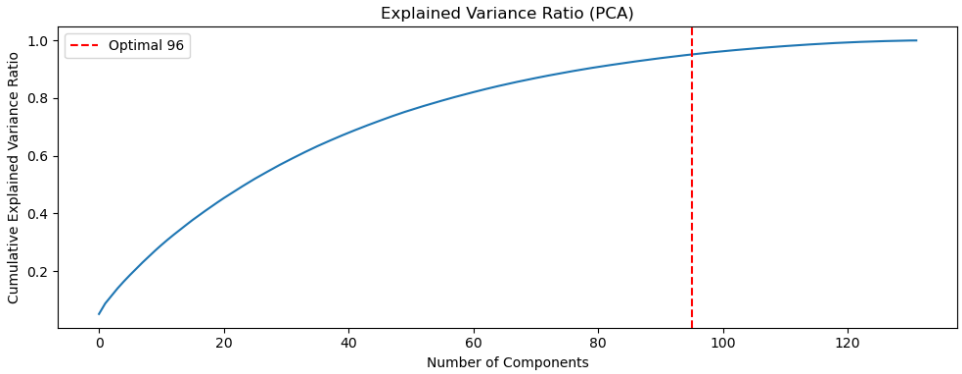


Figure 8: Optimal number of components for Geometry family

(3) Modelling

For the classification of the features into their corresponding classes, two models were analysed. These were mostly used due to their broad validation in this kind of application, in fact, both were kindly regarded as the *state-of-the-art* before Deep Learning came into the scene.

Support Vector Machine (SVMs), are a type of supervised machine learning model that separates data points into different classes by finding the optimal hyperplane in an N-dimensional space. The SVM aims to maximise the margin between the classes. When the data is not linearly separable, a kernel function can be used to map the data points to a higher dimensional space where the points can be separated by a hyperplane. Common types of kernel functions include linear, polynomial, Radial Basis Function (RBF), and sigmoid.

Random Forest Classifier, is a kind of learning algorithm based on the ensemble of several "weaker" models. In the case of Random Forest, the weak models are *Decision Tree*, each one being trained with random splits of the train data. The final classification for a sample is made by aggregating the classification on all the trees, usually by majority voting. The more trees a forest has, the more the capacity to fit the data it has and the more robust it should be (under good treatment of the data).

Both of these models have different parameters that need to be tuned. This step is crucial and is discussed in the Results section.

2.4 Deep Learning

Deep Learning (DL) as a learning technique, changed vastly the traditional machine learning paradigm. The crucial difference is in the *Feature Engineering* process, with DL the discovery of meaningful features

is done in a *data-driven* way. This means, the algorithm itself learns which features are more relevant for the task.

For images, the most used model are *Convolutional Neural Networks (CNN)*, a type of neural network based on local-connectivity and parameter sharing. This two concepts, based on the the actual mechanism of vision neurons, make these family of model particularly well-suited for image recognition tasks.

These networks are designed to process data through a series of layers that learn to identify patterns in images. The layers in a CNN typically include convolutional layers, pooling layers, and fully connected layers.

Convolutional layers, are used to identify features, such as edges, corners, and textures. These layers use a mathematical operation called *convolution* to scan the image and extract the meaningful values. It is relevant to mention that, depending on how deep is the layer, the features extracted are usually an aggregation, reaching the point of a layers recognising meaningful structured.

The ***pooling layers*** are used to reduce the dimensionality of the data by down-sampling the feature maps. This helps to reduce computational cost and prevent overfitting.

Finally, ***fully connected layers*** are used to classify the image based on the features that have been identified in the previous layers. These layers connect every neuron in one layer to every neuron in the next layer. The final layer of the network typically carries a softmax operation to output a probability distribution over the different classes.

In this analysis, three different CNN architectures are studied, with additional techniques as transfer learning and fine-tuning. Methods that are able to port the general features learnt on different datasets to achieve better results. The following models were used:

Hand-written CNN model

This model is a smaller, hand-written architecture composed by only `Conv2D` layers, `BatchNormalization` layers and `MaxPooling` layers. This model is trained from scratch and does not utilise pre-trained weights. It is designed with the purpose of being simple yet efficient.

The classification head on the model, takes the features extracted by the convolutional layers and, using a FF Neural Network, classifies into one of the three classes. In particular, the head is composed by:

- `Dense layers`
- `Dropout layers:`[6] The Dropout layer randomly sets input units to 0 during training time to help prevent overfitting.
- `Batch normalisation:`[3] is a technique used to improve the stability and speed of training by normalising the layers inputs.

ResNet

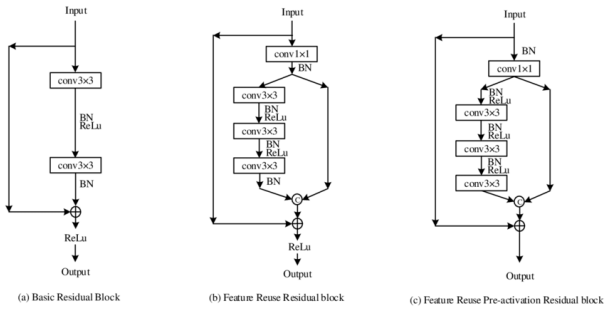
This model is based from the ResNet-50[1] architecture, which is a 50-layer convolutional neural network that uses residual blocks to form the network. The ability of this model to generalise comes from the gradient passing mechanism allowed by the residual blocks. In practice, this mean that the model is able to maintain more stable gradients during the training phase, even on early layers. This model was used to perform *transfer learning* for our task, so the weights trained on *ImageNet* dataset were used.

DenseNet

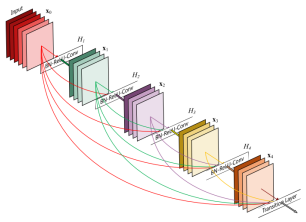
This model is built using the DenseNet-121[2] architecture, which is known for its ability to make deep learning networks more efficient by using shorter connections between layers. Similarly to the ResNet model, only the backbone was used, after which a classification head was added. The weights from *ImageNet* were also used.

2.5 Explainable AI techniques

In this section, the two methods for *Explainable AI* used in the anylisis are presented.



(a) Residual block from ResNet



(b) DenseLayer from DenseNet

2.5.1 Shapley Additive Explanations (SHAP)

Shapley Additive explanations (SHAP)[4] is a model agnostic approach to explain the output of a machine learning model that originates from game theory. SHAP computes the importance of a feature by seeing how well the model performs with and without each feature, for every other combination of features. The shapley value is the average marginal contribution of a feature value across all the possible combinations of features.

2.5.2 XAI with GradCAMs

For the deep learning approach, GradCAMs[5] (Gradient-weighted Class Activation Mapping) of the output class were computed. This technique is a method for visualising the most relevant regions in an image for the prediction of the CNN model. It is a type of explainable artificial intelligence (XAI) method that can be used to understand the decision-making process of a CNN, and can help to identify the specific features in an image that the network is using to make its predictions.

The method works by computing the gradient of the class score with respect to the feature maps in the final convolutional layer of the CNN. These gradients are then averaged across the different feature maps and are used to weight the feature maps, effectively highlighting the regions in the image that have the strongest influence on the class prediction.

The importance of this technique lies in its ability to provide a visual explanation of the decision-making process of a CNN, which can be particularly useful when working with medical images, where the interpretability of the model is crucial. Additionally, it can be used to identify the regions in the image that are most relevant for a given class prediction, which can be useful for further data analysis and model improvement.

3 Results

To assess the proposed models two splits, at different levels, of the dataset were performed. Firstly, the dataset was divided into a *training* and *testing* sets, setting the development and parameter tunning of the models exclusively on the training split. The percentage of the dataset kept for testing purposes was 20%.

Secondly, and for a more robust model selection process, *cross-validation* was used to create different validation scenarios from the training data. This method is vastly used and allows to have less biased statistics for the performance of the models as opposed to only have one *validation* set extracted from the training data.

Furthermore, for both kind of splits, the testing and the ones from the cross-validation, a stratified sampling was used based on the unique patients and the labels. This is crucial to avoid any data leakage from the training to the testing settings and to maintain a similar distribution of classes among the different splits.

For the class imbalance problem, two main ideas were analysed: (1) Up-sampling the minority class, in this case the tuberculosis class and pneumonia classes and (2) use weights on the loss functions of the learning algorithms. The second approach was selected, mostly because a data-intensity problem, up-sampling would require more computational power to train the models, in addition of an augmentation technique to avoid just repeating the samples.

3.1 Features-based models

For each of the models, a grid-search was performed to obtain the best set of parameters and hyper-parameters. This second set of settings consisted on preprocessing steps and the kind of feature reduction technique to be used.

Particularly, for the preprocessing, two image sizes were validated: 224 (medium) and 128 (small) pixels per side. Additionally, each of the preprocessing steps were tested, except for the equalisation and the sharpening. This was done with the idea to understand which techniques were more relevant for the models, specially, among techniques that changed drastically the presented image.

The following are the grids used on the fine-tuning of each model. The `DATA_SUBSET` are all the different combinations of image size and preprocessing steps applied.

```python
# gridSearch for SVM
param_grid = {
    'C': [1, 5, 10],
    'gamma': ['scale', 'auto'],
    'kernel': ['poly', 'rbf'],
    'subset': DATA_SUBSETS
}
```

(a) GridSearch on SVM

```python
# gridSearch for RandomForest
param_grid = {
    'n_estimators': [50, 100, 150]
    'subset': DATA_SUBSETS
}
```

(b) GridSearch on RandoForest

Support Vector Machine

Based on the grid search, the results on Table 1 were obtained. Note that *F1 Tub* value is the F1 score obtained only for the tuberculosis class, while *F1 Macro* considers the three classes as equally relevant.

These results were obtained only from the cross-validation assessment. It is clear that the general performance of these models on the *Tuberculosis* class is low, being the maximum *F1 Score* achieved of about 0.51.

C	Gamma	Kernel	Subset	Features	ACC	F1 Tub	F1 Macro
100	0.001	rbf	small_with_noise_inverted	raw	0.798635	0.517533	0.734064
5	scale	rbf	small_with_noise_inverted	raw	0.789016	0.509271	0.720798
100	0.001	rbf	medium_with_noise_inverted	raw	0.795158	0.509091	0.729100
10	scale	rbf	small_with_noise_inverted	raw	0.797704	0.497382	0.723493
10	scale	rbf	medium_with_noise_inverted	raw	0.791809	0.486346	0.718469

Table 1: Top SVM results

Apart from the performance, some interesting insights are observed. A great analysis can be done from the fact of the model learning better using the *RBF* kernel over the un-reduced features (without any feature reduction technique). Additionally, fro the best subsets, the models performed best when using small images and without noise reductions techniques.

When analysing the results with the reduction techniques applied, the performances were dropped significantly. Specially in the case of *PCA*, being at the bottom of performance. The best PCA and VT results are presented in Table 2:

C	Gamma	Kernel	Subset	Features	ACC	F1 Tub	F1 Macro
10	scale	rbf	medium_with_noise_inverted	var	0.694074	0.391459	0.617206
1	scale	rbf	small_without_noise_not_inverted	pca	0.499535	0.243493	0.440062

Table 2: SVM best results for PCA and VT

Random Forest

The Random Forest classifier results are presented in Table 3. They follow a similar pattern to the ones obtained from the *SVM* model, with very low results for the *F1* score on tuberculosis. Nevertheless, is worth noting that while the results here are far worse in terms of $F1$, the accuracy is on pair and even greater than the previous models.

Furthermore, it is also interesting to check that the performance is better for bigger images and, again, without the noise removal technique. The results for PCA and VT techniques, following the SVM trend, yield lower performance values.

N_Estimators	Subset	Features	ACC	F1 Tub	F1 Macro
50	medium_with_noise_inverted	raw	0.806165	0.246940	0.640821
150	medium_with_noise_inverted	raw	0.808647	0.244107	0.641193
100	medium_with_noise_inverted	raw	0.808440	0.238673	0.641472
50	small_without_noise_not_inverted	raw	0.796235	0.167748	0.611370
100	medium_without_noise_not_inverted	raw	0.805751	0.153460	0.616160
150	medium_without_noise_not_inverted	raw	0.806372	0.153399	0.621993

Table 3: Random Forest results

Testing values

The results of training the best model from each family with all the training set are presented in Table 4. These results are from trying to predict the testing data separated before.

Params	Subset	Features	ACC	F1 Tub	F1 Macro
(100, 0.001, 'rbf')	small_with_noise_inverted	raw	0.8116	0.5345	0.74996
(50)	medium_with_noise_inverted	raw	0.8093	0.2431	0.64351

Table 4: Best SVM and RF results on test

In summary, the best SVM model out-performs the best Random Forest model, both in overall scores and per-class statistics. Nevertheless, both models don't really accomplish reliable results on the Tuberculosis class.

Another important result, is RF models performing the best when trained on bigger images, while the best SVM models were driven by smaller images. It also, relevant to mention that both models seems to get confused by the noise removal technique.

3.2 Deep Learning models

As in the previous experiments, a grid-search with cross-validation for the hyper-parameters of the model was performed. The DenseNet and ResNet models were fine-tuned using pre-trained weights from `imagenet`, while the SimpleCNN was trained from scratch. The metrics to evaluate the performance of the models were F1-Score and Accuracy, given higher priority to the Tuberculosis part of the score.

For the DenseNet and ResNet models, different combinations for the number of layers in the classification head and the number of layers unfrozen in the fine-tuning process were tested . For the SimpleCNN, the grid was trying to optimise the number of convolutional layers on the backbone.

As the analysis presented analysis was on the *structure* of the architecture, the learning rate and weight decay were set at the default values, $1e^{-3}$ and 0, respectively. To avoid overfitting a callback for *Early Stop* was used.

The results on the SimpleCNN grid optimisation are on Table 5, while the ones from ResNet and DenseNet are on Table 6. When comparing between the proposed architectures, it can be noticed that the DenseNet model is the one yielding the best results, both for Tuberculosis and overall.

Layers	F1 N	F1 P	F1 T	Acc N	Acc P	Acc T
3	0.880	0.919	0.641	0.828	0.951	0.777
7	0.878	0.922	0.612	0.834	0.931	0.755
5	0.865	0.916	0.586	0.814	0.908	0.770

Table 5: SimpleCNN search results

Model	Head L	Unfrozen L	F1 N	F1 P	F1 T	Acc N	Acc P	Acc T
DenseNet121	2	10	0.95	0.97	0.85	0.94	0.98	0.89
DenseNet121	3	10	0.95	0.97	0.84	0.94	0.98	0.87
ResNet50	2	10	0.96	0.97	0.84	0.96	0.98	0.83
DenseNet121	3	5	0.95	0.97	0.84	0.94	0.98	0.86
ResNet50	4	10	0.95	0.97	0.83	0.95	0.98	0.84
ResNet50	3	10	0.95	0.97	0.83	0.95	0.98	0.82

Table 6: ResNet and DenseNet search results

The best result for DenseNet model, is with 2 layers in the classification head and 10 layers unfrozen for fine-tuning. This model has the best F1-Score for the Tuberculosis class, with a value of 0.85.

For the ResNet family, the best model is the one with 2 layers in the classification head and 10 layers unfrozen for fine-tuning. It also has a high F1-Score for the Tuberculosis class and the same macro F1 as the best DenseNet model with a value of 0.92.

Finally, the best results for the SimpleCNN model were obtained with 3 CNN layers and reaching a a F1-Score for the tuberculosis class of 0.641.

The results are in line with the expected from the fine-tuned models, as they were initialised using a pre-trained weights and therefore, have a better feature extraction capability. Also, the fact that DenseNet and ResNet have more layers than the simple CNN, increases the representational power of these models, and that is why they performed better.

Furthermore, the best results were obtained with a relatively low number of layers in the classification head and a relatively high number of layers unfrozen for fine-tuning. This is because having a large number of layers for classification may lead to overfitting, while unfreezing more layers allows the model to adapt better to the specific task of X-ray image classification.

The metrics for the best model of each family trained in the complete training split and predicting testings are presented in Table 7. The values itself don't vary much from the cross-validation set-up, but some changes are visible, probably from the inclusion of new data for the algorithm.

Params	F1 N	F1 P	F1 T	Acc N	Acc P	Acc T
(2, 10)	0.95	0.96	0.85	0.94	0.97	0.86
(2, 10)	0.96	0.95	0.84	0.96	0.97	0.83

Table 7: Best ResNet and DenseNet results on test

Performance on Pre-processed Images

A final test was performed training the best model with the fully preprocessed images (medium_without_noise_inverted subset). Overall, the performance didn't increase, but it was relatively the same. For the best DenseNet model, the F1-Score for tuberculosis, had only a difference of a 0.01 less than when training with the raw data.

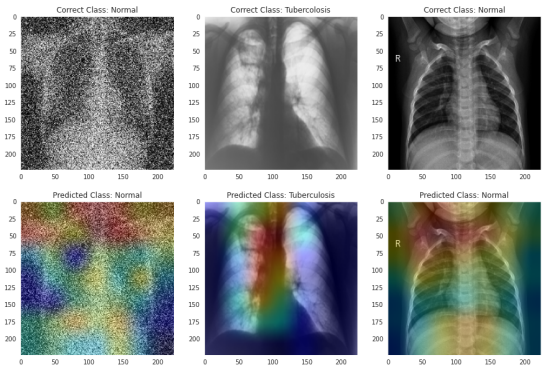
3.3 XAI analysis

Shapley Additive Explanations

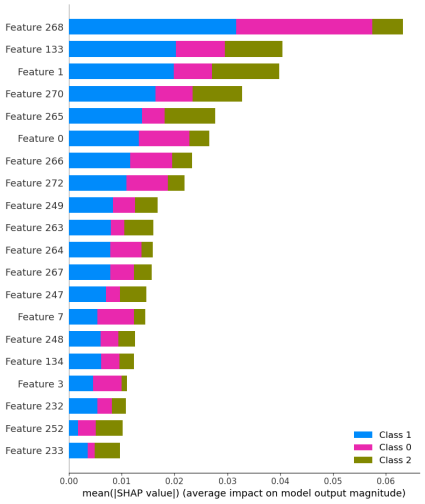
The SHAP analysis was applied to the best SVM model from the grid search. The bar plot in Figure 11b presents a summary of the mean SHAP value for each feature among the top 20 most important features. The top 3 features associated with the index are *Haralick*, *Mean Pixel-value* and the first *Hu moment*, after that most of the features come from *LBP*.

XAI with GradCAMs

The Grad-CAM visualisation for the tuberculosis class indicates that the model is effectively identifying regions of the image that contain anomalous formations, which is consistent with the high performance metrics observed for this class. For the normal class, the model appears to be focusing on more general regions outside of the lungs, even in the presence of noise. This may indicate that the model is robust to the presence of noise, however it may also suggest overfitting to certain features present in the training set.



(a) GradCAMs on the three classes



(b) SHAP on best SVM model

4 Discussion and Conclusion

One of the main challenges of the provided dataset is the **class imbalance**. This problem makes the models more prone to overfitting to the majority class, as they see more examples of it during training. Even when trying to avoid this problem with a technique like class weights on the loss, the overall decrease in performance is notorious in most of the cases.

This is common problem, specially in the biomedical sector, as the frequency of people presenting diseases is lower than healthier individuals. On other settings, this could also be a problem to the other side, when "sampling biases" tend to over-represent specif diseases.

Two future alternatives for this problem can be proposed. Firstly, to perform up-sampling on the minority classes, at the cost of higher computational demand and with the risk of loosing generalisation. A second approach would be to divide the problem into two sub-tasks: (1) Classify Normal vs Disease, having a more similar distribution of the classes and (2) Classifying between Tuberculosis and Pneumonia. Both approaches need to be carefully analysed and tested to get to a satisfactory conclusion.

Feature-based methods

In terms of preprocessing, it is important to note that these techniques can be used to enhance the features that are relevant to the classification task. For example, histogram equalisation can be used to enhance the contrast of an image, making it easier for the model to identify important features such as edges and textures. This step is especially important when we are extracting features based on the pixel intensity or the texture in the images. In the future, an interesting experiment would be to add more fined-grained preprocesses steps, to understand deeper how each hyper-parameter, can vary the performance of the models.

In terms of the results, it is possible that the **Support Vector Machines** models performed better than the Random Forest because SVMs are a linear model, and therefore they tend to be more robust to high-dimensional data. Additionally, SVMs have the ability to handle non-linearly separable data through the use of kernel functions. Overall, the results achieved by this family of models were poor compared to Deep Learning methods.

The **Random Forest**, on the other hand, as an ensemble method, can be highly sensitive to the quality of the data and the number of samples, which can lead to overfitting. In fact, it is this ability to fit really well to complex distribution what makes the a vastly used model. This behaviour can also be seen in the performance of the model with bigger images (more features) and with less preprocessing applied to them, acting as a regularisation.

These results are consistent with the literature in X-ray image classification, where SVM have been frequently mentioned as great approach outside of deep learning. In the analysed case, they don't yield the best results, but this could be a issue with the perturbations being to strong and poorly corrected. Extra analysis on this end is needed to understand a better way to improve the performance of these methods.

Lastly, regarding the poor performance of PCA as a dimensionality reduction technique, the main reason for this behaviour can be attributed to the low correlation between features. This added to the irremovable noise in the images, probably disturbs the correlation between common components, making it difficult to identify PCA components that describe the variance between them. Furthermore, the group-based reduction developed could have intensified this problem as, using PCA in this way, some cross-group relations may be lost during the reduction process.

Deep Learning Models

The CNN models used in this study, DenseNet and ResNet, are well-used architectures for image classification tasks. As expected, the results obtained with these models were superior to those obtained with the feature-based methods and the SimpleCNN model. The best results were achieved with the DenseNet-CNN model, which got to an average F1-score of 0.92. These results demonstrate the power of CNNs in capturing the complex patterns present in chest X-ray images.

Furthermore, analysing the performance against the preprocessed images, it is clear that the presence of perturbations don't really affect the model. This can be explained by the fact that models such as ResNet and DenseNet, are able to learn complex features from the raw data by themselves, and the need for any preprocessing technique may be reduced compared to traditional methods. In other words, they

also learn a "internal" preprocessing when they are extracting the features, therefore, this early steps may be skipped.

Additionally, more experiments could be conducted on the regularisation induced by the perturbations to the models. In theory, this kind of operations could be acting as a "data augmentation" step, making the model more robust and helping to generalise better.

However, despite the good performance of these models, there are some limitations that need to be addressed. One of the main concerns when using DL models is their lack of interpretability. While the GradCAM visualization provides some insight into the regions of the image the model is focusing on, a more comprehensive interpretability method could be explored.

Finally, to improve the performance of the models, some potential future improvements include incorporating more data augmentation techniques, using more complex fine-tuning methods such as Bayesian optimisation, and experimenting with different architectures like Inception and EfficientNet. Additionally, it would be interesting to evaluate the performance of these models on larger datasets and different types of chest X-ray images, such as high-resolution images or images from other modalities such as CT scans.

References

[1] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[2] Gao Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[3] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[4] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI].

[5] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: https://doi.org/10.1007%2Fs11263-019-01228-7.

[6] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.