Data Mining, Machine Learning, and Deep Learning

CXR Age and Gender Classification

Effectiveness of Machine Learning Classifiers to Determine Gender and Age Based on Chest X-Rays

Grischa Tobias Blaich (158371)
Mikołaj Antoni Barański (158286)
Jenő Tóth (158386)
Julius Wirbel (158289)

Characters count: 34,120
Page count: 15
May 19, 2023

Machine Learning Written Exam

MSc. in Business Administration and Data Science

Abstract

In radiology chest X-rays (CXRs) are an important diagnostic tool in determining the health of a patient. In this paper we are testing the performance of Convolutional Neural Network (CNN) models in classifying the age and gender of people from their CXR. To achieve this a baseline Support Vector Machine (SVM) with Principal Components Analysis (PCA) feature extraction, a full CNN model adapted from the VGG-16 architecture, and a SVM with CNN feature extraction are tested in different specifications. The models are trained on the SPR X-Ray Age and Gender Dataset (de Aguiar Kuriki et al., 2023) (n = 10700) labelled with age and gender, utilising the best performing image preprocessing methods including Gaussian blur and histogram equalisation. Our research has indicated that a SVM classifier based on features extracted by a CNN model achieves the best classification performance of both age-cohort and gender. The model achieved a 66.8% and 98.7% accuracy on the respective classification tasks. Dividing the classification task into gender and age-cohort separately and resampling the dataset had no significant effect on the accuracy of the SVM. Based on our results, CNN would need a more balanced and perhaps bigger dataset to perform well. However, this result is not robust to a more diverse pool of CXR images varied by additional factors such as ethnicity or different pathology types. Models such as the ones proposed by our research can be instrumental in providing labels in unlabelled CXR databases, which is especially important to ensure datasets have equal gender and age distributions.

Keywords: Age and Gender Classification, Convolutional Neural Network, Image Classification, PCA-SVM, CNN, SVM, CNN-SVM, VGG-16

1 Introduction

Chest X-rays (CXRs) are an important diagnostic tool in radiology. With the advent of machine learning algorithms, it has become possible to use these images to identify the gender and age of patients. Being able to quickly determine these variables on new data is important as oftentimes CXR datasets are not labelled, which may lead to undiscovered bias in data. Using datasets with hidden imbalances in gender may lead to errors down the line (Larrazabal et al., 2020), therefore a tool is necessary to filter such datasets out and understand its true gender composition. The purpose of this research paper is to compare performances of different machine learning models for classifying age and gender using CXR images. A number of papers looked at this problem, using different models and datasets. This paper will look at a new dataset published by The Radiology and Diagnostic Imaging Society of São Paulo (SPR) (de Aguiar Kuriki et al., 2023). We will be comparing different models to the baseline Support Vector Machines (SVM) based on a Principal Component Analysis (PCA) feature extraction model. The test models include; two Convolutional Neural Networks (CNN) adapted from VGG-16 architecture and four specifications of SVM using a feature extractor based on the aforementioned CNNs. The paper will also examine how our model compares against other models in the literature, by fitting our model on the widely used ChestX-ray14 dataset. Through this analysis we will be able to build upon current research on optimal CXR age and gender classification algorithms to advise researchers on the optimal method they can apply.

2 Literature Review

A number of previous papers have been published which developed machine learning models for classification of CXR images. Earlier research, such as Xue et al. (2018), only looked at gender detection. The paper compared SVM and Random Forest classifiers with four CNN architectures. Their data consisted of only 2,066 images, and the highest accuracy they achieved was 86.6%. This number was improved by later papers. Yang et al. (2021) looked at both gender and age, using over 66,000 images. The paper used the InceptionResNetV2 CNN architecture, and achieved an almost perfect, 99.9% accuracy for classifying gender. The paper determined the age prediction accuracy by measuring the root mean squared error, which was 2.8 years.

Apart from age and gender, other factors have also been examined, such as race, insurance status, and pathologies. Adleberg et al. (2022) focused on predicting age, self-reported gender, self-reported ethnicity and insurance status based on the MIMIC-CXR database, containing over 55,000 images. They used a pre-trained EfficientNet B4 architecture by Tan and Le (2019), and achieved as good results for gender classification as Yang et al. (2021). Rather than looking at the difference between a patient's real and predicted age, the paper categorised ages by 10-year intervals, therefore their results (91.1% accuracy on predicting age cohort) cannot be compared to the previous paper. For race, their model was 92.5% accurate, while for insurance statutes it was lower, 70.5%.

Perhaps the most used dataset for CXR classification is the ChestX-ray14 dataset (Wang et al., 2017). It contains 112,120 images. The main purpose of the data is not to be used for predicting age and gender (although these variables are available as well), but to be used for pathology classification. The dataset contains 15 labels for sicknesses, 14 pathologies and 1 for "No Finding". Despite not being specifically created for gender and age prediction, a number of papers look at models which predict the variables based on the dataset. These papers, and their results can help us with cross-validation, seeing how the accuracy of our model compares to the ones in the literature. They introduce a number of different models, such as a pre-trained DenseNet-121 CNN (Gozes and Greenspan, 2019), a pre-trained ResNet-50 CNN (Baltruschat et al., 2019) and a newly created CNN model (Ali and Ali, 2021).

3 Conceptual Framework

There are many different models to perform an image classification task of this complexity. One of the most researched and performant architectures is the CNN. These models use convolutional layers to extract relevant features from input images. Those are then classified by a classifier stage, often of fully-connected layers which end in a layer which gives a prediction for each possible class. During our literature research, we found several approaches using these networks.

Almezhghwi et al. (2021) compared different CNN and CNN-SVM approaches to classify diseases from CXR images. The authors tested two common CNN model architectures, VGG-16 and Alex-Net, but with a SVM as the final classifier. Their results show that this approach can be more accurate than using a 'traditional' fully-connected classifier.

Solomou and Kazakov (2021) utilise a CNN with a custom classifier stage to predict age and gender from CXR images. Instead of Alex-Net or VGG-16, they utilise a more modern architecture called 'EfficientNetB0' (Tan and Le, 2019) as their base classifier. Furthermore, they create two classifiers, one for age and one for gender prediction. This approach also yields good results, although the age classification is only really accurate in a range of approx 10 years (Solomou and Kazakov, 2021, p.4).

Xue et al. (2018) also compare different CNN architectures with a custom SVM classification stage. Their results also show that using a SVM as the final classifier yields very accurate results. Furthermore, their results indicate that the VGG-16 architecture despite its age and simplicity outperforms more modern architectures like LeNet or ResNet on this task. Hence, we chose to utilise a CNN based on the VGG-16 architecture as our feature extractor.

The last architecture we considered was 'ChestNet' Wang and Xia (2018). They also utilise a CNN, more specifically ResNet-152, model with a custom classification stage. What sets them apart from the other papers is the use of a 'Attention Branch', a further branch which tries to learn the 'regions of pathological abnormalities' (Wang and Xia, 2018). The result of this extra branch is combined with the output of the actual ResNet-152 model and leads to a measurable improvement in AUC accuracy. Still, this network architecture was not feasible for us, as it is extremely computationally complex and furthermore would overfit our very limited age and gender classes due to its complexity.

Based on previous literature we wanted to verify the performance of the previously applied techniques on the SPR dataset. To see their relative performance we will compare the accuracy of a full CNN model, to that of a SVM based on a CNN feature extraction, and to that of a baseline model a SVM based on PCA feature selection.

4 Data Overview

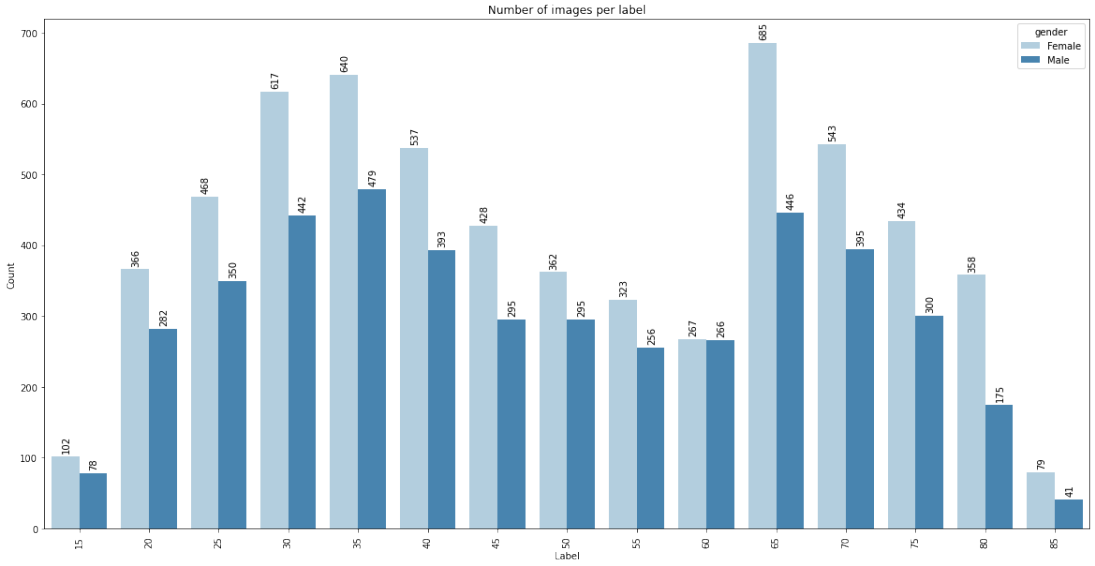


Figure 1: Age and gender distribution of the dataset after grouping

Our dataset consists of 10,700 high quality CXR images from the SPR labelled by age and gender (de Aguiar Kuriki et al., 2023). The images in the dataset consist of grayscaled CXRs from patients. These are scans of the X-Ray sheets and come in a resolution of 1024x1024 pixels. These images, while all having the same resolution, differ on where the scan is within the boundaries of the image, containing black bars to fit the resolution.

The dataset consists of CXRs covering 72 different ages ranging from 18-89 as well as two genders. This distribution mirrors a standard age pyramid in a high-income country. Notably for nearly all ages, there are more samples for women than men. In our analysis 5-year cohorts

were used for classification tasks. This groups the ages into 15 classes, which combined with the genders creates 30 classes in the final dataset as per Figure 1.

5 Methodology

5.1 Data preparation

5.1.1 Data cleaning

In the initial stage of our analysis we investigated the dataset for errors and corrupted files. Due to the size of the dataset it was not feasible to assess the accuracy of the labels in the dataset as differences between gender and age of a CXR are not easily recognizable.

Our investigation indicated only three images in the dataset were corrupt and causing errors. These images and one more image have been removed to keep the class number even to use batch sizes larger than one. Further image cleaning was not necessary, as the rest of the images all were confirmed to conform to the image sizes and deliver complete image data.

5.1.2 Data preprocessing

To optimise the accuracy of classification models used in our paper we have implemented a series of data preprocessing and augmentation steps. Given that the preprocessing steps ultimately impact the performance of our final classification model, different combinations of various presumed effective image preprocessing methods were tested on a basic CNN. The subsequent sections will provide further explanation, including the considered preprocessing methods.

Cropping: The first step in the preprocessing process is to crop the images, to remove their black borders, as can be seen in Appendix B.1.1. The cropping maintains the images aspect ratio. This is required to later scale all the images to the same size to be processed by the CNN. Cropping allows the model to focus on the region of interest and discard unnecessary information. By following this preprocessing step, it is ensured that all the CXRs are consistently cropped, preserving the aspect ratio, and ready for further processing and analysis using classification models.

Gaussian blur: Gaussian blur is a frequently used technique for noise reduction in medical images (Cadena et al., 2017, p. 5). It performs far better than alternatives such as Wiener filters (Ramadan, 2019). Medical images, such as CXRs, and MRIs, can often suffer from various types of noise, including random variations in pixel values. Which can degrade image quality and make accurate analysis and interpretation of the image more difficult. The Gaussian blur utilises a Gaussian function, which is associated with normal distribution, to determine the transformation applied to each pixel in the image (Cadena et al., 2017, p.6).

The kernel size implemented determines the amount of blurring applied to the image. When applied to medical images, it helps to reduce high-frequency noise, while preserving essential image features, such as edges and structures. The kernel acts as a smoothing function, averaging out pixel values, which helps to suppress noise and create a smoother appearance. Larger kernel sizes result in more aggressive blurring and stronger noise reduction, potentially causing a loss of fine details. (Cadena et al., 2017)

Various papers, including Giełczyk et al. (2022) and Cadena et al. (2017), suggested a 5x5 Gaussian blur kernel size for CXRs. To determine the most suitable kernel size for our data,

we decided to test one kernel size below (3x3) and one above (7x7) (Appendix B.1.2). Ensuring that we use the best kernel for reducing noise in the dataset, without losing fine details.

Edge enhancement: Edge enhancement is another widely used image preprocessing method, aiming to improve the visibility of edges or boundaries between different structures (Chaira, 2012). However, before applying edge enhancement, it is often beneficial to perform some basic preprocessing steps on the image (Chaira, 2012). This may include cropping, to remove any irrelevant regions, and Gaussian blur to reduce noise. Such preprocessing steps help prepare the image for effective edge enhancement.

Excessive enhancement may amplify noise or introduce artefacts that may be misleading. Therefore, a careful evaluation of the enhanced image is necessary, to guarantee that it improves the clarity and interpretability of the CXR without misrepresenting important features (Appendix B.1.3). To ensure that edge enhancement has a positive effect on the classification accuracy of our model, preprocessing process tests were run with and without it.

It should be noted that the edge enhancement function used is relatively simple, without adjustable parameters. This deliberate simplicity ensures that the edge enhancement improves the classification ability of the model without introducing complexities or subjective biases.

Normalisation: Normalising CXR images is an important step in image preprocessing to ensure consistency and enhance the performance of subsequent machine learning algorithms. The goal of normalisation is to adjust pixel values of an image to a standard range or distribution, allowing for better comparison and interpretation of data. In our case, we explored histogram equalisation (HE), which was suggested by Giełczyk et al. (2022) and Contrast Limited Adaptive Histogram Equalization (CLAHE).

HE aims to redistribute the pixel values of an image to span a desired range (Al-Ameen et al., 2015). It achieves this by stretching or compressing the pixel value distribution using the histogram of the image. By applying HE, the visibility of image details can be enhanced and the overall image quality is improved. In some situations it can perform poorly, losing detail, over enhancing and amplifying noise (Al-Ameen et al., 2015). However, it may also reveal hidden details and improve the appearance of images (Temiatse et al., 2018).

On the other hand, CLAHE is a variant of HE that adapts to the local characteristics of an image. It divides the image into smaller regions and applies HE individually to each region, taking into account the local contrast (Shi et al., 2020). An excessive increase in the brightness of the image may lead to an unbalanced contrast in the final image (Al-Ameen et al., 2015; Shi et al., 2020). This adaptive approach aims to overcome the limitation of traditional HE, which can lead to over-amplification of noise (Shi et al., 2020). Both approaches were tested as part of our preprocessing, as well as not normalising the images (Appendix B.1.4).

5.1.3 Preprocessing Testing:

To optimise the performance of our classification model, the most effective preprocessing process had to be determined. In order to achieve this, tests were conducted using the possible combinations of the previously mentioned preprocessing methods, to determine the process that yields the best classification results in a simple CNN model. Allowing us to efficiently evaluate the performance of various preprocessing processes.

The tests performed are summarised in Table 1 (Appendix B.2.1). It is important to note that the input images for these preprocessing variations were always the cropped images obtained

through earlier preprocessing steps. These results provide valuable insights into the performance of the preprocessing combination.

Process	Accuracy	Precision	Recall	F1 score
1.4	0.204206	0.209754	0.204206	0.2069428221
1.6	0.200467	0.169793	0.200467	0.1838594141
3.4	0.189252	0.194351	0.189252	0.1917676111
4.1	0.173364	0.184669	0.173364	0.1788380206
4.4	0.167757	0.21268	0.167757	0.1875661871

Table 1: Selected results of preprocessing testing

Based on these results we decided to choose to use Gaussian Blur with a 3x3 kernel and HE as the optimal preprocessing process, illustrated in Figure 2. This process demonstrated the highest accuracy (0.204), recall (0.204) and f1 score (0.207). While other processes showed similar performance, this particular combination stood out, performing well across the board.

Figure 2: Preprocessing Steps

a) Cropped b) Gaussian Blur (3x3) c) Histogram Equalisation

5.1.4 Data Augmentation

To increase the diversity and quantity of the data our models can be trained on, we decided to augment our data. One commonly used data augmentation technique for images is flipping. By horizontally flipping the existing images, the number of images in the dataset was effectively doubled. This technique is particularly useful when the orientation of an object or the direction of a certain feature is not crucial for classification. In the case of CXRs the anatomical structures remain the same regardless of left-right orientation. By applying horizontal flipping to the images, we introduce additional examples with slightly altered appearances while preserving the underlying information (Appendix B.3.1). This augmentation strategy significantly increases the size of our training set, which contributes to improving our model's performance.

However, it is important to note that not all types of transformations are suitable for every dataset or task. In our case, rotations were not implemented as part of data augmentation to increase the dataset size, as it was not effective in improving performance.

5.2 Classification Algorithms

5.2.1 Model Choice

Based on our literature review we decided to compare different commonly applied model architectures to be able to compare and contrast their performance. Our baseline model is a SVM classifier based on CXR data fed through a PCA feature extractor. The PCA is based on the first 200 components which were able to preserve 90% of the variance of the image. This classifier jointly classifies gender and age-cohort by predicting which one of the 30 classes is most fitting to each image.

Our second model is a CNN model based on VGG-16 architecture (Simonyan and Zisserman, 2015), which also jointly predicts age cohort and gender. The third model is a variation of the VGG-16 architecture model which has a separate gender and age cohort classifier.

Our fourth model uses the above described CNN model as a feature extractor, while conducting the final classification using a SVM, jointly predicting age cohort and gender. The fifth model splits the fourth one into three sub-models: the first predicts the gender, whereas the other two predict the age cohort based on the specific gender.

Finally due to the class imbalance present in the dataset we also conducted analysis models with the same architecture as four and five, but which we trained on a resampled dataset. These resampled models are our sixth and seventh models.

Model	Feature Extractor	Classifier	Joint prediction gender & age
1 - Baseline	PCA	SVM	Yes
2 - Full CNN	CNN	CNN	Yes
3 - Full CNN Dual	CNN	CNN	No
4 - SVM	CNN	SVM	Yes
5 - SVM Dual	CNN	SVM	No
6 - SVM Resampled	CNN	SVM	Yes
7 - SVM Resampled Dual	CNN	SVM	No

Table 2: Full list of analyzed models with their specifications

5.2.2 Feature extraction model

Based on the literature review, we chose to create a model based on the VGG-16 architecture (Simonyan and Zisserman, 2015). This architecture consists of blocks of Convolutional Layers followed by a MaxPooling layer. However, as the goal is to learn 30 classes and not 1000 classes of the ImageNet dataset, the architecture has been simplified slightly to prevent overfitting. More specifically, the deeper layers of our feature extractor contain only two Convolutional Layers instead of three, building on the VGG-16-B structure proposed in the original paper (Simonyan and Zisserman, 2015, p. 5). The final feature extractor architecture is shown in Appendix C.1. Furthermore, the convolutional layers within the feature extractor are normalised by the L2-Norm with a factor of 0.001, initialised using 'he-normal' initialization (He et al., 2015) for the weights and activated using the ReLU function. Regularization is used to prevent the exploding gradient problem, in which the gradient values increase significantly during training, but also to prevent overfitting.

5.2.3 CNN Classifier Models

First Classifier: Single classifier for age and gender The model with a combined classifier has a classification stage consisting of seven layers, three of which are Fully Connected / Dense layers with 128, 64 and 30 neurons respectively. Like the layers in the feature extractor, the weights of the layers in the classifier are initialised using the he-normal distribution. Furthermore, the bigger layers are activated using ReLU, while the classifier is activated using the Softmax function. Within the classification stage, there are also two Dropout-Layers, one in the beginning and one before the final classifier, with a rate of 0.4 and 0.2 respectively. As well as a GlobalAveragePooling layer to flatten the output from the final convolution of the Feature Extractor and a BatchNormalization layer to standardise the data before passing it to the classifier stage. The two fully-connected layers before the final classifier are also regularised using the L2-norm.

This structure was chosen based on the VGG-16 classifier (Simonyan and Zisserman, 2015), but with measures to prevent overfitting added to it. Furthermore, adding the Dropout-Layers also helps the model to generalise better. The dropout rates were chosen based on a small scale, manual feature search and showed the best results of the tested models. Lower dropout rates have led to harsh overfitting, while higher rates prevented the network to converge further than 20% accuracy on the validation data.

Second Classifier: Separate classifiers for age and gender The combined classification task has a two-output classifier that branches into one classifier for the age and one for the gender. Building on the network with the single classifier, the branch for the age is the same. As the branch for the gender only needs to perform a binary classification, the additional fully-connected layers from the age branch were omitted and the final classifier is activated by a Sigmoid function instead of the Softmax for multiclass classification.

This is in contrast to final classifiers used by Wang and Xia (2018), which contain more fully connected layers with more neurons and the split at the end of the network, resulting in more computational complexity. Further differentiating our model, we chose to predict the ages as a class and not as a regression like Wang and Xia (2018) did.

5.2.4 CNN Training

Images and labels are split into a Train and Validation dataset using a 80/20 split, resulting in 17120 training and 3280 test images. To account for the class imbalance of the dataset, the individual class weights are also calculated and passed to the training function. This controls the influence each class has on the update of the gradient, so that the classes with less samples are not underrepresented. Unfortunately, this only works for the CNN with the combined classifier, as Tensorflow does not support that function for multi-output networks.

The training itself was performed using the Adam optimizer by Kingma and Ba (2017). In contrast to Stochastic Gradient Descent, this optimizer computes adaptive learning rates for each parameter in the model. This leads to improved convergence, e.g. finding the minimum of the loss function. To further improve the convergence quality, a stepped learning rate was used, halving the initial learning rate of 0.001 every 16 epochs.

5.2.5 SVM Classifiers

To choose the optimal SVM classifiers we applied a grid search with five-fold cross validation to find the optimal kernel, C and gamma values. They were applied if they outperformed the

automatically chosen hyperparameters of the SVM class. In the end, all SVMs used the RBF kernel, C=1, and gamma of either 0.01 or an automatically calculated gamma.

For the resampled data set, 4 resampling techniques were applied. These include two undersampling techniques - random unders-sampling and NearMiss. The first technique artificially decreases the size of the overrepresented classes by removing similar data points in the over-represented class, while the NearMiss method deletes data points from the larger class which are close to the under-represented class. We also included two oversampling techniques - Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). SMOTE creates additional samples of the low class based on prevalent features of existing data points, while ADASYN synthetically creates additional samples of data points which are difficult to classify between classes.

5.2.6 Evaluation Metrics

To evaluate the performance of our models we compared the accuracy, precision, recall, and f-1 scores achieved by the models. Accuracy measures the proportion of correct predictions to the total number of predictions. Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances out of all actual positive instances. The f-1 score combines precision and recall, which may be useful with imbalanced datasets. We measured the accuracy of predicting combined age-gender classes, as well as predicting only gender.

6 Results

6.1 Training Results

Table 3: Performance Metrics of Models - Age-Cohort and Gender

Model	Precision	Accuracy	Recall	F1-score
1 - Baseline	0.212	0.205	0.205	0.180
2 - CNN	0.344	0.326	0.326	0.307
3 - CNN dual	0.271	0.276	0.276	0.264
4 - SVM	0.683	0.668	0.668	0.670
5 - SVM dual	0.683	0.668	0.668	0.670
6 - SVM Resampled	0.675	0.667	0.667	0.668
7 - SVM Resampled dual	0.675	0.667	0.667	0.668

As can be seen in Table 3 and 4 the SVM classifier based on CNN features has the highest performance on our dataset reaching an accuracy of 67% on combined age-gender and over 98% on gender itself. This is significantly better than the baseline model and better than the full CNN models. Notably, there is no significant difference between models 4,5,6, and 7. This shows that in the core dataset resampling and divided prediction do not yield different prediction accuracies for the SVMs.

Table 4: Performance Metrics of Models - Gender

Model	Precision	Accuracy	Recall	F1-score
1 - Baseline	0.900	0.900	0.900	0.900
2 - CNN	0.961	0.959	0.959	0.959
3 - CNN dual	0.913	0.905	0.905	0.903
4 - SVM	0.986	0.986	0.986	0.986
5 - SVM dual	0.987	0.987	0.987	0.987
6 - SVM Resampled	0.986	0.986	0.986	0.986
7 - SVM Resampled dual	0.988	0.988	0.988	0.988

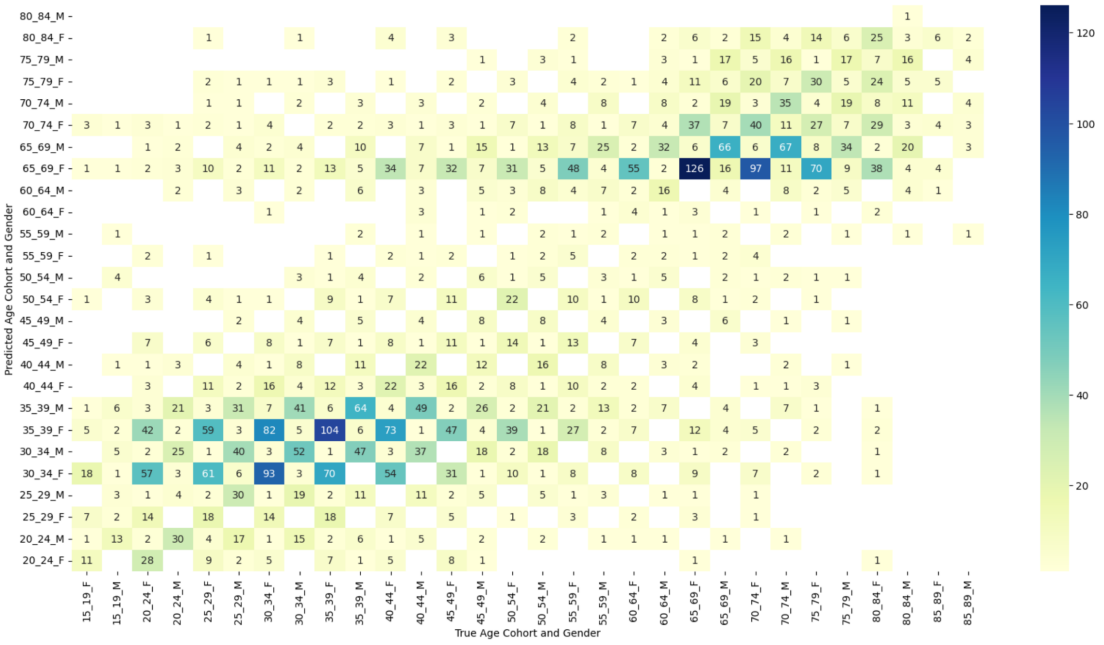


Figure 3: Age-gender predictions of baseline PCA-SVM (Model 1)

6.1.1 Model 2 - CNN

The CNN with the classifier predicting a combination of the age cohort and gender achieved an accuracy of 27.76%. This is comparable to the baseline PCA-SVM, although the CNN took way longer to train with a total training time of nearly 3 hours. The lack of pretrained weights allowing the CNN to start with a better baseline also likely contributed to the slower and worse convergence, although the ImageNet dataset which is used for pretraining in the models we compare to does not contain any X-Ray data.

The heatmap suggests that the gender is almost always predicted correctly, while the age group is what brings the model down. The big spread around the plane shows two peaks which do correspond to the skewness of our input data. This suggests that to improve the accuracy of the CNN, more data augmentation for the underrepresented classes needs to be done and balancing the weight of the classes alone is not enough.

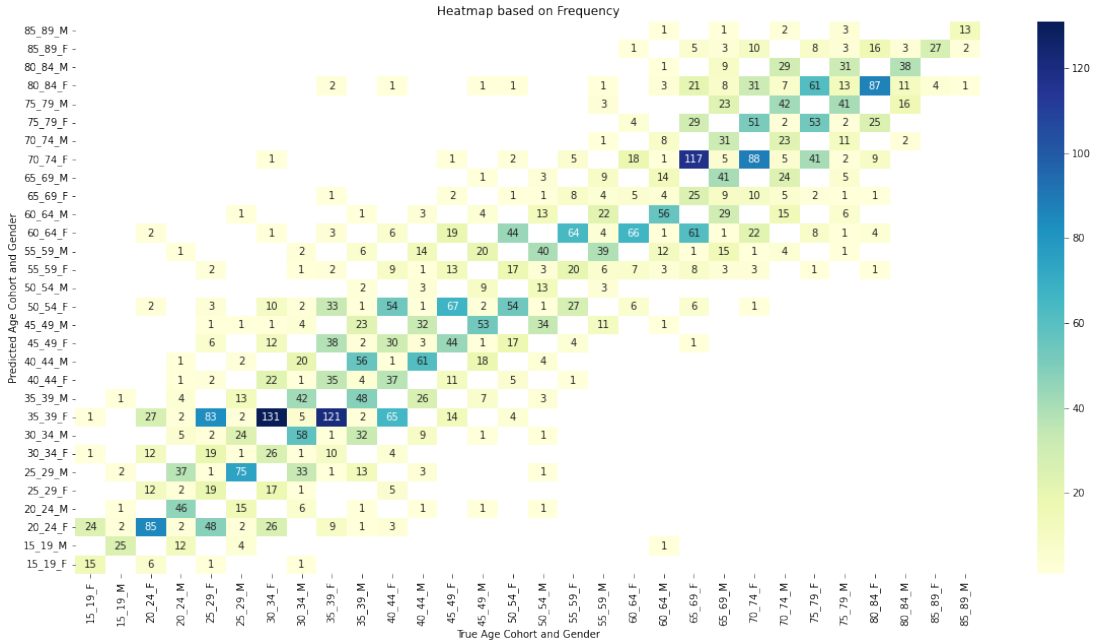


Figure 4: Age-gender predictions of CNN (Model 2)

6.1.2 Model 3 - CNN Dual

While the model with separate predictors for age and gender does well on the gender, it falls apart on the age classification, barely outperforming the baseline PCA-SVM model in the combined age/gender prediction. This is evident when looking at the heatmap showing where the misprediction happen.

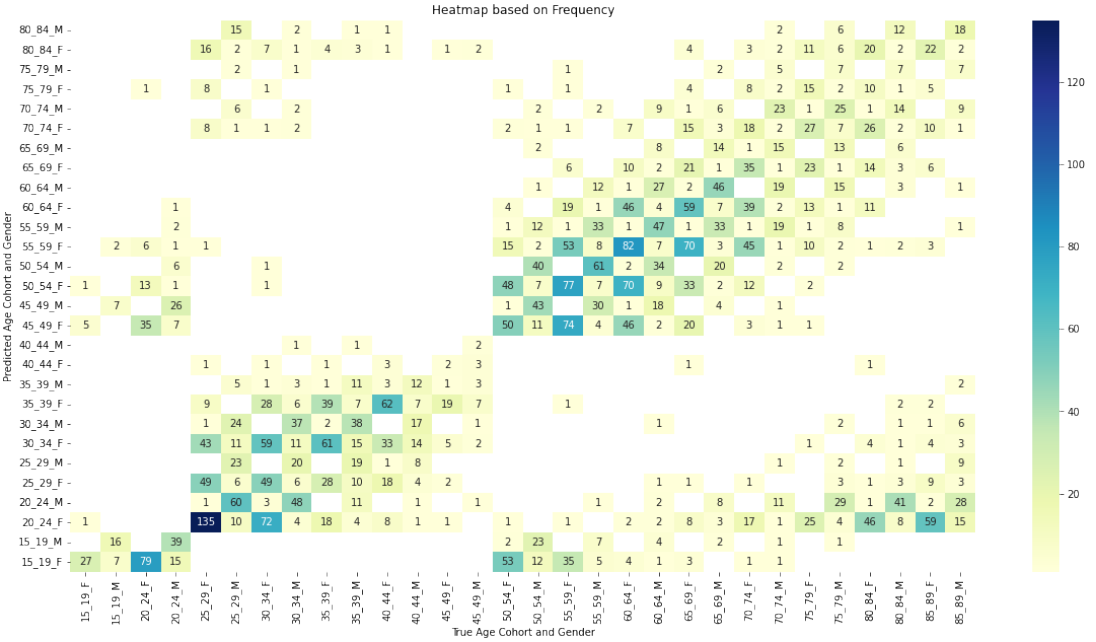


Figure 5: Age-Gender predictions of Model 3

Unlike the SVM, which has a significantly lower spread, this model tends to misclassify the ages, especially in the middle of the age range. Furthermore, the spread around the correct age group is relatively big, as the age only ever reaches a weighted average accuracy of 27.6%. This is due to bad performance on the age cohort and persistent overfitting after about 55 epochs, c.f. Appendix F.

Even with a comparatively simple classifier and measures to combat overfitting, the CNN with the split prediction for the classes fails to perform as hoped. Running the network with even fewer neurons in the fully-connected layers might mitigate the overfitting, and further data augmentation for the underrepresented classes might be needed.

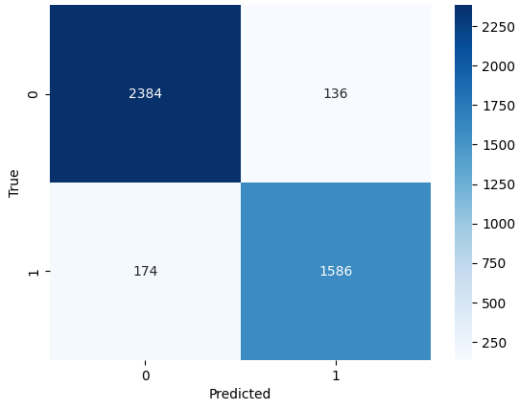


Figure 6: Baseline gender prediction (Model 1)

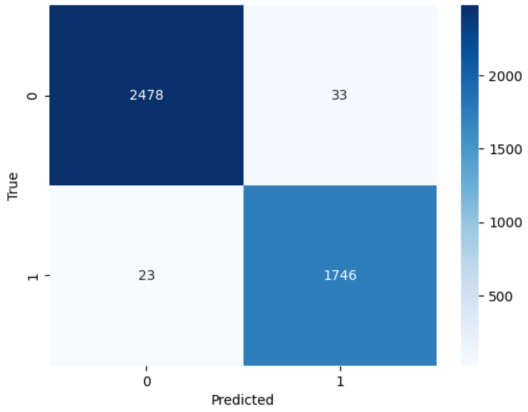


Figure 7: SVM-CNN gender prediction (Model 4)

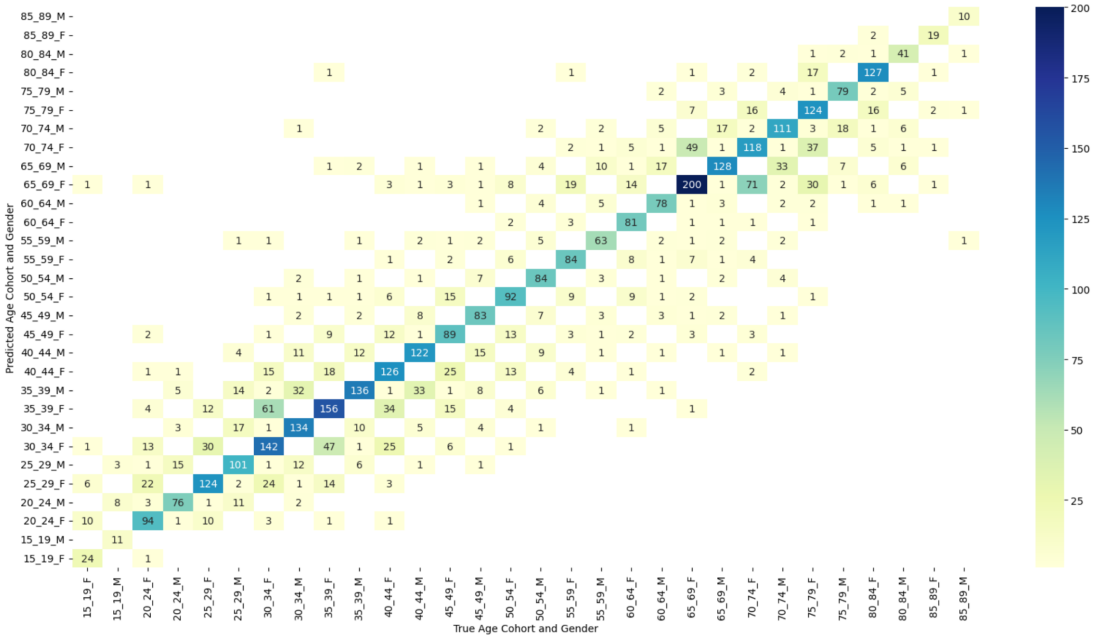


Figure 8: Age-gender predictions of SVM-CNN (Model 4)

6.1.3 Models 4,5,6,7 - SVMs with CNN features

As shown on the confusion matrices and heatmaps, gender predictions of the CNN-SVM models are very precise with over 98% accuracy. Age predictions of these models are also very well aligned with the data, with the majority of mispredictions occurring within one or two age cohorts. This comes in stark contrast to the results of the baseline model, which only achieves a 20% accuracy, and models 2,3 with CNN classifiers. Moreover, the results are only weakly biased towards the 35-39 and 65-69 age cohorts which are predicted disproportionately compared to the other age cohorts. In contrast the baseline model rarely predicts any cohorts between 35-65 and has a very high bias towards the two main cohorts. This follows the uneven distribution of samples in these age-cohorts, hence indicating that the baseline model is having a hard time getting a good understanding of less represented cohorts.

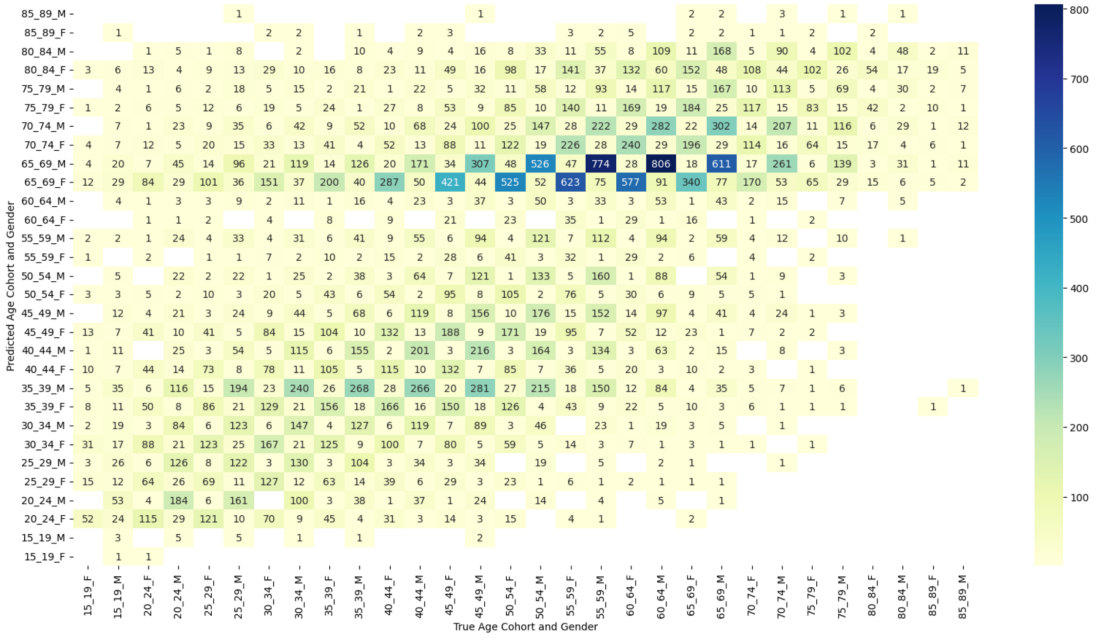


Figure 9: Prediction performance of SVM-CNN with resampling (Model 6)

6.2 Cross Validation

CXR images do not differ greatly between datasets. As such, our model should be useful in later research, with new data. We decided to cross-validate how useful our best models (4-7) are comparing them to three other models using a common dataset. The dataset used for the cross validation is the ChestX-ray14 dataset from the National Institutes of Health (NIH) Clinical Center (Wang et al., 2017). The dataset contains 112,120 images belonging to 30,805 patients. The images' labels contain the patients' age, gender, and 14 types of pathologies, as well as image-related data, such as image parameters.

The CNN-SVM applied to the cross validation dataset performs significantly worse than on the original dataset, achieving only 13% accuracy on age cohort-gender classification and 90% on gender itself. The result here appears to mimic the low power of the baseline classifier described in the previous section. The heat-maps (Appendix E) indicate that the 35-39 and

65-69 age cohorts are being predicted for an exceeding number of instances. This is marginally less prevalent on the CNN-SVM model trained on the re-sampled dataset.

The three articles we use for comparison measure age-related accuracies differently. Gozes and Greenspan (2019) and Baltruschat et al. (2019) look at age as a continuous variable, and calculate the mean absolute error between predicted and actual values. Ali and Ali (2021) use classes, however, their age groups are 1-10; 11-20; 21-30; and 31-120. Because of this, we only look at the gender prediction accuracies. The papers report four metrics: accuracy, sensitivity, specificity and the Area Under the ROC Curve (AUC). The values for our best model were 0.896, 0.905, 0.89 and 0.896 respectively.

Gozes and Greenspan (2019) uses an ImageNet pre-trained DenseNet-121 CNN model. They published the AUC for their gender-predicting model, whose value is 0.997. Baltruschat et al. (2019) looked at AUC, sensitivity and specificity. Their AUC was 0.9435, the sensitivity 0.878, and the specificity 0.859. Finally, Ali and Ali (2021) published 0.99 AUC and 0.95 accuracy scores. Our model, while having higher sensitivity and specificity than Baltruschat et al. (2019), did not perform as well as state-of-the-art models.

7 Discussion

The baseline model's low performance can be attributed to its use of PCA for extraction of key image features. While PCA can drastically reduce the size and complexity of a datapoint while capturing the majority of its variation, the chosen variation may not be the defining element of the image under analysis. The architecture of a CNN is specifically designed to focus on the features of an image, which are most important for the classification task. This allowed the CNN-SVM models (4-7) to perform better than the baseline.

However, both CNN networks performed considerably worse than the CNN-SVM models. Based on the heatmaps shown in Figures 4-5, one can see that the CNN struggles to correctly fit the age groups in the 'middle' of our dataset. Those happen to be the classes, which have a lower number of images per class. While the combined classification model performs better, probably due to the individual class weights, the split CNN model struggles to match this performance. This difference in performance between the CNN and the CNN-SVM might be due to the limited amount of data available to train the CNN, more specifically the imbalance in the dataset.

To gain further insights into the performance of our single classifier CNN model a Grad Cam analysis was carried out. This method enables us to visualise and interpret where the CNN model focuses during its classification process. In Figure 10, we can see the highlighted regions of interest to the CNN model. Upon examining this output, it becomes evident that the CNN model tends to look for features in areas that are not medically relevant for determining the age or gender from a CXR. To improve the model's performance further cropping of the top and bottom of the image is worth considering. This change would allow the model to focus on more medically relevant features, located in places that are more applicable to classifying the age and gender from a CXR.

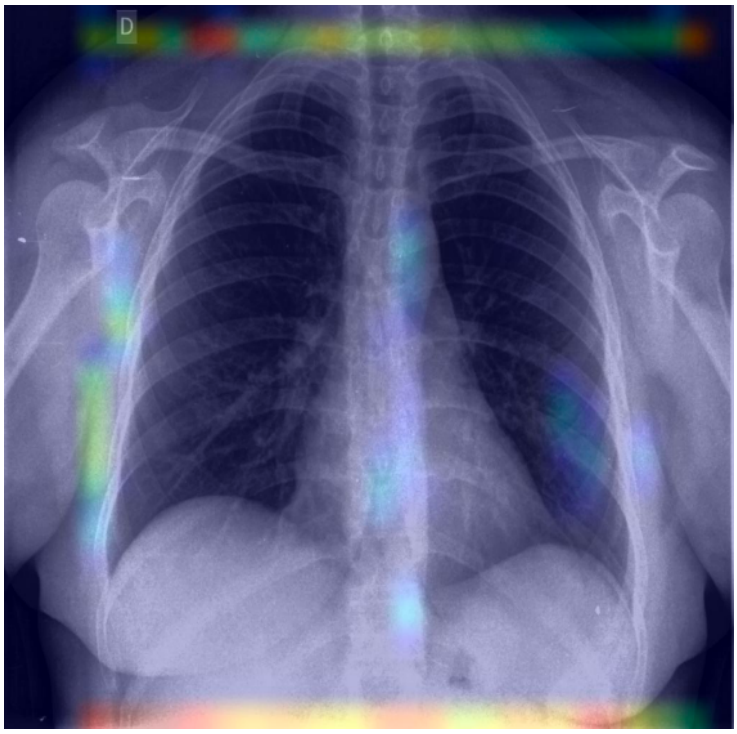


Figure 10: Example of Grad Cam from predictions

8 Conclusion & Future Work

Our work has demonstrated the power of different Machine Learning classifiers on the SPR CXR Age and Gender Dataset (de Aguiar Kuriki et al., 2023), indicating that optimal performance can be achieved by a SVM classifier run based on image features extracted by a CNN model. We showed this model has achieved superior performance compared to the baseline SVM based on a PCA feature selection, and better performance than the CNN classifier. The optimal SVM was able to overcome the skewed nature of the dataset with an unequal distribution of observations per class. The chosen SVM classifier was additionally less computationally intensive than corresponding full CNN models.

From the cross validation it is clear that there is still significant scope for future work and improvements. The current models demonstrate a good performance for data from the same region as the training data, but their performance suffers when classifying ages in other regions and with varying pathologies. This discrepancy can be attributed to differences in people's structural development in different regions as well as an imbalance of training data. To address these limitations, it would be necessary to train the model using data from diverse regions and balanced classes to ensure its robustness and generalizability.

Additionally, future work could involve exploring alternative and additional preprocessing processes, such as contrast, to refine the approach to preprocessing CXR images. Cropping could also be explored further as the Grad Cam showed the CNN looking for features in areas that are unimportant for determining age and gender. Furthermore, future research could also involve classifying the person's health, identifying sickness in CXRs and classifying them accordingly. Such advancements in medical image analysis and classification could have a profound impact on the healthcare industry, enabling more accurate and efficient diagnosis and treatment planning.

References

Adleberg, J., Wardeh, A., Doo, F. X., Marinelli, B., Cook, T. S., Mendelson, D. S., and Kagen, A. (2022). Predicting patient demographics from chest radiographs with deep learning. *Journal of the American College of Radiology*, 19(10):1151–1161.

Al-Ameen, Z., Sulong, G., Rehman, A., Al-Dhelaan, A., Saba, T., and Al-Rodhaan, M. (2015). An innovative technique for contrast enhancement of computed tomography images using normalized gamma-corrected contrast-limited adaptive histogram equalization. *EURASIP Journal on Advances in Signal Processing*, 2015(1).

Ali, M. and Ali, R. (2021). Gender and age detection assist convolutional neural networks in classification of thorax diseases. *PeerJ Computer Science*, 7:e738.

Almezhghwi, K., Serte, S., and Al-Turjman, F. (2021). Convolutional neural networks for the classification of chest x-rays in the IoT era. *Multimedia Tools and Applications*, 80(19):29051–29065.

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*, 9(1).

Cadena, L., Zotin, A., Cadena, F., Korneeva, A., Legalov, A., and Morales, B. (2017). Noise reduction techniques for processing of medical images. *Proceedings of the World Congress on Engineering*, 1:5–9.

Chaira, T. (2012). A rank ordered filter for medical image edge enhancement and detection using intuitionistic fuzzy set. *Applied Soft Computing*, 12(4):1259–1266.

de Aguiar Kuriki, P. E., Farina, E., Abdala, N., Aragão, B., Coelho, M., Straus, M. T., Bianco, G., and Kitamura, F. C. (2023). Spr x-ray age and gender dataset.

Giełczyk, A., Marciniak, A., Tarczewska, M., and Lutowski, Z. (2022). Pre-processing methods in chest x-ray image classification. *PLOS ONE*, 17(4):e0265949.

Gozes, O. and Greenspan, H. (2019). Deep feature learning from a hospital-scale chest x-ray dataset with application to tb detection on a small-scale dataset.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

Ramadan, Z. M. (2019). Effect of kernel size on wiener and gaussian image filtering. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 17(3):1455.

Shi, Z., Feng, Y., Zhao, M., Zhang, E., and He, L. (2020). Normalised gamma transformation-based contrast-limited adaptive histogram equalisation with colour correction for sand–dust image enhancement. *IET Image Processing*, 14(4):747–756.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Solomou, C. and Kazakov, D. L. (2021). Utilizing chest x-rays for age prediction and gender classification. © IEEE, 2021. This is an author-produced version of the published paper. Uploaded in accordance with the publisher?s self-archiving policy. Further copying may not be permitted; contact the publisher for details.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

Temiatse, O. S., Misra, S., Dhawale, C., Ahuja, R., and Matthews, V. (2018). Image enhancement of lemon grasses using image processing techniques (histogram equalization). In *Data Science and Analytics*, pages 298–308. Springer Singapore.

Wang, H. and Xia, Y. (2018). Chestnet: A deep neural network for classification of thoracic diseases on chest radiography.

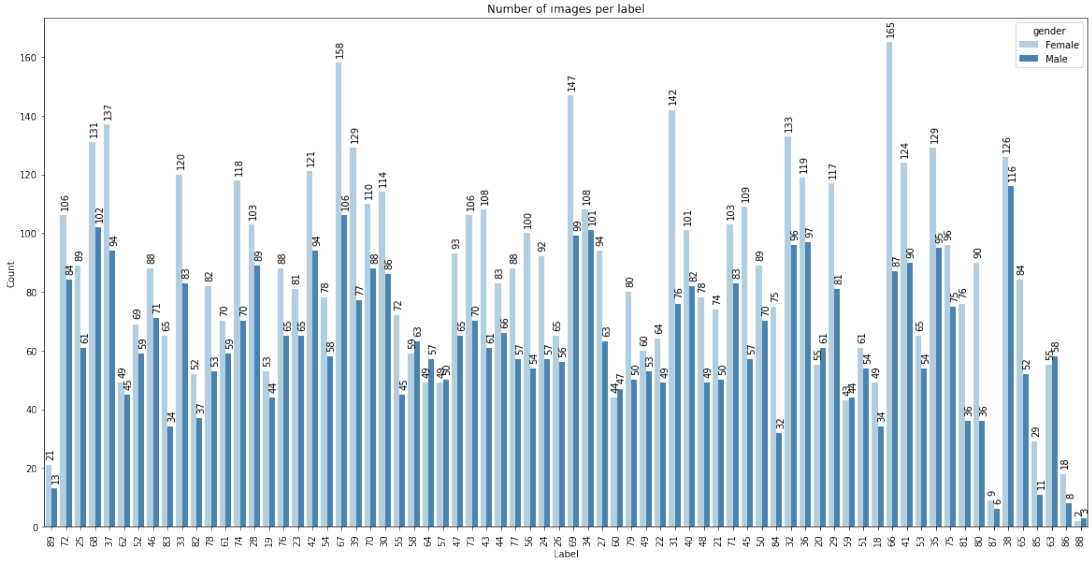
Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Xue, Z., Antani, S., Long, R., and Thoma, G. R. (2018). Using deep learning for detecting gender in adult chest radiographs. In Zhang, J. and Chen, P.-H., editors, *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. SPIE.

Yang, C.-Y., Pan, Y.-J., Chou, Y., Yang, C.-J., Kao, C.-C., Huang, K.-C., Chang, J.-S., Chen, H.-C., and Kuo, K.-H. (2021). Using deep neural networks for predicting age and sex in healthy adult chest radiographs. *Journal of Clinical Medicine*, 10(19):4431.

A Data Exploration

A.1 Age and gender distribution before grouping

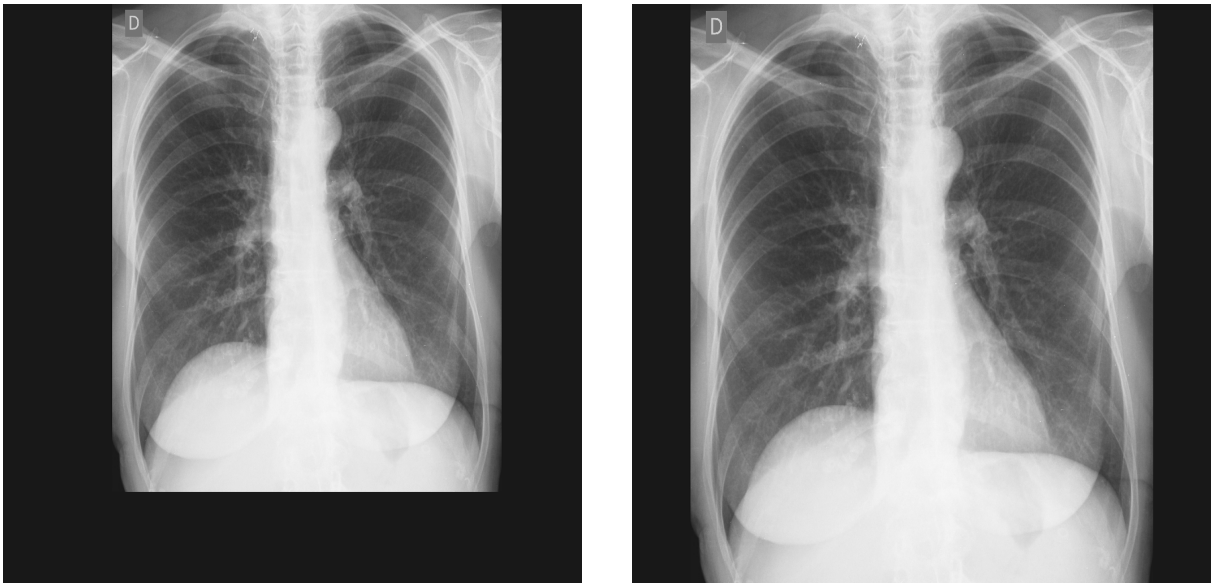


Age and gender distribution of the dataset before grouping

B Dataset Preparation

B.1 Data Preprocessing

B.1.1 Cropping



a) Original

b) Cropped

B.1.2 Gaussial Blur

a) Cropped

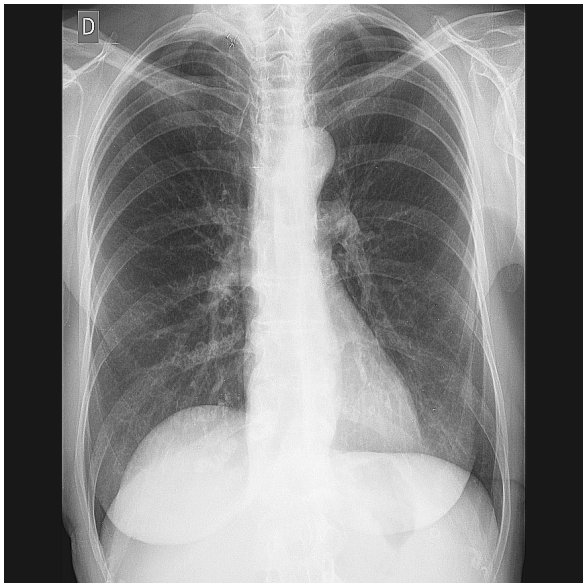
b) 3x3 Kernel

c) 5x5 Kernel

d) 7x7 Kernel

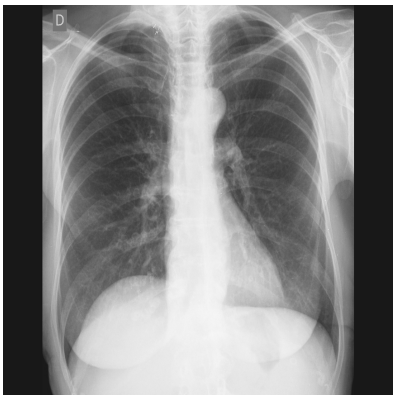
B.1.3 Edge enhancement

a) Cropped



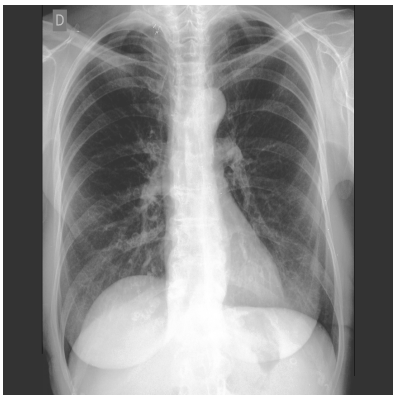
b) Edge Enhanced

B.1.4 Normalisation



a) Cropped

b) Histogram Equalisation



c) CLAHE

B.2 Preprocessing Testing

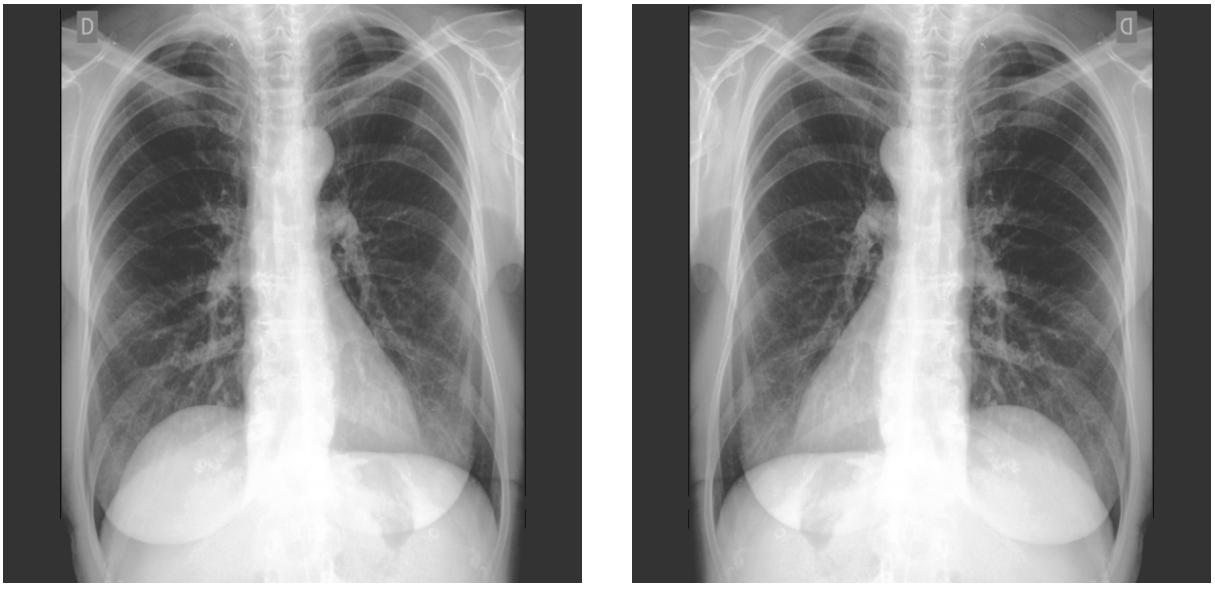
B.2.1 Preprocessing options

Process	Gaussian Blur	Edge enhancement	Normalisation
1.1	3x3	TRUE	Histogram
1.2	3x3	TRUE	CLAHE
1.3	3x3	TRUE	-
1.4	3x3	FALSE	Histogram
1.5	3x3	FALSE	CLAHE
1.6	3x3	FALSE	-
2.1	5x5	TRUE	Histogram
2.2	5x5	TRUE	CLAHE
2.3	5x5	TRUE	-
2.4	5x5	FALSE	Histogram
2.5	5x5	FALSE	CLAHE
2.6	5x5	FALSE	-
3.1	7x7	TRUE	Histogram
3.2	7x7	TRUE	CLAHE
3.3	7x7	TRUE	-
3.4	7x7	FALSE	Histogram
3.5	7x7	FALSE	CLAHE
3.6	7x7	FALSE	-
4.1	-	TRUE	Histogram
4.2	-	TRUE	CLAHE
4.3	-	TRUE	-
4.4	-	FALSE	Histogram
4.5	-	FALSE	CLAHE
4.6	-	FALSE	-

Process	Accuracy	Precision	Recall	F1 score
1.1	0.149065	0.151324	0.149065	0.1501860059
1.2	0.168692	0.135574	0.168692	0.1503306265
1.3	0.164486	0.170889	0.164486	0.1676263768
1.4	0.204206	0.209754	0.204206	0.2069428221
1.5	0.138318	0.206059	0.138318	0.1655259716
1.6	0.200467	0.169793	0.200467	0.1838594141
2.1	0.181308	0.176547	0.181308	0.1788958292
2.2	0.147196	0.126163	0.147196	0.1358703313
2.3	0.148131	0.181327	0.148131	0.163056595
2.4	0.179907	0.191723	0.179907	0.1856271548
2.5	0.178505	0.189206	0.178505	0.1836997916
2.6	0.154206	0.175661	0.154206	0.1642357688
3.1	0.188318	0.163103	0.188318	0.1748058924
3.2	0.157944	0.175681	0.157944	0.1663410108
3.3	0.183645	0.173415	0.183645	0.1783834519
3.4	0.189252	0.194351	0.189252	0.1917676111
3.5	0.133645	0.183255	0.133645	0.1545668317
3.6	0.178972	0.176359	0.178972	0.1776558924
4.1	0.173364	0.184669	0.173364	0.1788380206
4.2	0.129439	0.085002	0.129439	0.1026163269
4.3	0.17757	0.190488	0.17757	0.1838023038
4.4	0.167757	0.21268	0.167757	0.1875661871
4.5	0.179907	0.209638	0.179907	0.1936379297
4.6	0.163084	0.159065	0.163084	0.1610494303

B.3 Data augmentation

B.3.1 Final Image and Flipped Final Image



a) Final Image

b) Flipped Final Image

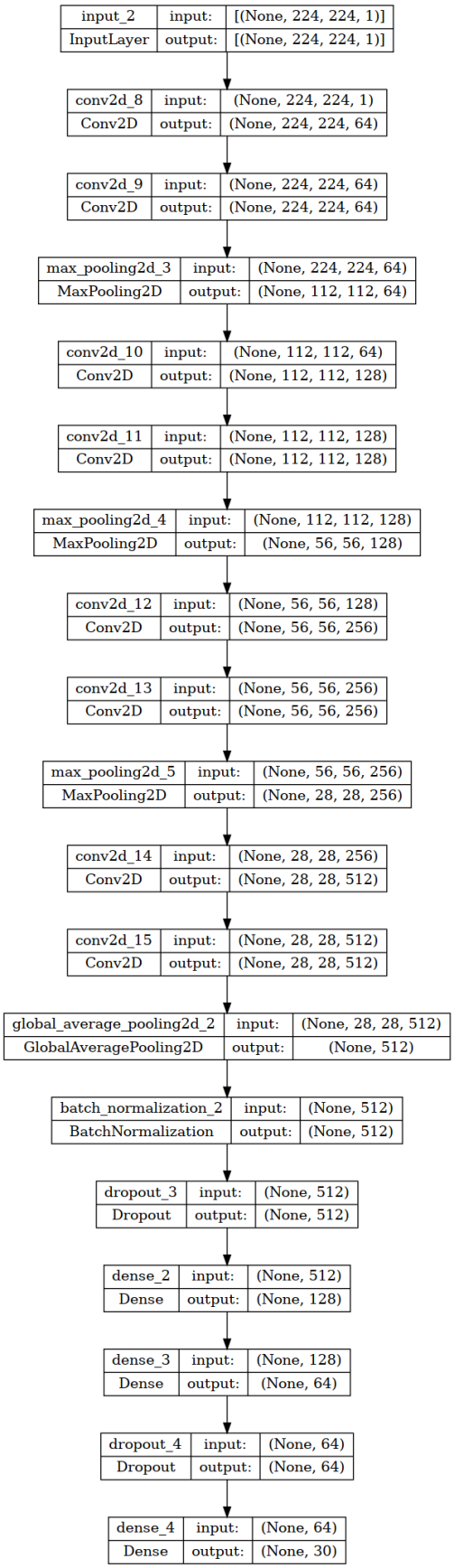
C Classification Algorithms

C.1 Full model specification

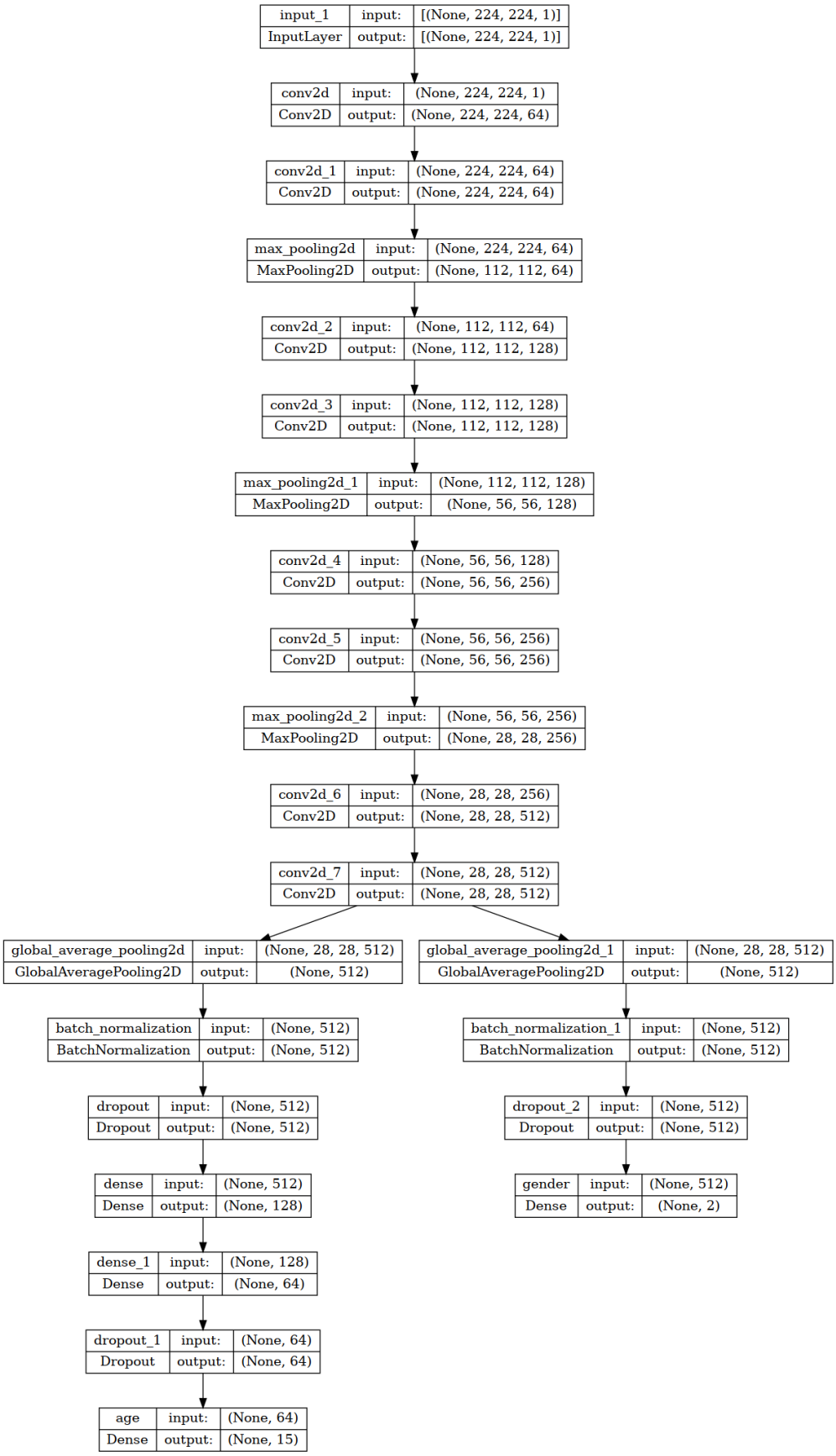
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 1)]	0
conv2d (Conv2D)	(None, 224, 224, 64)	640
conv2d_1 (Conv2D)	(None, 224, 224, 64)	36928
max_pooling2d (MaxPooling2D)	(None, 112, 112, 64)	0
conv2d_2 (Conv2D)	(None, 112, 112, 128)	73856
conv2d_3 (Conv2D)	(None, 112, 112, 128)	147584
max_pooling2d_1 (MaxPooling2D)	(None, 56, 56, 128)	0
conv2d_4 (Conv2D)	(None, 56, 56, 256)	295168
conv2d_5 (Conv2D)	(None, 56, 56, 256)	590080
max_pooling2d_2 (MaxPooling2D)	(None, 28, 28, 256)	0
conv2d_6 (Conv2D)	(None, 28, 28, 512)	1180160
conv2d_7 (Conv2D)	(None, 28, 28, 512)	2359808
max_pooling2d_3 (MaxPooling2D)	(None, 14, 14, 512)	0

Total params: 4,684,224
Trainable params: 4,684,224
Non-trainable params: 0

D CNN Model Overviews



Single Net Layer Overview



Dual Net Layer Overview

D.1 CNN Training Class Weights

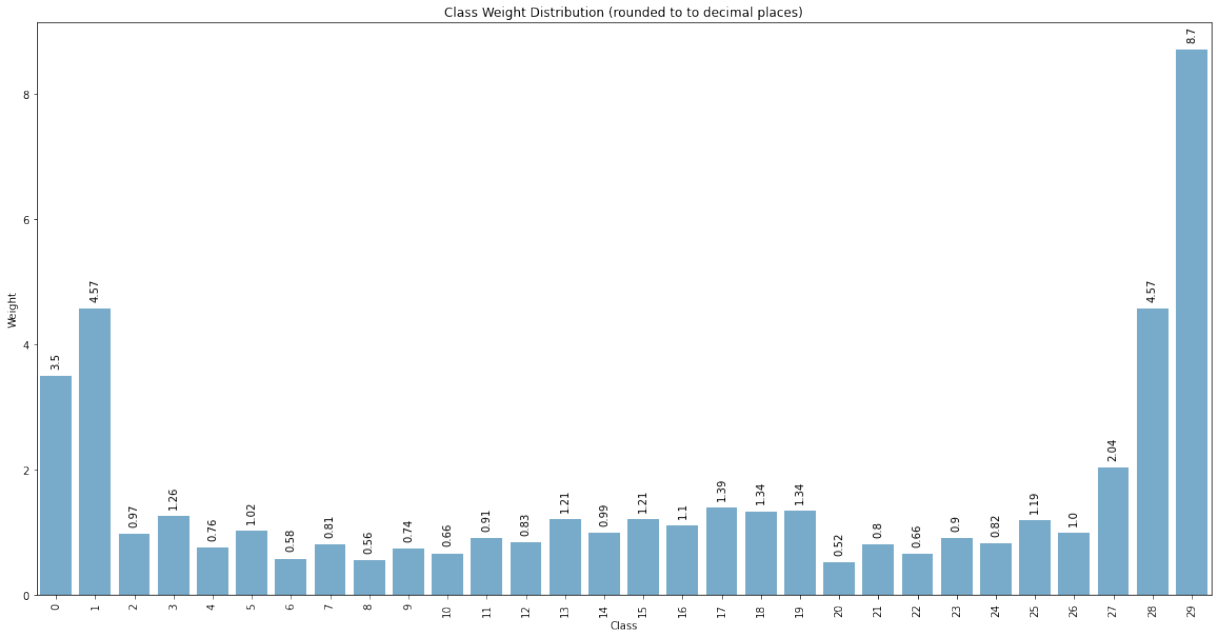


Figure 11: Class weights

D.2 CNN Training Results

D.2.1 Single Model

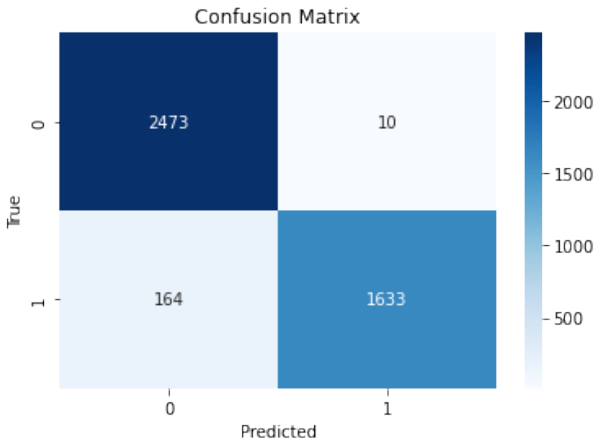


Figure 12: Gender predictions of Model 2

D.2.2 Dual Model

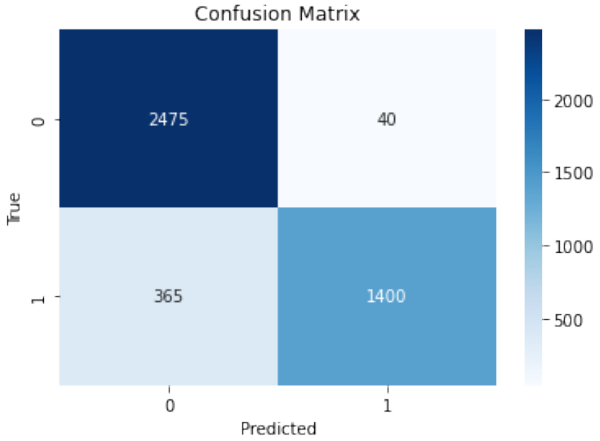
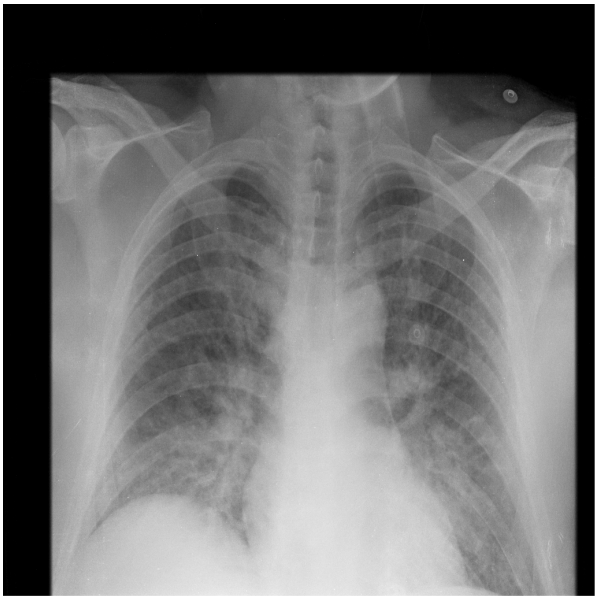


Figure 13: Gender predictions of Model 3

E Results

E.1 Cross Validation

E.1.1 Cross Validation images



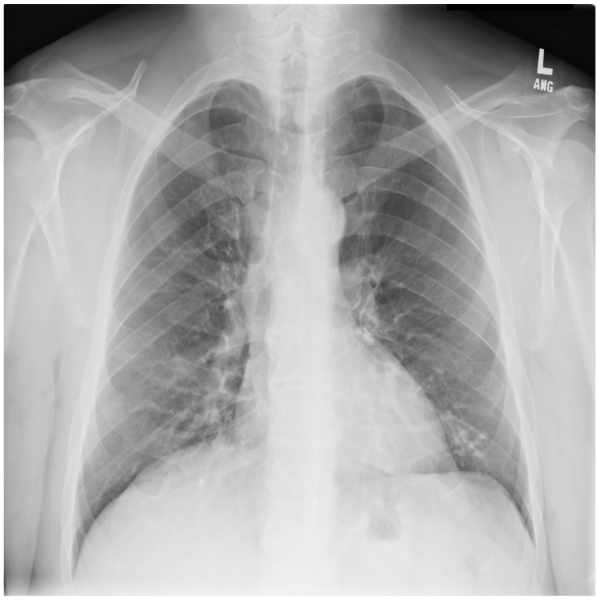
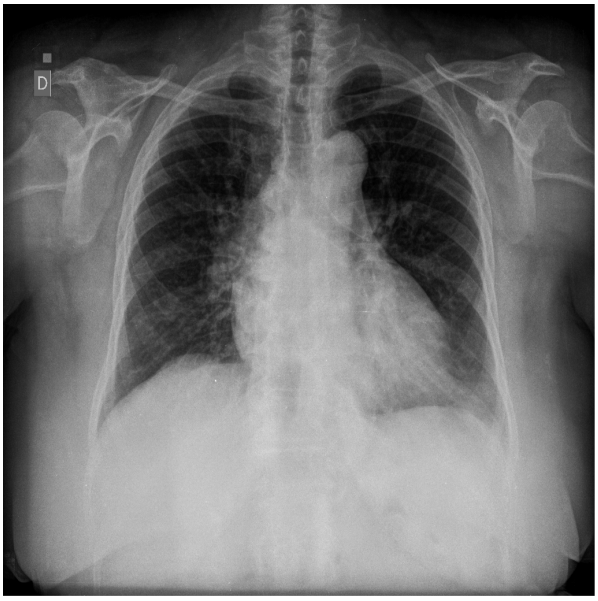


Figure 14: 52-year old males (left: our data, right: ChestX-ray14)



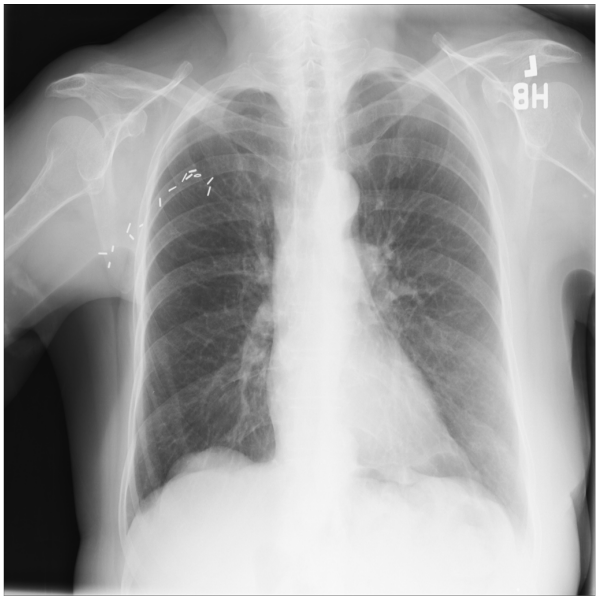


Figure 15: 74-year old females (left: our data, right: ChestX-ray14)

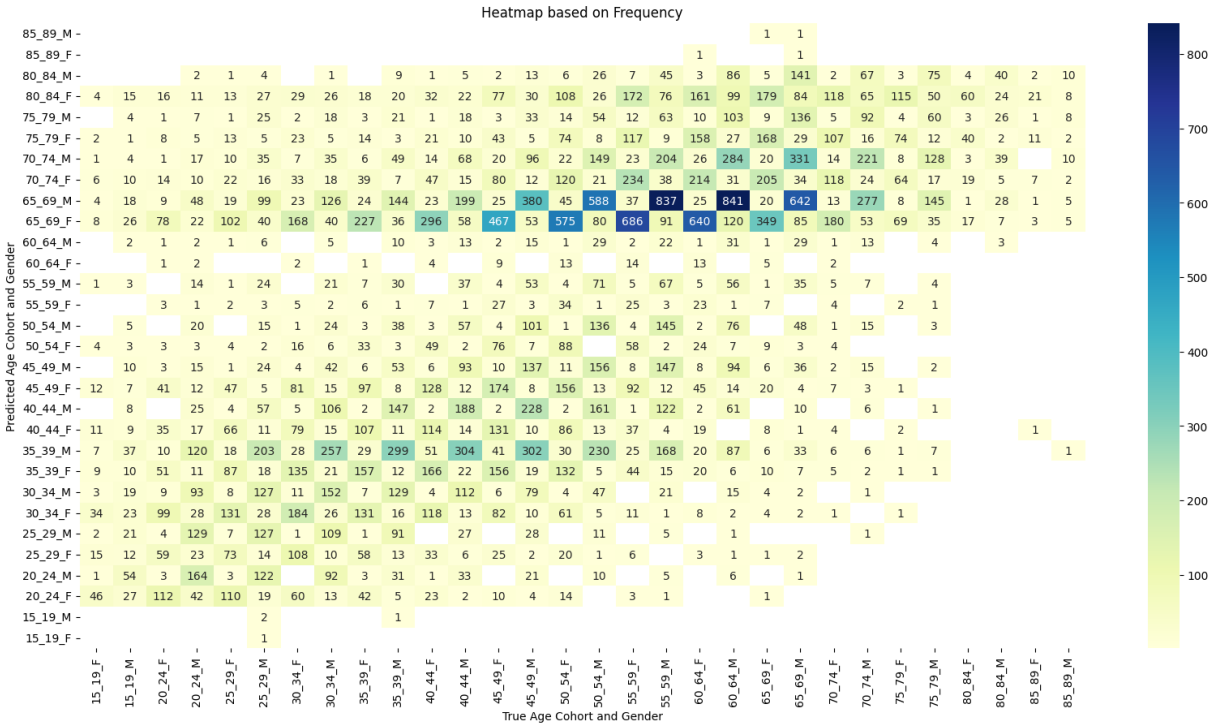


Figure 16: Age-Gender predictions of Model 4 on the cross-validation dataset

E.1.2 Cross Validation Results

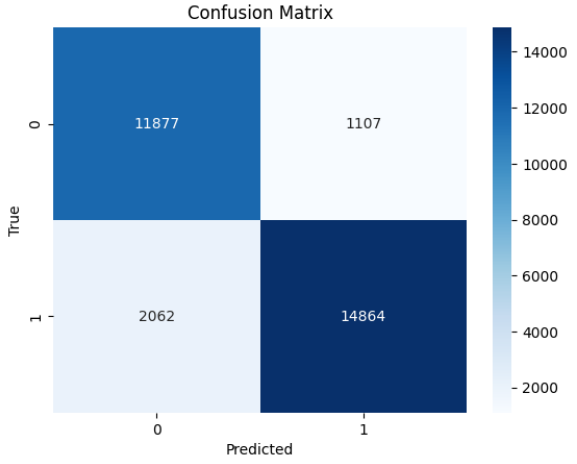


Figure 17: SVM-CNN gender prediction (Model 4) on the cross validation dataset

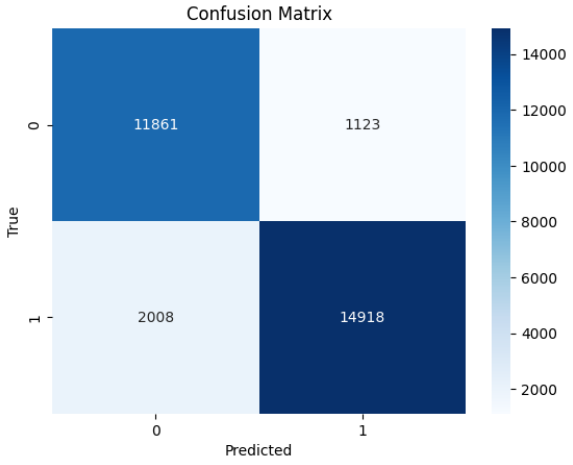
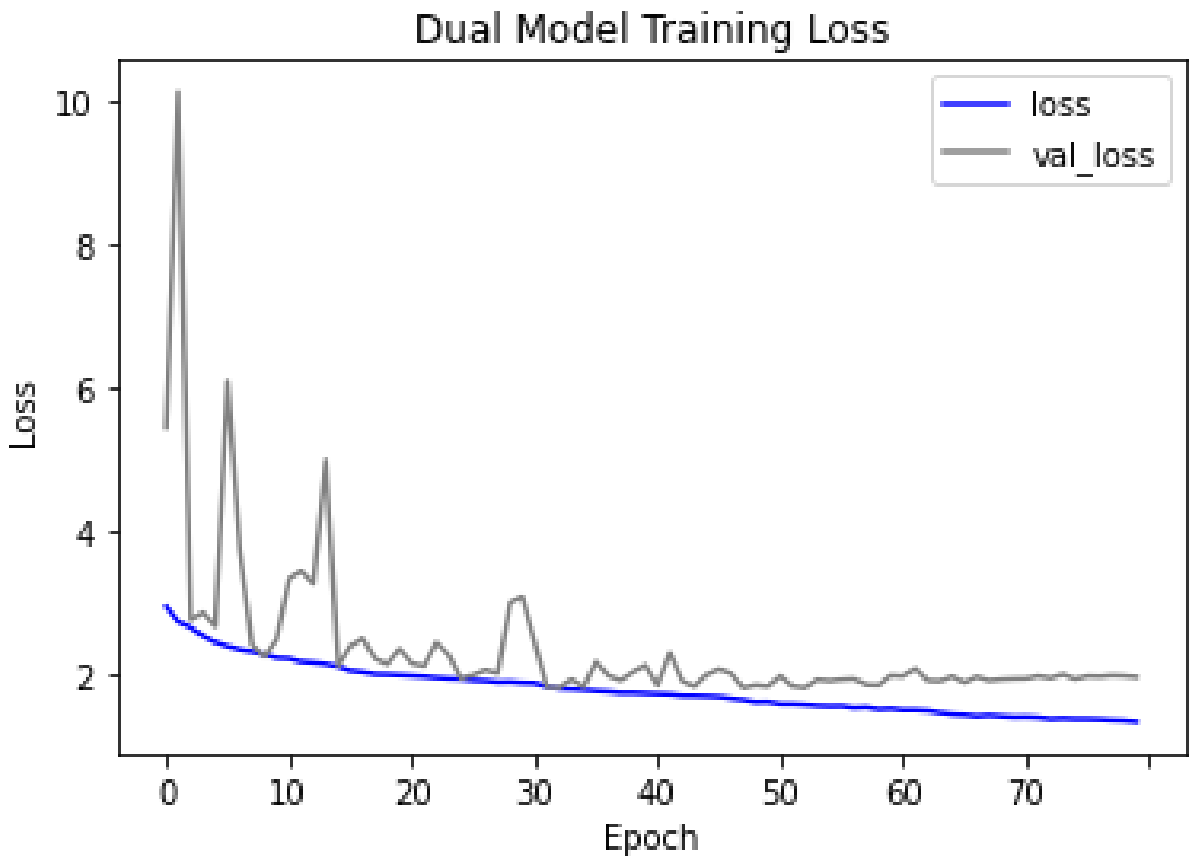


Figure 18: SVM-CNN gender with resampling prediction (Model 6) on the cross validation dataset

F Loss Graph


Dual Model Training Loss
loss
val_loss
Loss
Epoch
10
8
6
4
2
0
10
20
30
40
50
60
70