Feature extraction techniques for localization of region of interest of abnormalities in biomedical images

Dorodi Afroze Krishty, Ludwig Wilhelm Wall, Osazee Ero
SYDE 675 – Group 9, Option B
Systems Design Department
University of Waterloo, Canada
{dkrishty, lwall, oero}@uwaterloo

***Abstract*—Advances in digital medical imaging technologies are generating large archives of data containing valuable information. This presents an opportunity to find automated methods for enhanced medical decision support systems to assist medical experts in decision making. In this paper, we present an assessment of three unsupervised feature learning approaches for detection and classification of diseases using medical images. These techniques are weakly supervised feature localization with CNNs, unsupervised learning with ORB-SVM and Autoencoder networks. The models implementing each of the three techniques are trained on image data from the *CoronaHack Chest X-Ray-Dataset* with 5800 chest X-ray images of healthy patients and patients with pneumonia.**

***Index Terms*—ORB, PCA, k-means, SVM, CNN, U-Net, Autoencoder**

I. Introduction

Many machine learning techniques rely on the availability of labelled training data. Ideally, this training data should be available in abundance to simplify the training procedure and improve the performance of the trained models. In the domain of medical diagnostics, annotating medical image data with useful information is an arduous process that requires medical expertise and specific domain knowledge. As such, labelled training data of high quality is not always available and acquiring the labels can be costly. In some cases, such as classifying rare diseases, sufficient training data might just not exist. Finally, training data with labels for new and unknown classes is always unavailable. Faced with the challenge of an exhaustive archive of medical images without consistent annotations, unsupervised learning presents the means for feature extraction and localization of regions of interest that will assist in the retrieval of key information of biomedical concepts. We compare automated methods regarding reliability and efficiency in detecting abnormalities, and whether they can ultimately support decision-making in clinical treatment.

The chosen task is classifying signs of pneumonia in X-rays chest image data. X-rays are one of the most effective methods to diagnose pneumonia. Pneumonia causes a staggering number of mortalities every year across the world. Classifying pneumonia using chest X-rays requires professionals like radiologists and such expert knowledge has limited availability. Both traditional machine learning and deep neural networks have been applied to various tasks in the medical domain extensively and they have achieved varying levels of accuracy in detecting biomedical insights from X-ray images. However determining suitable or the best approach, given the many options in both traditional and deep neural network based learning methods, remains an open challenge.

We address this challenge by implementing and evaluating three different unsupervised feature learning approaches. One of these approaches relies on a key-point extraction and description algorithm, such as ORB, to train a support vector machine (SVM). The second one is a weakly supervised feature localization with a convolutional neural network (CNN) and finally, a U-Net based autoencoder network. The learned features are validated through qualitative feature analysis and trained by classifiers for biomedical concept detection.

The paper is organised in the following order: the 'Background' section reviews relevant literature describing the advances in the medical image classification to date. The 'Methods' section describes the design of the unique experiments. 'Results & Analysis' discusses the results of our experiments. Finally, conclusions are drawn with future prospects of advancing the employed current techniques.

II. Background

A. Feature Detection & Classification with ORB-SVM

Traditional machine learning algorithms such as SVM have been in use in medical image classification for a very long time. In several research methods for classifying benign and malignant tumours by way of extracted features used to construct a classification with SVM achieved accuracies of 85% [1] and 91.84% [2]. On mammographic images using the SVM classifier, the method achieved 88.75% [3] and 89.3% accuracies [4]. However, SVM has its demerits too – the feature extraction and selection are time-consuming and vary according to different objects [5] whereas deep neural networks such as CNN have achieved significant performance since 2014 [6]. The combination of SIFT and SVM classifier to classify medical images has shown to get 67% accuracy [7], however SIFT is a patented technology. Then Rublee et al. [8] proposed the freely accessible ORB feature detection technique. Jhadav et al. [9] achieved accuracy of 77.6% using

SVM with ORB for image classification of normal versus pneumonia X-ray dataset. ORB (Oriented FAST & Rotated BRIEF) is a fusion of FAST key point detector and BRIEF descriptor [10]. A feature detector is an algorithm that detects feature-points (also called interest points or key-points) in an image [11]. The process of finding unique patterns dictated by changes in gradient and direction is called feature description as it describes each feature by assigning it a distinctive identity enabled for effective recognition for matching [12]. Several types of feature detectors are available such as SIFT, SURF, ORB, BRISK, KAZE. All these algorithms are invariant to scale, rotation and affine changes [12] but ORB is known to be faster and require less computational resources [13]. Therefore, this report uses ORB and SVM as the representation of the traditional machine learning method for classification of medical images.

B. Weakly supervised feature localization with CNN

Understanding how convolutional neural networks (CNNs) make decisions could help us locate certain regions within an image that contain the most relevant information and could summarize the contents of the whole image. This could be very useful in the biomedical field, as it could assist medical experts in locating salient features within a large scene of medical features, thus greatly reducing wait time in medical diagnosis. An approach for achieving this could be visualizing the activation maps after training a CNN classifier, as it could offer us useful insights on certain important features within an image. A recent research in this field is called weakly supervised learning [14], where using only weak supervision such as image class labels, we obtain regions of certain actions within such an image. This could be very beneficial in object localization tasks as it reduces the cost of image annotations. Several authors in the literature have proposed different approaches, one very popular paper was by [15] where the authors proposed a technique called class activation mapping (CAM) that uses a deep CNN for performing object localization without using any bounding box annotations. The technique involved using a global average pooling layer, and a SoftMax activation function as a classification head of a deep CNN, then after training, taking a weighted linear sum of the learned features of the penultimate convolutional layer of the network. The authors explained that the class activation mapping which was simply a weighted linear sum of the presence of visual patterns at different spatial locations in an image, such that by simply resizing the class activation maps to the size of the input image, we can identify the image regions most relevant to a particular class. This technique was demonstrated to achieve about 37.1 % top-5 error for object localization in ILVSVRC 2014 [16] without requiring any bounding box annotations. The drawback of the previously mentioned approach was that it may not be effective in locating small features or structures within an image. This was further improved upon by [17], where the authors proposed a method for learning fine pixelwise features from image-level labels by combining global pooling with the popular U-Net [18] architecture. Unlike the previous work that uses a SoftMax classification network, the weights of the network for this architecture were optimized to solve a regression task, where the objective is to predict the number of lesions. They demonstrated the potential of the method to detect enlarged perivascular spaces in brain MRI.

C. Autoencoder Network

There are more healthy people than people with a specific disease. Medical image data from the larger group of healthy people is more readily available and employing experts to classify and label this data not always necessary. Autoencoders can be useful in this situation as they can be used for classification purposes as long as training data for one class is available. The basic classification principle is that these autoencoders simply identify anomalous data compared to training data. To achieve this, the input data is compressed into a lower-dimensional representation, and then reconstructed as well as possible using knowledge about the core features inherent to the data set. Autoencoders are expected ”to obtain the true nature of the data, without uninteresting features and noise.” [19] This can be used to find and then utilize low dimensionality features that represent the data well. As a side effect however, input data that lacks key features, or has alterations of the key features, will fail to be encoded and decoded properly. Poor reconstructions therefore indicate anomalous input data that differs in the key features identified by the autoencoder. To detect a specific class, one might train the autoencoder with training data from all classes except from the target class. To detect new, unseen classes, the autoencoder can be trained with data from all known classes. The autoencoder could also be trained on all ”normal” training data with the goal of detecting any type of anomaly, and then analyzing the specific reconstruction errors further to differentiate between different types of anomalies. This was our goal for the autoencoder in this project.

In basic autoencoders, excessive noise and the existence of uninteresting features can also result in poor reconstructions. The performance and robustness of autoencoders can be improved by differentiating between noise and outlier features, and noise-free data. This separation can be performed on-the-fly by integrating additional filter layers. Such a filter layer performs a function similar to a Principal Component Analysis. The filter layer makes use of the assumption that ”noise and outliers are essentially incompressible” [20] to cull out anomalous parts of the data that are difficult to reconstruct. Apart from making identifying the core features of the data set more robust, the outputs of the filter layer itself could also be useful in detecting anomalies.

For some medical imaging classification applications it is beneficial to make use of both the broad context provided by the sparse core feature representation of a basic autoencoder and detailed, localized insights of a convolutional layer kernels. The U-Net architecture [18] combines both ideas to enable precise pixel by pixel classifications. Without making use of pixel-wise classifications, our project tests whether this

method of adding localized data can improve binary anomaly detection as well.

III. Method

A. Feature Detection & Classification with ORB-SVM

The choice of ORB as the feature detector-descriptor was a critical decision in our application and based on literature review of high quality research articles that provide comparison results of various types of feature detection techniques (SIFT, SURF, ORB, KAZE). The key-points were extracted with the respective descriptors computed using skimage library. The descriptors generated by ORB is 256-dimensional and each image yielded a variable number of descriptors, so we fixed the number of detection to 100 key-point descriptors. In order to train an SVM classifier we require the feature matrix and label vector, where the feature matrix or descriptors are calculated from interest points of the ORB technique descriptors followed by labels retrieved by k-means clustering. The clustering method is able to deal with binary string nature of the ORB descriptors. However the corresponding descriptors are high dimensional for each interest point therefore, we subsequently perform dimensionality reduction using PCA to retain 80% variance. It is common practice to apply PCA before a clustering algorithm (such as k-means), with the expectation to improve the clustering results by reducing noise. Finally, labels were obtained by performing k-means clustering on the descriptors where each cluster represented features present in abnormal, normal and those that are present in both normal and abnormal. Then we trained the SVM classifier, a notable benefit of SVM algorithm is that it performs well with high dimensionality data. In the testing stage, samples were randomly chosen from the dataset and their respective key-point descriptors retrieved using ORB then fed to SVM as a feature matrix with labels obtained by performing k-means clustering in the training stage. An additional step was introduced to sketch patches to indicate regions of interest in the sample image. In this way, each time new samples were introduced for testing, we simply extracted the feature-descriptors using ORB around regions of interest followed by classification with the trained SVM without the additional steps for PCA and Clustering for each sample x-ray image. The proposed workflow is demonstrated in Figure 1.

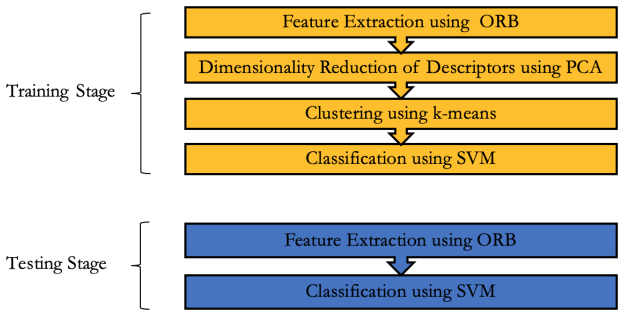

Fig. 1: Feature detection and classification steps

B. Weakly supervised feature localization with CNN

Taking inspiration from the previously mentioned work on feature extraction using deep CNNs, we present an approach for salient features localization using only image class labels. The methodology involves three steps:

Step 1:

We first train a CNN autoencoder architecture for learning the latent space for capturing relevant information within an image. The architecture used is the popular U-NET architecture which uses a series of down-sampling and up-sampling layers concatenated together for reconstruction of the given input image. This architecture was selected because of its high performance in segmentation tasks as reported in the literature [18]. The architecture is shown in the appendix in Figure 10.

Step 2:

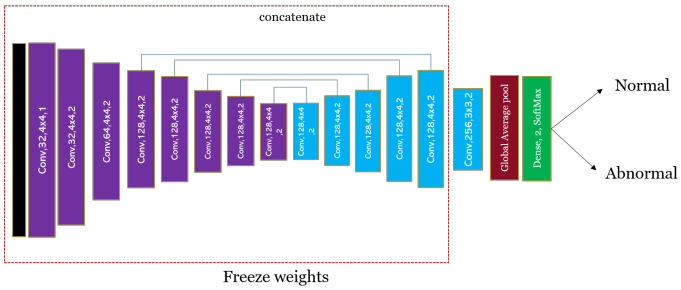

Fig. 2: Transfer learning

After training the autoencoder, we then experimentally remove the topmost layers of the decoder network while freezing the remaining layer weights. After which we attached a new convolutional, global average pooling and SoftMax activation layer as shown in Figure 2. Then we retrained the newly introduced layers for predicting the class labels of the given classification task.

Step 3:

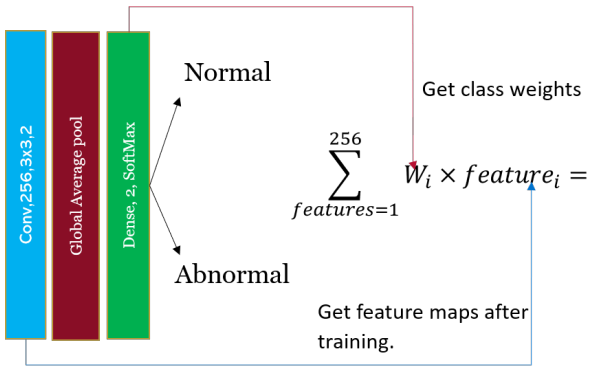

Fig. 3: Compute feature activation's maps

During live testing of the network for feature extraction, we take a weighted sum of the feature maps of the penultimate convolutional layer as shown in Figure 3, then resize the resulting map to the size of the input image to get a heat-map

that shows how several regions of the image are activated and which could be used to locate certain salient regions within such images.

C. Autoencoder Network

We implemented and tested three different autoencoders: Two versions of a simple autoencoder, and a version of the U-Net architecture. The preprocessing steps are the same for all autoencoders. Outliers, i.e., images that do not depict front facing X-rays, are pruned manually. Images are resized to a size of 1024x1024 pixel, converted to gray scale in float32 representation, and pixel values are scaled down from a range of 0 to 255 to a range between 0 and 1. All available images that were classified as "normal" are used for training, and all images classified as "pneumonia" are used for validation. All autoencoders were trained using the adam optimizer with a starting learning rate of 0.0001 and use mean squared error as the loss function.

The first simple autoencoder contains three sets of layer combinations including a 2D convolutional layer with kernel size 3x3, a batch normalization layer, and a max pooling layer with stride size (2,2). The result of this encoding path is then decoded by another three sets of layer combinations, in which the max pooling layer is replaced by a 2D up-sampling layer with the same stride size. All of these six convolutional layers use the relu activation function and are initialized using He normal initialization to mimic the U-Net architecture and to make results more comparable. A final sigmoid acticated convolutional layer provides the image shaped output. The number of kernels per convolutional layer starts at 16 and doubles with the encoding progression before returning to 16 during decoding. This autoencoder encodes the input image down to a representation that is 64x smaller than the original size.

The second version of the simple autoencoder mimics the first, but it scales the image data four times as much at every set of layers by increasing the stride size to (4,4). In this version, the deepest encoded representation is 4096x smaller than the original.

Finally, we implemented a version of the U-Net. Figure 15 in the appendix depicts the details of this architecture. This autoencoder uses the same kernel size and activation and initialization functions for convolutional layers, and follows a similar progression in the number of kernels, starting with 16 kernels. Apart from handling input and output of the network, this is the only change we made compared to the original U-Net architecture. Instead of using batch normalization, the U-Net makes use of dropout layers. It features four tiers of encoding with a stride size of (2,2) instead of three, but the deepest encoded representation reaches roughly 1000x smaller than the original input data due to the way the architecture handles the edges of layers to promote accuracy in overlapping regions.

IV. Results & Analysis

The three feature extraction techniques discussed were tested on a Kaggle dataset, which consisted of chest x-ray images of about 5800 healthy and pneumonia infected patients, categorized into Covid-19, SARS (Severe Acute Respiratory Syndrome), Streptococcus & ARDS (Acute Respiratory Distress Syndrome), collected from different sources [21]. The data distribution ratio was 27% healthy and 73% unhealthy chest x-ray images. The pattern recognition task was to classify healthy and abnormal x-ray images and to determine regions of interest that influence the distinction. Each of the above described methods was tested on a different system configuration, which are specified in the following subsections.

A. Feature Extraction & Classification with ORB-SVM

In this paper the experiment with the ORB feature extractor with an SVM classifier which was trained on an ordinary performance system with the following configuration of RAM:8G, GPU:2GB consumer grade, CPU:i5 (2.7GHz) dual processor. This traditional approach of ORB with SVM has many features and classifying algorithms that can be evaluated. The following results were obtained for 500 samples with each image 256x256 pixels in dimension. After locating key-points (pre-set to 100), BRIEF takes all the key-points found by the FAST algorithm and converts it into binary feature vector so that together they can represent an object. It is then accompanied by dimensionality reduction using PCA which maps the existing descriptors from 256 dimensions to 110 dimensions with 80% variance. The main objective of the PCA is to retain existing information after reducing dimension of the descriptors. The accuracy is greatly improved by using ORB-PCA compared to just ORB. Since no ground truth labels of the data are known, we use k-means algorithm for the formation of precise clusters to increase accuracy. With the possible domain knowledge of normal, abnormal and mutual features present in both samples we chose a partition of k=3 clusters. Where the least frequency of occurrence was assigned as abnormal features followed by normal features and highest occurrence assigned to features present in both normal and abnormal samples. SVMs are effective when training high dimensional spaces, when training the SVM we considered one of two hyperparameters, such as the Radial Basis Function (RBF) : c and gamma. The choice of the regularization parameter c, is crucial to avoid overfitting, especially if the number of features is higher than the number of samples. In our case setting the regularization parameter, c = 1 improves accuracy the most. After fitting the classifier for cluster predictions, we achieve an r-squared cross validation score of 92.79%. Finally, in the testing stage we manually test normal and abnormal images where orb extractor obtains the descriptors which is then fed into the SVM for end-to-end feature learning. For visualization (Figure 4) the last step of the test draws circles around regions of interest in the image where green indicate normal, red indicate abnormal and blue indicate the features belonging to both classes.

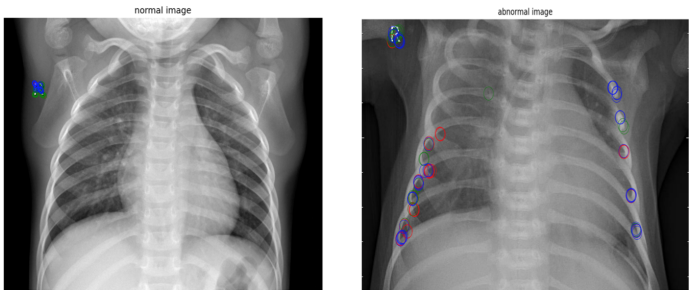

Fig. 4: Patch drawing on healthy and unhealthy chest x-ray

A notable observation is that the feature detector also detected the usual medical labelling on x-ray images, careful consideration needs to be made to avoid false detection as such.

B. Weakly supervised feature localization with CNN

The experiment was carried out using the tensor-flow deep learning framework on a system configuration of RAM:16GB, GPU:4GB, NVidia GTX 960M, and core-i7 quad core processor. During the autoencoder training phase, the input-image was resized to 256 x 256, then each of the encoder and decoder layers excluding the output layer was followed by series of Batch Normalization and ReLu activation function. The training was carried out for 5 epochs using the Adam optimizer and the mean-square error loss function. During training, the model validation mean-square error loss was monitored using the keras callback function in order to obtain the best model. During the transfer learning training phase, we employed a 90% train-test split, using same Adam optimizer as before with the categorical crossentropy loss function. We also employed the keras callback validation loss monitoring function for obtaining the best model and trained for 100 epochs. The training curve for the model can be shown in Appendix, and Figure 7 shows several activation maps gotten using this technique. From the heat-map we observed that for the healthy images, regions close to the stomach and liver gets activated, while regions close to the pulmonary artery gets activated for unhealthy images.

C. Autoencoder Network

All three autoencoders were trained for 500 epochs on a system configuration of RAM:32GB, GPU:24GB NVidia RTX3090, and CPU:Ryzen 5900X twelve core processor. Training the simple autoencoder architectures took less than 2 hours, while training the U-Net based architecture took about 10 hours.

Out of the three autoencoders, the U-Net based architecture achieves the best separation between the loss distributions of autoencoded healthy and abnormal X-ray chest images, as shown in Figure 6. Since the distributions still overlap, this autoencoder does not perform well enough to reliably detect anomalous images from the data set. The histograms for the two simple autoencoders can be found in the appendix. The simple autoencoder with a max. size reduction of 4096

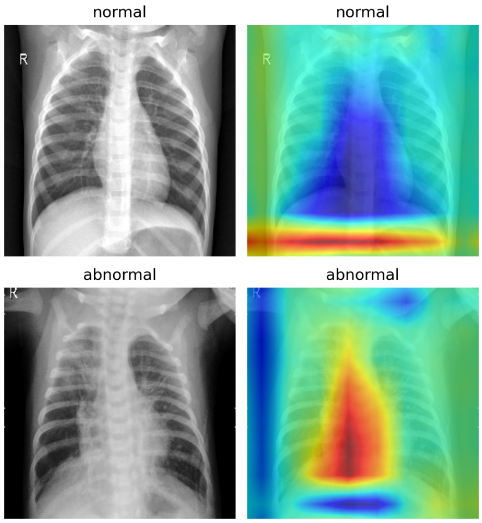

Fig. 5: heat-map of healthy and unhealthy chest x-ray

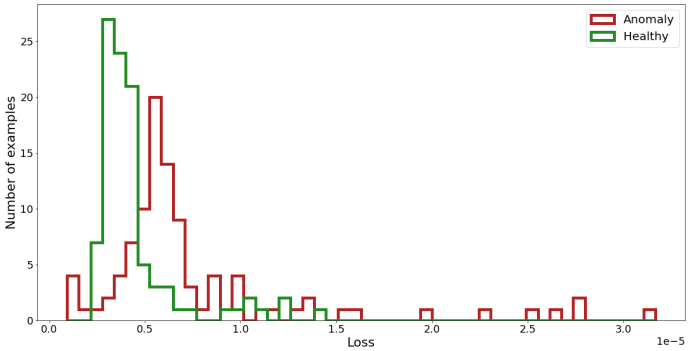

Fig. 6: Loss distribution for reconstruction differences using the U-Net based autoencoder.

times separated the loss distributions more than the one with a reduction of only 64 times, for which healthy and abnormal results are essentially indistinguishable. However, anomaly images were reconstructed with on average less error than healthy images by the 4096x autoencoder, which is a surprising result.

We further performed visual inspections of the shapes of the remaining losses, i.e., of the differences between input images and reconstructed images. Both the 64-times reduction autoencoder and the U-Net based architecture achieved reconstructions that are visually indistinguishable from the original, as shown in Figure 7. The reconstructed images of the 6094-times reduction autoencoder are noticeably blurry and the remaining losses focus uniformly around high-frequency and high-contrast areas of the original images. It is possible that the healthy images were reconstructed less well by this autoencoder since the pneumonia anomalies might cloud the X-ray images, which conceals high-frequency details. For the other two autoencoders, the remaining losses seems to center around artifacts in the images, such as letters included in the images. Since the letters are essentially noise artifacts

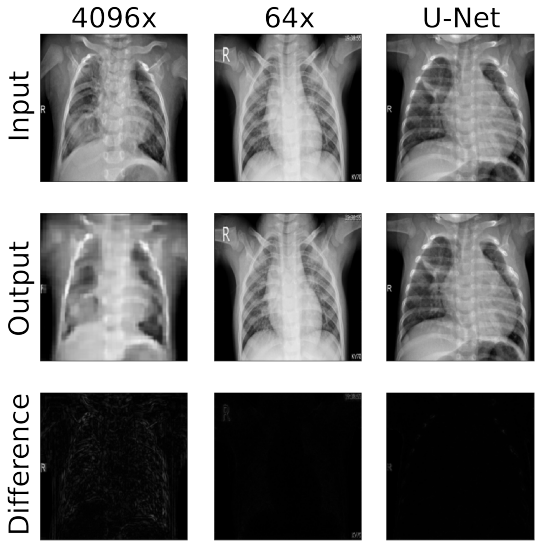

Fig. 7: Comparison between pneumonia input images, reconstructed images, and their differences, for three autoencoders.

in the images, this is not a promising result. Filtering these features as described in [20] might alleviate the problem. It seems that overall, the visual distinctions between anomalous and healthy images might fall within the range of features that an autoencoder has to learn to reconstruct X-ray images considered healthy. For example, differences in the contrast based on the imaging equipment, differences in body types between children and adults, and also patients with a common cold rather than pneumonia, which might already exhibit similar, albeit less severe, features. It appears that the strengths of the U-Net based architecture could not come to fruiting in this set-up and that it might be more successful in classifying infected area pixels in the lungs, rather than simply autoencoding the given images. The progression of training and anomaly validation losses does not indicate signs of traditional overfitting either. A plot of the progression can be found in the appendix, Figure 14. Though consistency would be much lower, it is possible that training this autoencoder for less epochs could yield beneficial overall performance results. When applying the U-Net based autoencoder on a gray scale image of a human face, i.e., a definite anomaly, the resulting image was a visually almost perfect reconstruction of the input image. The reconstruction loss was comparatively extremely high, and could easily be used to classify the image as an anomaly. Overall, autoencoders may not be well suited for anomaly detection in X-ray chest image data.

V. Conclusions

We present classification results for three different machine learning approaches. We demonstrate that it is possible to obtain powerful representations of biomedical concepts via both deep learning and traditional techniques, partly achieving results on par with previously investigated literature in the field. All three approaches studied in this work offer a variety of further enhancement features and classification algorithm options.

The deep learning approaches like weakly supervised feature localization with CNN and Autoencoder networks can provide decent results with additional complexity of convergence. Nonetheless, this is an issue that can be further evaluated by training and comparing different representational learning models such as ResNet and Inception and other forms of autoencoders such as Sparse or Denoising. For the traditional method using popular computer vision methods such as SIFT, Bags of (Visual) Words model and different Clustering algorithms can be studied in a comparative analysis. The scope for further enhancement is manifold and likely to lead to considerable improvements as well as expand the scope for research in automating anomaly detection in digital medical images. Overall, all three methods address the challenge with some level of certainty.

Future work will consider better evaluation tools for a comparative analysis between the performances of the different approaches. Subsequently, if there are partly known annotations, more descriptive representations can be studied by means of semi-supervised learning. Finally, for realising the performance in a practical realm, these models can be evaluated through information retrieval systems in clinical setups.

VI. Data Availability Statement

The *CoronaHack-Chest X-Ray-Dataset* used in this paper is publicly available and downloadable from the following link: https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset. [21]

VII. Authors Contributions & Remarks

The three approaches were based on traditional machine learning and deep neural networks for feature learning where authors Osazee Ero worked on weakly supervised feature localization with CNN, Ludwig Wilhelm Wall on variations of the Autoencoder network and Dorodi Afroze Krishty on ORB with SVM.

References

[1] Andy Chiem, Adel Al-Jumaily, and Rami N Khushaba. A novel hybrid system for skin lesion detection. In *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*, pages 567–572. IEEE, 2007.

[2] Ilias Maglogiannis, Elias Zafiropoulos, and Christos Kyranoudis. Intelligent segmentation and classification of pigmented skin lesions in dermatological images. In *Hellenic Conference on Artificial Intelligence*, pages 214–223. Springer, 2006.

[3] Y Rejani and S Thamarai Selvi. Early detection of breast cancer using svm classifier technique. *arXiv preprint arXiv:0912.2314*, 2009.

[4] Leonardo de Oliveira Martins, Aristófanes Corrêa Silva, Anselmo Cardoso De Paiva, and Marcelo Gattass. Detection of breast masses in mammogram images using growing neural gas algorithm and ripley's k function. *Journal of Signal Processing Systems*, 55(1):77–90, 2009.

[5] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[6] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[7] Juan C Caicedo, Angel Cruz, and Fabio A Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 126–135. Springer, 2009.

[8] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[9] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18, 2019.

[10] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer, 2010.

[11] Mahmoud Hassaballah, Aly Amin Abdelmgeid, and Hammam A Alshazly. Image features detection, description and matching. In *Image Feature Detectors and Descriptors*, pages 11–45. Springer, 2016.

[12] Shaharyar Ahmed Khan Tareen and Zahra Saleem. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In *2018 International conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–10. IEEE, 2018.

[13] Eduardo Pinho and Carlos Costa. Unsupervised learning for concept detection in medical images: A comparative analysis. *Applied Sciences*, 8(8):1213, 2018.

[14] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[15] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[17] Florian Dubost, Gerda Bortsova, Hieab Adams, Arfan Ikram, Wiro J Niessen, Meike Vernooij, and Marleen De Bruijne. Gp-unet: Lesion detection from weak labels with a 3d regression network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–221. Springer, 2017.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[19] O. Lyudchik, J. Vlimant, and M. Pierini. Outlier detection using autoencoders. In *CERN-STUDENTS-Note-2016-079*, 2016.

[20] Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 665–674, New York, NY, USA, 2017. Association for Computing Machinery.

[21] Praveen. CoronaHack-Chest X-Ray-Dataset. https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset/, 2020. [Online; accessed 24-April-2021].

APPENDIX

Fig. 8: K-means inertia plot for partitioning data into clusters.

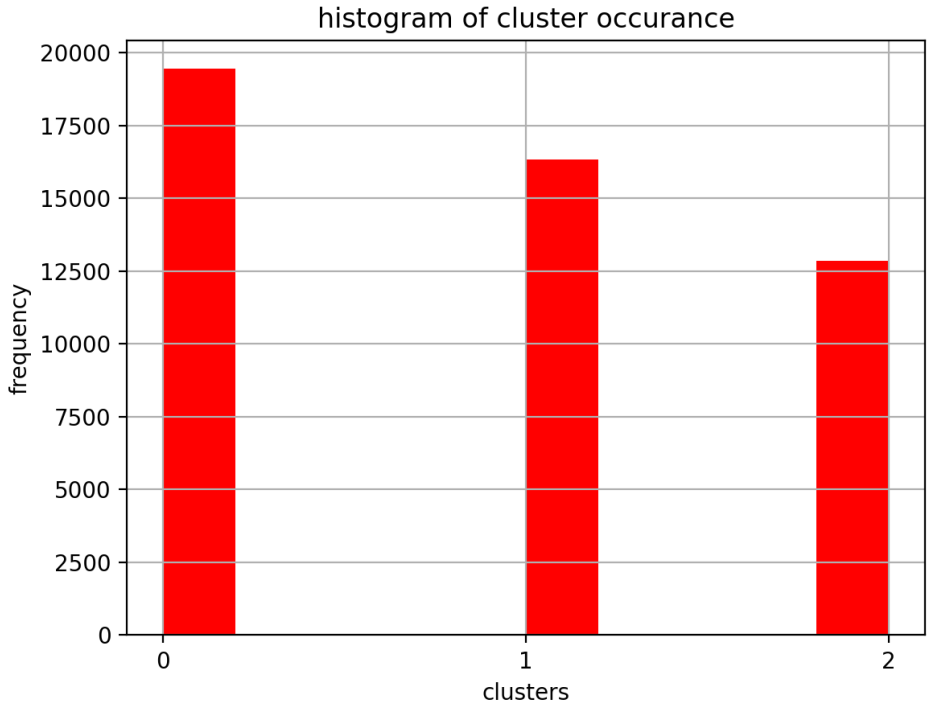

Fig. 9: K-Means clustering for data without ground-truth label.

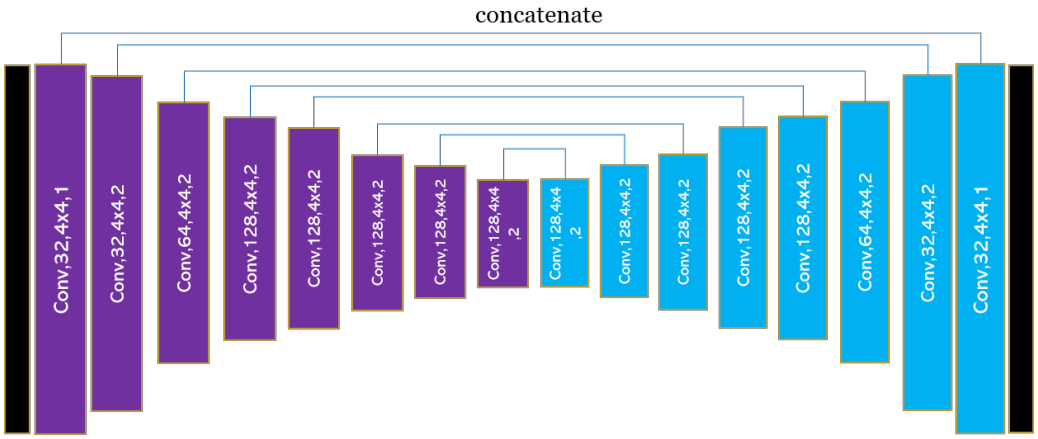

Fig. 10: U-Net like auto-encoder

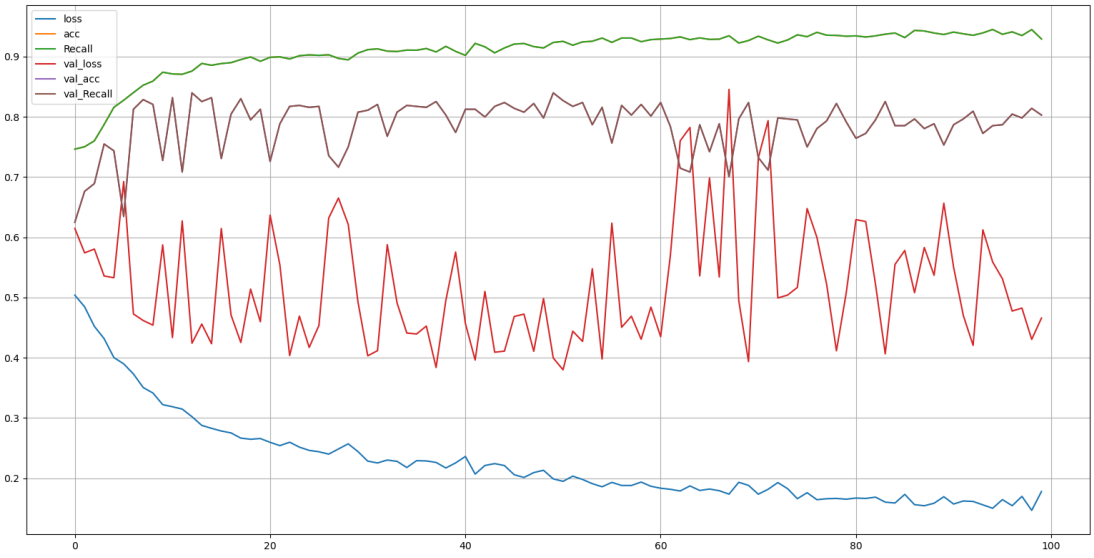

Fig. 11: Learning curve for Weakly Supervised Feature Localization with CNN

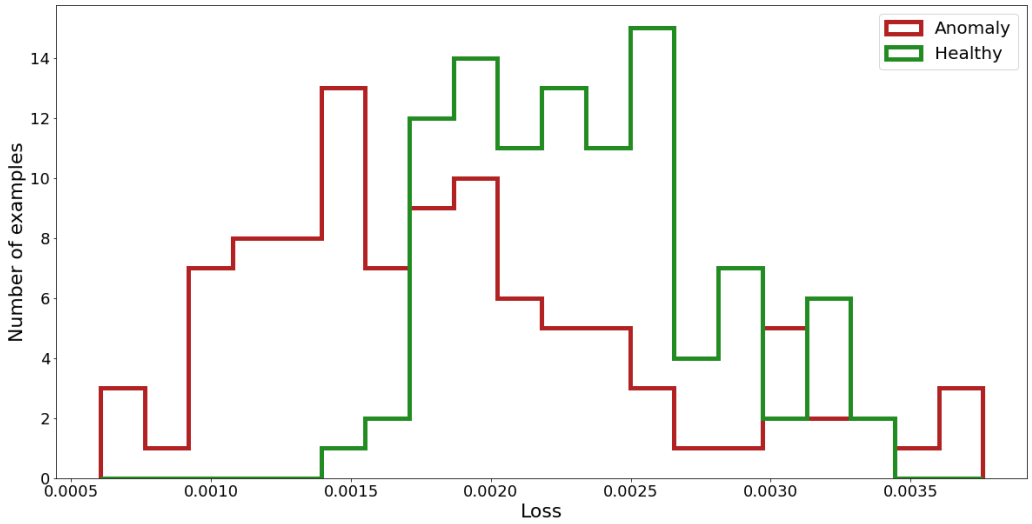

Fig. 12: Loss distribution for reconstruction differences using the 4096x reduction autoencoder.

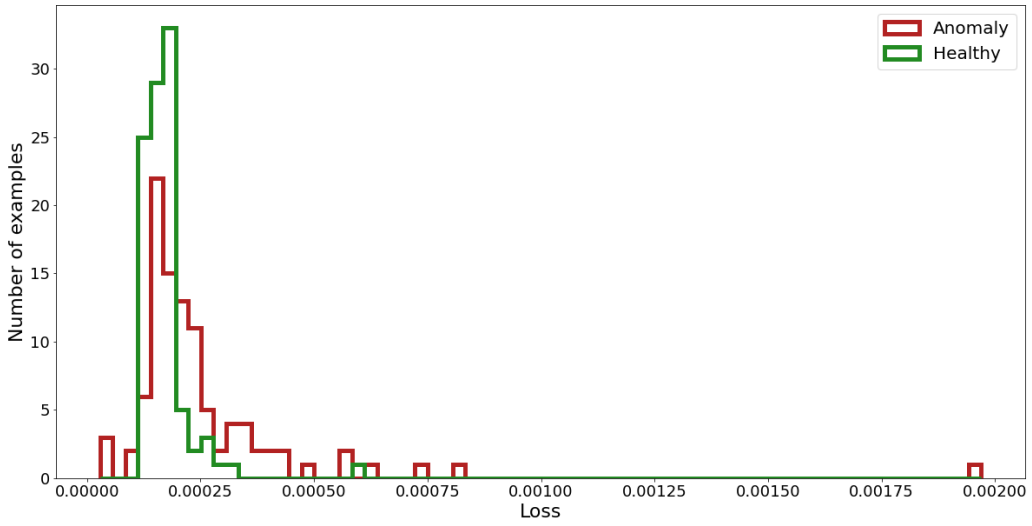

Fig. 13: Loss distribution for reconstruction differences using the 64x reduction autoencoder.

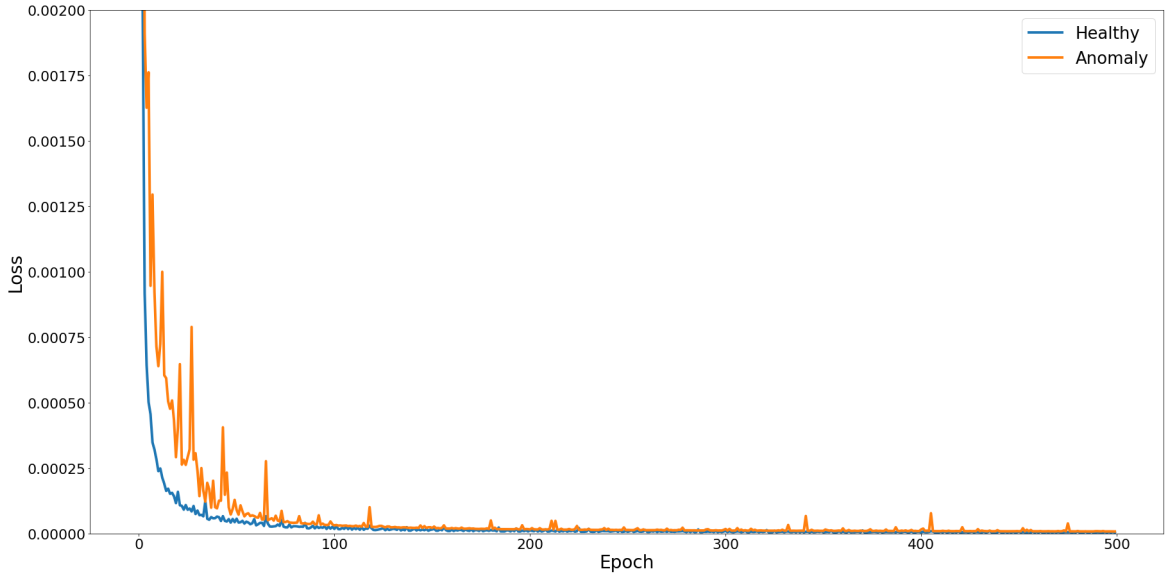

Fig. 14: Progression of training and anomaly validation losses for the U-Net based autoencoder.

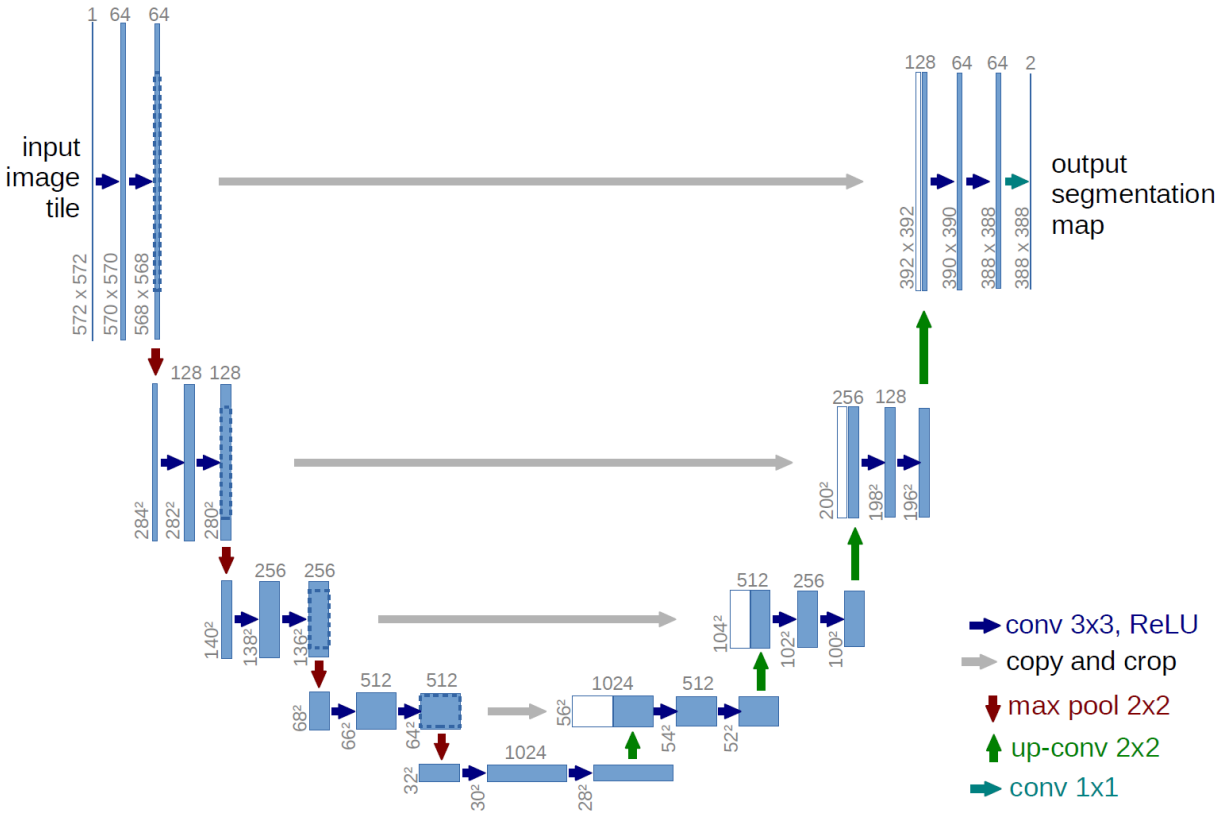

Fig. 15: U-Net architecture for medical image classification. Image was part of previous work. [18]