Comparing Supervised and Unsupervised Models for Disease Detection

VIVAN BIJU, University of Twente, The Netherlands

The automation of detecting rare diseases accurately has proven to be challenging due to the lack of enough data on cases with such diseases. This research explores and compares the effectiveness of supervised and unsupervised machine learning methods to identify rare disease patterns in medical imaging. Supervised learning methods are known to be very accurate but unsupervised methods are adept at handling unlabelled data and identifying the anomalies that exist. Thus, a thorough comparison will bring more clarity on what methods to prefer for anomaly detection. This research also focuses primarily on using different types of autoencoders to detect anomalies in medical images. To thoroughly assess the supervised models(Resnet50 and Densenet121) and the unsupervised models(Autoencoders and Variational Autoencoders), this study will make use of multiple datasets: Retinal OCT Images, Brain Tumor MRI Scans, COVID-19 Radiography images and ISIC 2018 HAM10000 dataset. By conducting this comparative analysis, this research aims to shed light on suitable machine learning models for the use of detection of diseases in medical images.

Additional Key Words and Phrases: Supervised machine learning, Unsupervised machine learning, Autoencoders, Variational autoencoders, Masked autoencoders, Convolutional Neural Networks

1 INTRODUCTION

Despite being "rare", rare diseases on the whole impact a large percentage of the population worldwide, even if each condition alone affects a relatively small number of people. This means that while we have a lot of cumulative data when we look at all diseases if we look into certain diseases in specific, the datasets become a lot sparser. Due to this, there is an immense need for different approaches to analyse data and detect and treat these diseases effectively.

In this context, machine learning(ML) emerges as a potent tool. Using ML, we can make predictions or decisions when trying to detect rare diseases. This thesis explores two primary categories of ML, namely supervised and unsupervised methods of learning. Supervised learning methods are those that depend on labelled datasets being available as a prerequisite so that models(algorithms) can be trained on them to produce effective results when dealing with new data. Unsupervised learning methods, on the other hand, do not involve any prerequisite labelled data. The models look for patterns and try to classify them based on these patterns. Many models like Resnet50 and Densenet121, for supervised methods, and autoencoders(AE) and variational autoencoders(VAE), for unsupervised methods, have been used in the field of medicine for this purpose. We shall discuss more in detail how these models work later, but for now, let us look at the uses in different research.

An issue that does arise from this is that it is not immediately apparent which models would be suitable for use in different datasets of medical images. By conducting a thorough analysis of various models, it is possible to gain a better insight into the most suitable models for disease detection in medical images.

Hamlili et al. explore the use of the Resnet50 model, a type of convolutional neural network(CNN) to derive an effective model for the classification of x-ray and CT images of lungs for the detection of COVID-19[3]. Another research conducted by Zebari et al. focus[16]. Both Resnet50 and Densenet121 are supervised models which proved to give results of high quality.

In terms of unsupervised models, this research mainly focuses on the use of autoencoders. These are models which encode images to represent encodings that capture the most essential features of the images, and then decode them and attempt to reconstruct the original images using just the encoded data. Applying a different approach for the detection of COVID-19 in radiography images, a method involving the use of deep CNNs along with a regular AE yielded promising results[4]. Another research also uses a VAE - a type of autoencoder that outputs a distribution of probable outputs, rather than just one reconstructed image - on the HAM10000(Human Against Machine) dataset, provided by ISIC as a challenge in 2018 for the detection of skin cancer in images of skin lesions which showed fruitful outcomes[8].

All these models are valid means of detection of the different diseases. This research, however, takes the experiments a step further to test the effectiveness of the different models on various datasets, rather than just one. Performing a strong comparative analysis in this manner allows for a better assessment of the effectiveness of the different models as well as how applicable they are for the different datasets, by examining the pros and cons of the different models used.

2 PROBLEM STATEMENT

It is undeniable that there has been a substantial amount of research in the field of anomaly detection, both, for supervised and unsupervised learning methods. However, an understanding of which methods to use in which context requires an in-depth analysis of how effectively the different methods work. This brings up the research questions we will aim to answer in this paper.

2.1 Research Question

What is the comparative effectiveness of supervised and unsupervised learning methods in detecting rare diseases from limited data?

This question can be challenging to answer on its own and can be better answered through the following sub-questions:

2.1.1 ***SQ1:*** In terms of the various metrics available, how do supervised and unsupervised models perform on the datasets?

41st Twente Student Conference on IT, July 05, 2024, Enschede, The Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2.1.2 ***SQ2:*** To what extent can fine-tuning and transfer learning aid the models in producing better results?

3 RELATED WORK

In this section, we will delve deeper into the related work on disease detection using various learning methods. It is worth understanding the concepts that apply when discussing the different models that were experimented with previously as this will also help understand the models used in this research.

Resnet50 and Densenet121 are two supervised methods and are convolutional neural networks(CNNs). CNNs are a type of model that makes use of layers of convolutions that essentially take images as matrices of pixels and perform mathematical operations using certain types of matrices (called "filters") which gather the more important features in the images. These are layered to grasp and extract as many features as possible. Edges and brightness are good examples of features that are extracted using this method. With each layer, the image is reduced further to the most important features with the help of a pooling layer. There are also non-linearity layers to represent the images better since images cannot be represented as linear functions[9]. As their name implies, Resnet50 has 50 layers, while Densenet121 has 121 layers. Resnet50 makes use of skip connections which allow it to bypass some layers if required, whereas the layers in Densenet121 are densely connected, where each layer receives all information for every preceding layer and provides its information over to succeeding layers[13][14].

The paper written by Hamlili et al. discusses an attempt at transfer learning with Resnet50 for use on radiography images of the chest, to handle a multi-class classification of a dataset that consists of COVID-19 positive, viral pneumonia, bacterial pneumonia and normal images[3]. The dataset was preprocessed to change the contrast, and denoised so that the anomalies were easier to see, and the images were resized to the dimensions that are required by Resnet50(224x224 pixels). The images also underwent transformations(rotations, reflections, translations) and normalisation. Then, using the base model of Resnet50 combined with a deep CNN, the model was trained and was able to produce close to perfect results. Specifically, the overall accuracy was 95.2%, with an average sensitivity of 95.6%, specificity of 98.4%, and precision and F1-score of 95.3%.

Similarly, for a dataset for brain tumour classification, Zebari et al. develop a model, with Densenet121 playing a vital role in their architecture[16]. This dataset was meant for a binary classification problem, so the images were either indicative of a tumour, or not. Similar to the research involving Resnet50[3], the images were preprocessed with various transformations and resized as per Densenet121's requirements(224x224 pixels). These images were also denoised and normalised. For this research, Densenet121 along with the final layers for the classification of the image were sufficient to yield effective results. When tested on unseen data, the model had a recall of 89.04%, which was slightly below their train recall of 92.87%, for the brain tumour class, and the normal image class also seemed to exhibit a slight reduction in metrics.

The principles of CNNs allow us to better understand how autoencoders work. There are many types of autoencoders(this paper focuses on regular and VAEs). They work on a similar idea; an encoder uses convolutions to encode images into their most important features and a decoder reconstructs the image from this encoded form as best as it can. In its encoded form, the image exists in what is called a "latent space". The encoder and decoder model together are what builds an autoencoder. These have been crucial for research regarding anomaly detection in medical images.

There has been research done to detect skin diseases using a VAE [8]. Here the dataset used was the ISIC 2018 HAM10000 challenge dataset. The decoder in this type of autoencoder produces a Gaussian distribution from which it samples the latent state. By doing this, the resulting output is a reconstruction of the input but is not part of the input set. Using this model yielded high AUC values(0.77) for the reconstruction scores and showed promising results for what seems to be one of the first attempts at using unsupervised learning methods to detect skin diseases. This score for reconstruction allowed for melanoma to be detected with 86% accuracy. Interestingly, during preprocessing, the images were normalised between a range of -1 to 1, rather than the standard 0 to 1. The images were resized to 128x128 pixels here.

Finally, a study done on a different dataset for images of chest x-rays to detect COVID-19 made use of a combination of a CNN and a regular AE[4]. The data is preprocessed to a size of 254x254 pixels and sent through similar augmentations as the other studies, including rotations, translations, reflections, and also normalized and denoised. The model is an AE with a large number of convolutional layers that bring down an input image of 254x254 pixels reduced to 31x31 pixels, which hold the extracted features. These are then used by the decoder to reconstruct the data. Once again, promising results were obtained, with a high accuracy of 98%.

Clearly, there have been many areas explored, all of which have been proven to produce effective results for the detection of different diseases. However, it is still unclear as to which methods work best for which scenarios, and there is no analytical comparison of which methods perform better than the other. This paper aims to get a better insight on what models could be the most suitable in different scenarios.

4 METHODOLOGIES

This section will highlight the methods taken to answer the research question. We will make use of existing models which can be fine-tuned to work for the datasets we will use. This research makes use of four datasets and four models for a comparative analysis. Below is a list of these datasets and models:

Datasets

- Retinal OCT Images (optical coherence tomography)[7]
- Brain Tumor MRI Dataset[10]
- COVID-19 Radiography Database[1][11][12]
- ISIC 2018 HAM10000[15]

Models

- Resnet50[5]
- Densenet121[6]
- Regular Autoencoder

- Variational Autoencoder

The implementations primarily make use of Tensorflow Keras. The "applications" module available with Keras provides the models for Resnet50 and Densenet121. The autoencoders were made from scratch using Keras too. The Retinal OCT Images, Brain tumour MRI, and COVID-19 Radiography datasets were obtained through Kaggle, and the ISIC2018 HAM10000 dataset was obtained from the official ISIC archive website.

Each of the models is built to perform a multi-class classification on each dataset used. For the supervised methods, each dataset is organised to have "train", "val", and "test" sub-directories, for the training, validation and test data respectively. The split of the Covid dataset is done such that 70% data is utilised for training, 10% for validation and 20% for testing(unseen data). For the tumour, retina and melanoma datasets, the split was already predetermined and the exact number of images will be mentioned later in the experiment section. For the unsupervised models, the datasets are organised such that the model is trained only on normal images. So, for ISIC, the model is trained and validated only on the benign keratosis class and tested with each anomaly to see how many are detected. A similar split is made for the tumour dataset with the "notumor" class, "NORMAL" class for the retina dataset, and "covid" class for the COVID dataset. The AE and VAE are then tested on the anomaly classes to see how many of the anomalous images are indeed recognised to have anomalies. Below is the list of classes in each of the datasets:

- **Retinal OCT Images** [7]
  - **Dataset:** Optical Coherence Tomography (OCT) images of the retina.
  - **Classes:**
    - CNV (Choroidal Neovascularization)
    - DME (Diabetic Macular Edema)
    - DRUSEN
    - NORMAL
- **Brain Tumor MRI**[10]
  - **Dataset:** Magnetic Resonance Imaging (MRI) scans for brain tumor detection.
  - **Classes:**
    - Glioma
    - Meningioma
    - Pituitary
    - notumor (No tumor)
- **COVID-19 Radiography**[1][11][12]
  - **Dataset:** Radiographic images used to detect COVID-19 and other respiratory conditions.
  - **Classes:**
    - COVID
    - Lung_Opacity
    - Viral Pneumonia
    - Normal
- **HAM10000**[15]
  - **Dataset:** Dermoscopic images from the HAM10000 dataset.
  - **Classes:**
    - Melanoma
    - Melanocytic nevus
    - Basal cell carcinoma
    - Actinic keratosis / Bowen's disease (intraepithelial carcinoma)
    - Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis)
    - Dermatofibroma
    - Vascular lesion

Moving on to the sub-research questions, below is an overview of how the experiments are set up and conducted to answer each of the questions.

4.1 On answering SQ1:

For each model, we take a collection of metrics. Firstly, we determine the accuracy, precision and recall with which each of the classes is determined. This helps us understand if the classes of images from the unseen data are correctly and consistently identified, and check if the classes identified are truly relevant to the results we need to see, which in our case, are cases where the anomalies(diseases) exist in the image.

$$Accuracy = \frac{Number of correct predictions}{Total number of predictions}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Apart from these standard metrics, we also want to measure F1-scores for the supervised models. This allows us to balance the precision and recall and make sure that underrepresented classes - which in our case are the positive cases of diseases, since they are rare - are identified well. The unsupervised models do not require this metric since they are already being assessed to detect one anomaly at a time against the normal class, so we do not have to worry about other classes that might make the performance of the model seem better by a weighted average.

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Finally, we also keep an eye on losses to make sure that the models are not losing out on too much of the information obtained from the images while training. We also plot ROC graphs to observe the model's AUC for the different classes. A neat classification report will also show the scores for the metrics mentioned so far for each of the classes in each dataset.

4.2 On answering SQ2:

This sub-question will call for various tweaks in the model followed by observation of the metrics mentioned in SQ1. This includes varying learning rates, varying number of epochs for which the model trains, experimenting with different optimisers, and freezing and unfreezing layers in the pre-trained models.

5 APPROACH

5.1 Objective

The objective of this experiment is to evaluate and compare the performance of four machine learning models—ResNet50, DenseNet121, Autoencoder (AE), and Variational Autoencoder (VAE)—in classifying images of diseases across four distinct datasets. This study aims primarily to measure each model's effectiveness through accuracy, loss, precision, recall, and F1 score, and, we will also observe training times if the previously mentioned metrics are nearly equal between models. The outcomes will provide insights into the potential of these models to enhance diagnostic accuracy in medical imaging, guiding optimal model selection for specific disease contexts and contributing to the advancement of personalized medical interventions.

5.2 Experimental Setup

The supervised models required a different experimental setup compared to the unsupervised models, especially with regard to how the datasets are organised for training, validation and testing. In general, all images in every dataset were resized to 128x128 which allowed for models to be trained faster. The models were run on servers with NVIDIA A10 and A16 GPUs which allowed the supervised models to train at a speed of approximately one minute per epoch, whereas the unsupervised models trained at thirty seconds per epoch.

Image Augmentations

For all the models, all images undergo the same augmentations for training, validation and testing. These include rotating with a range of ±20, width and height shift range of ±20% of the width and height respectively, a zoom range of ±10% of the original image, and random horizontal and vertical flips. The images are normalised between values 0 and 1 for the supervised models and the regular autoencoder and between -1 and 1 for the variational autoencoder. For the validation and test set, only the normalisation and nothing else is performed. Finally, apart from the HAM10000 dataset, the images used in the unsupervised models are augmented to be grayscale. The HAM10000 dataset deals with melanoma which can have many intricate features that the model needs to learn and these could get overshadowed if the images were made grayscale. The supervised models use colour images.

Supervised

Dataset Organisation

The split of the different datasets for supervised methods is elaborated below. Each of the folders contains a proportion of all the different classes.

- **Retinal OCT Images** [7]
  - train - 83484 images
  - validation - 968 images
  - test - 32 images
- **Brain Tumor MRI**[10]
  - train - 5143 images
  - validation - 569 images
  - test - 1311 images
- **COVID-19 Radiography**[1][11][12]
  - train - 14815 images
  - validation - 2117 images
  - test - 4233 images
- **HAM10000**[15]
  - train - 10015 images
  - validation - 193 images
  - test - 1512 images

Models

The models used in for this experiment involve the use of Resnet50 and Densenet121 as base models, pre-trained on the "imagenet" dataset[2], both of which can be imported through the applications module of Tensorflow Keras. To apply transfer learning, all layers of both models are frozen initially. We build layers on top of the frozen models which will handle the classification of the images into their respective classes. After experimenting with different possible values, at this stage, both models returned best results when they had their hyperparameters set to:

- Batch size = 32
- Learning Rate = 0.001
- Number of Epochs: 20
- Optimizer = Adam
- Target size = (64, 64) for Densenet121 and (128,128) for Resnet50

Once this finishes training, we unfreeze all layers for both cases. We then retrain the model, with slightly different hyperparameters for both.

- For Resnet50 we only lower the learning rate to 0.0001 and train for 20 more epochs.
- For Densenet121 we lower the learning rate to 0.00007 and train for 30 more epochs.

This allows the models to be trained enough to be tested on unseen data. Throughout the training process, we observe the accuracy, precision, recall and loss.

Unsupervised

The split of the different datasets for the unsupervised methods is elaborated below along with which classes are used to train the models and which were used in testing. We do this because we have to train the autoencoder to learn one type of class so that when it tries to reconstruct images from the anomaly classes, it has a high reconstruction loss, which is then detected to be beyond the acceptable loss threshold and classified to possess the anomaly.

- **Retinal OCT Images** [7]
  - train - 26315 images (all "normal" images)
  - validation - 242 images (all "normal" images)
  - test - 726 images (different anomalies)
- **Brain Tumor MRI**[10]
  - train - 1595 images (all "notumor" images)
  - validation - 405 images (all "notumor" images)
  - test - 906 images (different anomalies)
- **COVID-19 Radiography**[1][11][12]
  - train - 2531 images (all "covid" images)
  - validation - 723 images (all "covid" images)
  - test - 3510 images (different anomalies and the normal images)

- **HAM10000**[15]
  - train - 6705 images (all "nevus" images)
  - validation - 909 images (all "nevus" images)
  - test - 568 images (different anomalies)

Models To experiment with unsupervised models, we use autoencoders and variational autoencoders. The architecture for these primarily remains the same across datasets, but there are a few differences with the HAM10000 dataset since we deal with colour images for that.

- **Architecture for Retinal OCT Scans, Brain Tumor MRI, and COVID-19 Radiography Datasets**
  For these datasets, the autoencoder is structured to have an input shape of (64x64x1). We use 32, 64, 128, and 256 filters in the first, second, third and fourth convolutional layers of the encoder respectively, using max pooling after each layer to reduce the spatial dimensions by half. We also use batch normalization, L2 regularization with a factor of 0.0001 and dropout with a rate of 0.4 to prevent overfitting and stabilize learning. The final output shape from the encoder is (4x4x256). The decoder of this autoencoder upsamples the spatial dimensions, doubling them each time, a convolutional layer between each upsample to reduce the number of feature maps per upsample until the output shape is once again (64x64x1).
  The variational autoencoder is quite similar to the autoencoder with a few minor changes. The dropout rate in this is set to 0.5 and the L2 regularisation factor is 0.01. The convolutional layers remain the same. The main difference lies at the end of the encoding process, where the output of the final convolutional layer is flattened and passed through a dense layer of 128 units, followed by a dropout layer with a parameter of 0.5. We also now have two separate dense layers to represent the mean and log variance of the latent space. Finally, a sampling layer takes care of gradient-based optimisation using the reparameterisation trick, which is the process of sampling a point in the latent space, based on the outputs of the mean and log variance dense layers. The decoder, which accepts latent vectors of size 64, then accepts this output of the encoder and reshapes it to a (4x4x256) shape. From here, the decoder behaves the same as that in the autoencoder and reproduces the (64x64x1) shape.
  Below is a list of the hyperparameters used for the autoencoder:
  - Batch size = 32
  - Learning Rate = 0.0005
  - Number of Epochs: 20
  - Optimizer = Adam
  - Target size = (64, 64)

  For the variational autoencoder:
  - Batch size = 32
  - latent dimension = 300
  - Learning Rate = 0.0001
  - Number of Samples = 15
  - Number of Epochs: 50
  - Optimizer = Adam
  - Target size = (64, 64)
  - beta = 0.01
- **Architecture for HAM10000 dataset**
  The architecture for the HAM10000 dataset for the autoencoder and the variational autoencoder is similar to that for the other datasets too. While the AE and VAE for the other datasets operate on grayscale images, the ones made for the HAM10000 dataset deal with colour images. This is because the images in this dataset hold intricate detail and even these small differences in details change the diagnosis completely. Hence, the last layer needs to allow for three filters in the output, for the red, green and blue data, rather than one, like the the models for the other datasets, where the images are grayscale. Apart from this, everything, including the hyperparameters, is the same.

It should be noted that for the AE, we make use of the Structural Similarity Index Measure(SSIM) loss to calculate the reconstruction loss and measure the threshold to recognise anomalies, whereas the VAE uses a the sum of the Kullback-Leibler loss and the reconstruction loss. The reconstruction loss is calculated as the mean of the sum of the mean squared errors of the initial and the reconstructed image.

Once the AE and VAE models are complete, the training process is quite similar. Run the models to train on the established training class so that it learns to reconstruct images of that specific class very well. The validation is also done with images of the same class. In the training phase, we tweak the various parameters so as to obtain a greater difference in the losses between the model reconstructs images for the class it was trained on and classes that are anomalous so that the threshold is easier to determine. Once they are trained, they can be tested on the anomaly classes individually. We also produce ROC curves to observe the learning of the AE and VAE for each of the classes and measure the AUC values for the classes.

It is also worth mentioning that the hyperparameters selected for this study were optimized through a systematic trial and error process. This was primarily because the models and datasets used in this research were quite complex to work with within the given time frame. Future research could attempt to experiment with and fine-tune these parameters further to attempt to get better results.

6 RESULTS

Throughout the experiment, some metrics need to be kept track of. Below is a list of the metrics calculated, along with the reason for choice.

6.0.1 *Accuracy.*

This metric helps understand how well the models detect anomalies and classify the medical images correctly.

6.0.2 *Precision.*

This metric allows us to determine how consistently the model can predict the classes of different anomalies.

6.0.3 *Recall.*

This metric helps evaluate whether the model is truly learning how the different classes are different from each other and how to detect them.

6.0.4 f1-Score. This metric explicitly shows the balance and trade-off between precision and recall for the supervised models, which are both important metrics for this research

6.0.5 AUC ROC curve. This metric also measures the models' ability to correctly classify different images but is more specific to each class in the dataset, rather than the "Accuracy" metric which gives a weighted average.

6.1 Performance

In this section, we will go over the final results obtained using the models after testing them with unseen data. Each of the four models is tested with unseen data from each dataset.

	Resnet50	**Densenet121**
OCT		
CNV	99%	100%
DME	100%	100%
DRUSEN	100%	99%
NORMAL	100%	100%
Brain MRI		
Glioma	92%	71%
Meningioma	91%	86%
Pituitary	99%	91%
notumor	88%	81%
COVID-19		
COVID	91%	92%
Lung_Opacity	89%	89%
Pneumonia	92%	95%
Normal	93%	92%
HAM10000		
AKIEC	51%	57%
BCC	51%	85%
BKL	58%	37%
DF	31%	0%
MEL	52%	65%
NV	88%	92%
VASC	70%	80%

Table 1. F1-Scores for Classes Across Datasets

Table 1 draws a comparison between the Resnet and Densenet models based on their F1 scores for each of the classes in each dataset. In terms of the OCT Retinal scans, we see that the scores are nearly a perfect 100%. This can be explained by the fact that the training dataset is extremely extensive. The model has so many images to train on that it is able to extract all important features extremely well and use that on the test set, a significantly smaller set of unseen data in comparison to the size of the training set. Another point to note about the OCT dataset would be the fact that each of the classes has images that are quite distinct from each other, unlike with the HAM10000 dataset, where the images can be easily confused(which is also one of the reasons for why the models perform significantly poorer on those images). Therefore, the models can learn the differences a lot better, thereby producing great results. For the Brain MRI dataset, Resnet50 seems to return better results compared to Densenet. However, it should be possible to achieve such results, if not better, with Densenet, considering its architecture allows for more intricate features to be detected. However, this would be for future research to try and experiment with. The COVID dataset has similar results for both models. Finally, the HAM dataset shows significantly varying results. Some classes are better detected by Resnet while others are better detected by Densenet. However, what is more useful to note here is that this is a prime example of where supervised methods start to falter. The other datasets contained a reasonably large amount of training data with extensive labelling. However, with the HAM10000 dataset, the imbalance in datasets was too high for the models to keep up. The AE and VAE were able to perform significantly better in this regard by detecting anomalies effectively.

Moving on to the unsupervised methods of anomaly detection, the way we collect and compare the AE and VAE is slightly different. In the supervised methods, we perform a multiclass classification which means that the AUC for the ROC curves may result in high values, but this would be because one of the classes is easier to predict, while other underrepresented classes are not. This is what we see in Table 1 for the HAM10000 dataset.

However, the AE and VAE are dealt with differently. Since each model is trained on one class which is considered the non-anomalous data, and then tested on a mix of just one anomalous class and the non-anomolous class, the AUC values are more reliable and determine how well the model is able to predict the different classes. Table 2 shows the resulting AUC values for the AE and VAE for each of the classes in the different datasets.

	VAE	**AE**
OCT		
CNV	0.98	0.99
DME	0.92	0.92
DRUSEN	0.80	0.74
NORMAL	-	-
Brain MRI		
Glioma	0.70	0.55
Meningioma	0.88	0.64
Pituitary	0.85	0.67
notumor	-	-
COVID-19		
COVID	-	-
Lung_Opacity	0.55	0.52
Pneumonia	0.92	0.86
Normal	0.72	0.64
HAM10000		
AKIEC	0.76	0.74
BCC	0.74	0.68
BKL	0.71	0.65
DF	0.68	0.51
MEL	0.75	0.66
NV	-	-
VASC	0.82	0.57

Table 2. AUC Scores for Classes Across Datasets

As seen in Table 2, the autoencoders are very suitable for the job regardless of the dataset we look at. There are some classes that they do struggle to detect, for example, the lung opacity class, where neither the AE nor the VAE was able to go beyond 0.52 and 0.55 respectively. Further fine-tuning in future research could help bump these numbers up too. The classes with "-" for values are the ones on which the model was trained and were taken as the non-anomaly classes when training the models. Observing the AUC values for these classes is not necessary since the research focuses more on if the models can detect the anomalies well.

So far, we have compared the supervised models among themselves, and likewise for the unsupervised models. To compare all the models with each other, we can also take a look at the accuracy of each model. Table 3 is an overview of the final accuracy that each model was able to achieve, for each dataset.

	Resnet50	**Densenet121**	**AE**	**VAE**
OCT	100%	93.75%	92.15%	93.11%
Brain MRI	92.44%	84.38%	70.02%	87.15%
COVID-19	93.10%	92.68%	70.48%	65.61%
HAM10000	83.42%	72.82%	70.18%	82.52%

Table 3. Accuracy of Different Models on Different Datasets

Overall, the supervised models seemingly exceed the unsupervised models in terms of accuracy, based on Table 3. However, upon further inspection, there are some other points to consider. The accuracy of the different models present in this table is simply an overall accuracy and does not give a full idea of whether or not the models perform well on the datasets. To truly make a judgement, we must look back on Tables 1 and 2. As stated before, the datasets selected portray a degree of imbalance in all cases. This means that even if the accuracy of the different models is high, this might just mean that the models have adapted extremely well to detecting classes in the dataset that might be the majority. This still leaves the underrepresented data difficult to predict for the model. This is precisely what we see with Resnet50 and Densenet121. For the HAM10000 dataset, while the models have been showing an accuracy of 83.42% and 72.82%, their F1-scores for the underrepresented classes seem to be too low for the model to be able to effectively predict those specific classes. The "NV" and "VASC" classes have high F1-scores since those classes have high representation and the models train extremely well to learn their features and distinguish them. But in terms of other classes like AKIEC, MEL or DF, both Resnet50 and Densenet121 struggle to produce good results.

On the other hand, the AE and VAE seem to do a better job at predicting the different classes of the HAM10000 dataset. The accuracy for the AE and VAE on this dataset are 70.18% and 82.52% respectively, and this time, these numbers do reflect the performance of the models well because the AUC for most classes of this dataset are also consistently high. Both models produce a more balanced output for AUC values for all classes, with the VAE performing exceedingly better than the AE. This is expected considering that the VAE learns and reconstructs images using features on a distribution rather than discrete values, allowing it to capture the finer details in images better than the regular AE.

This does make room for one more question; if the AE and VAE display such reliable results for the HAM10000 dataset, why do Resnet50 and Densenet121 seem to outshine these models when we look at the results with other datasets? Clearly, the accuracy and the F1 scores of all classes in the other datasets are very high, and this does show that Resnet50 and Densenet121 are more suitable for these datasets. What must be considered here are the datasets rather than the models. The HAM10000 dataset is a case of a dataset with an extreme imbalance in the representation of classes. The remaining datasets, however, are not as extremely imbalanced. This makes it still possible to classify the different images in these datasets using the supervised models. While more datasets like HAM10000 would have been ideal to test the AE and VAE further, finding such datasets is also very difficult, given the extreme financial and labour costs of labelling data for the use of training supervised models. Therefore, unsupervised models would be the next best reliable option for the use of detection of diseases in medical imaging. In the case of datasets like OCT retinal scans, Brain Tumor MRI scans or COVID-19 Radiography images, the supervised models show extremely reliable results. This also directly ties to the fact that supervised models like Resnet50 and Densenet121 are designed to extract important, intricate features present in medical images. This allows the models to learn the data significantly better than the AE or VAE, which cater more towards reconstruction of images, and rely more on poor reconstruction of images to reliably detect anomalies. In other words, the AE and VAE do not extract as much information as Resnet50 or Densenet121 would from the given data, since they are not designed for such a task. However, with other datasets that may be as extremely imbalanced as HAM10000, if not more, the AE and VAE would be more likely to have the upper hand at producing useful results, in comparison to Resnet50 or Densenet121.

To further expand on why the AE and VAE might be a better choice for imbalanced datasets, we can also look at the time taken for the models to train. The table below gives an overview of the average times taken per epoch to train the different models on different datasets.

	Resnet50	**Densenet121**	**AE**	**VAE**
OCT	510	290	60	25
Brain MRI	40	20	4	4
COVID-19	100	42	4	3
HAM10000	62	35	16	4

Table 4. Time per Epoch of Different Models(in seconds)

As seen in Table 4, the AE and VAE in general display significantly better performance in terms of time taken to train. Combining this with the fact that the AE and VAE tend to produce better results with imbalanced datasets, the unsupervised models do seem like the wise choice for a task involving a dataset with extreme imbalances.

7 FUTURE WORK

Looking back on the goals of this research, there is plenty of room for further development, which could be reached with an increase in resources like time, number of researchers and more. Collaborative efforts from various researchers would be beneficial to the expansion of this research to deepen our views on the different machine learning models used to detect diseases.

7.0.1 Unsupervised Models. So far, there is ample evidence that shows that autoencoders and variational autoencoders might be a better choice when it comes to extremely imbalanced data, especially considering the costs that go into obtaining and labelling data for supervised models. We looked at the AUC values of different classes for the unsupervised models, and along with their accuracy, were able to conclude that they are an effective means of detecting anomalies in medical images. However, there are many varieties of autoencoders. For example, masked(MAE) and sparse autoencoders(SAE) could perform just as well, if not better than the VAE, and performing further experiments with these models would deepen our insight on what models are more suited for the different types of datasets.

7.0.2 Supervised Models. At the same time, we also looked at the accuracy of supervised models along with the F1 scores on the different classes. Once again, other supervised models like Deep-CNNs or other pre-trained models can also be tested to explore the possibilities and check if perhaps there might just be a model that produces fruitful results.

7.0.3 Enhancing Current Performance. Finally, the models that were explored in this research could still be further fine-tuned for use on the datasets used, to check if there may be ways of obtaining better results and making sure no class has low F1-scores or AUC values.

8 CONCLUSION

In this study, we attempt to enhance our understanding of the performances of different models, both supervised and unsupervised, in the field of the detection of diseases in medical images. We explore two supervised models - Resnet50 and Densenet121 - and two unsupervised models - autoencoders and variational autoencoders. The former two models are pre-trained models on which we apply transfer learning and fine-tuning to produce results that cater to the requirements of this study. The latter two models, are trained on one non-anomaly class and then used to detect the presence of different diseases in the anomalous images.

By doing this, we see that supervised models may perform effectively with a certain level of imbalance in datasets and even produce results better than an AE or VAE can since these supervised models can extract features from images better, considering they are designed for this purpose. However, after a certain threshold, when the imbalance is too extreme, AEs and VAEs seem to surpass the performance of the supervised models, not just by how well they predict anomalies, but also by the short period under which they achieve the task.

This helps us answer **SQ1**, which asks how well the supervised and unsupervised models perform in terms of the various metrics available since we now see that all models seem to produce results with high accuracy. The supervised models have a higher accuracy than the unsupervised models since they are meant to extract features from images. However, their effectiveness reduces with a greater imbalance in the datasets, as observed in Table 1, where the F1-scores are very low for HAM10000. On the other hand, even though the accuracy of the AE and VAE are slightly lower, they are still effective enough to detect anomalies and would be able to operate at this level consistently, regardless of the dataset.

To answer **SQ2**, which refers to how effectively we can use transfer learning and fine-tuning to produce better results, we experimented with the hyperparameters of the various models and arrived at the best possible results. The different metrics recorded in this paper are based on these models. Future research in this field can attempt to improve the models by tweaking the results further.

The research aims to highlight the need for the use of anomaly detection and multiclass classification for the diagnoses of rare diseases, since they become increasingly arduous to detect, the rarer they get. It portrays the effectiveness of such models and shows the potential of AI in the field of medicine as a means to detect health concerns quicker before extensive treatment is required, as well as to make sure that diagnoses are accurate.

This study lays a foundation for many future studies to come. With the collective action of multiple researchers on this topic, as well as deeper investigation of the models used and new models alike, we can pave a clearer path to further our understanding of the most effective means of machine learning and potentially revolutionise the methods by which we detect and treat rare diseases.

9 REFERENCES

[1] M.E.H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M.A. Kadir, Z.B. Mahbub, K.R. Islam, M.S. Khan, A. Iqbal, N. Al-Emadi, M.B.I. Reaz, and M.T. Islam. 2020. Can AI help in screening Viral and COVID-19 pneumonia? *IEEE Access* 8 (2020), 132665–132676. https://doi.org/10.1109/ACCESS.2020.3010287

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[3] Fatima Zohra Hamlili, Mohammed Beladgham, Mustapha Khelifi, and Ahmed Bouida. 2022. Transfer Learning with Resnet-50 for Detecting COVID-19 in Chest X-Ray Images. *Indonesian Journal of Electrical Engineering and Computer Science* 25, 3 (2022), 1458–1468. https://doi.org/10.11591/ijeecs.v25.i3.pp1458-1468

[4] Hanafi, Andri Pranolo, and Yingchi Mao. 2021. CAE-COVIDX: Automatic COVID-19 Disease Detection Based on X-Ray Images Using Enhanced Deep Convolutional and Autoencoder. *International Journal of Advances in Intelligent Informatics* 7, 1 (2021), 49–62. https://doi.org/10.26555/ijain.v7i1.577

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[7] Daniel Kermany, Kang Zhang, and Michael Goldbaum. 2018. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. https://doi.org/10.17632/rscbjbr9sj.2

[8] Yuchen Lu and Peng Xu. 2018. Anomaly Detection for Skin Disease Images Using Variational Autoencoder. *arXiv:1807.01349v2 [cs.LG]* (2018).

[9] Mayank Mishra. 2023. Convolutional Neural Networks Explained. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939. Accessed: 2023-06-19.

[10] Msoud Nickparvar. 2021. Brain Tumor MRI Dataset. https://doi.org/10.34740/KAGGLE/DSV/2645886

[11] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S.B.A. Kashem, M.T. Islam, S.A. Maadeed, S.M. Zughaier, M.S. Khan, and M.E. Chowdhury. 2020. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images. *arXiv preprint arXiv:2012.02238* (2020).

[12] Tawsifur Rahman, Ali Khandakar, Yazan Qiblawey, Al Tahir, Serkan Kiranyaz, Saima Kashem, Muhammad E. H. Chowdhury Islam, Somaya Al Maadeed, Susu M. Zughaier, and Muhammad Salman Khan. 2020. COVID-19 Radiography Database. Kaggle. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/data

[13] Pablo Ruiz. 2023. Understanding and Visualizing DenseNets. https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a. Accessed: 2023-06-19.

[14] Tanish Sharma. 2023. Detailed Explanation of Residual Network (ResNet50) CNN Model. https://medium.com/@sharma.tanish096/detailed-explanation-of-residual-network-resnet50-cnn-model-106e0ab9fa9e. Accessed: 2023-06-19.

[15] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* 5 (2018), 180161. https://doi.org/10.1038/sdata.2018.161

[16] Nechirvan Asaad Zebari, Ahmed A. H. Alkurdi, Ridwan B. Marqas, and Merdin Shamal Salih. 2023. Enhancing Brain Tumor Classification with Data Augmentation and DenseNet121. *Academic Journal of Nawroz University* 12, 4 (2023). https://doi.org/10.25007/ajnu.v12n4a1985

A USE OF AI TOOLS

During the preparation of this work, the author(s) used Grammarly to check for grammatical errors and formatting of the content of the LaTeX file. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.