# Deep Neural methods for medical imaging: An experimental study on COVID-19

**KTH - Kungliga Tekniska Högskolan**

**Antonio Frederico Nesti Lopes**
afnl@kth.se

**Matheus Oliveira Franca**
maof@kth.se

**Svenja Raether**
raether@kth.se

## Abstract

The dramatically fast evolution of COVID-19 has triggered scientists around the world to conceive fast effective responses in favor of a better understanding of this novel disease. Radiography examination is an available, accessible, and portable alternative screening method that enables rapid triaging. In this sense, deep neural nets can provide recognition and classification tasks to aid radiologist experts and improve Chest X-ray (CXR) images analyses for faster medical treatment. **Wang's (2020)** work deploys a customized network, a.k.a. COVID-Net, conceived from an automatic network search (AutoML) consisting of a deep-CNN composed of several PEPX modules. This work aims to reproduce **Wang's (2020)** network in PyTorch without accounting for pre-training on the ImageNet dataset. Furthermore, the challenge of dealing with small data and the duty to provide reasonable accuracy for medical purposes drives the investigation of deep neural methods to go on various strands. We provide a performance comparison with ResNet50 as we account for possible extensions and improvements through generative models. Further additions, such as calibration, GradCAM algorithm, and an interface for simplified usability were added on the top of results analysis. The code for this project is available at https://github.com/FredericoNesti/DD2424_COVID19_Project.
**Keywords:** COVID-19, COVID-Net, ResNet, CNNs, GradCAM, calibration, generative networks.

## 1 Introduction

The ongoing global crisis due to the spread of COVID-19 severily impacted hospitals capacity and logistics, as well as medical treatments. This problem raised the need for an effective response from the scientific community, with many ongoing open-source initiatives. Fast track works such as **Wang (2020)** [10] and **Fan (2020)** [2] aim to support medical researches by providing a way for identifying COVID-19 cases from X-ray images. This work aims to reproduce **Wang's (2020)** results, currently available as a TensorFlow implementation, by using PyTorch. Furthermore, topics such as data balancing, calibration, and explainability are covered and an interface for simplified usage is provided.

An additional reference for the comparison of the results is given in **Ilias Papastratis**[1] reproduction of COVID-Net and its comparison with ResNet50.

This report splits into the following sections: Section 2. Related Work, presents the literature that supports the methods implemented in this paper and states other relevant developments. Section 3. Data specifies the data source, as well as features and configurations of the data; Section 4. Methods,

[1] github.com/IliasPap/COVIDNet

covers the approaches undertaken for handling small data through training and testing, efficient metrics to analyze the results, and the Grad-CAM algorithm for auditing model explainability. It is followed by section 5. Experiments, containing results and corresponding discussions and, finally, section 6. Conclusion, where a summary of the key results is given and possible extensions mentioned.

## 2 Related Work

COVID-Net was designed by **Wang (2020)** [10]. It proposes a projection-expansion-projection design pattern for balancing computational efficiency and representational capacity. COVID-Net was created explicitly to detect X-Ray images of corona patients with particularly high sensitivity and to differentiate them from conventional lung diseases.

**Ilias Papastratis** provides additional analysis of Wang's work by making comparisons between COVID-Net and other well-established models such as Resnet50 and Resnet18. His work reveals more details about the architecture and parameters of COVID-Net and contributes a first realization of the model in PyTorch. However, with regard to comparable metrics and explanations of possible differences, this reveals gaps that need to be filled in order to make the two approaches more comparable.

Previous studies, however, didn't cover the calibration for their models. **Volodymyr**[4] and **Laves**[5] have already studied this aspect for medical analysis with other models. This is an important aspect in order to progress in the direction of real cases scenarios, where the model should indicated a coherent level on confidence on its prediction. Our work aims not only to replicate the previous results, but also add this analysis for assisting the progression towards a final product to be used by doctors.

**Tarroni (2018)** [1] provides another insight when it comes to handling medical datasets. His work proposes a generative model by using a CNN-VAE model to learn the latent representation of Cardiovascular images, where the learned mean representation is treated as an input to a MLP to perform classification tasks. His work gives support to the idea of using a VAE network so as to learn the latent representation of COVID-19 CXR images so as to come up with a more powerful technique for data augmentation.

## 3 Data

All datasets used were collected from public databases used by Linda Wang.[2] The publicity of CXR images regarding COVID-19 is ongoing. Moreover, Wang's report on COVID-Net test are due on to 4 different versions from March to May. In this sense, the given implementations and stated results in this report are made based on the Covidx version of the available data. The decision for the Covidx dataset is based on the availability of comparable key figures and metrics at the beginning of this project.

The CXR dataset is composed by 3 classes: normal, pneumonia-related diseases and COVID-19. Particularly, the dataset used integralizes 358 CXR images from 266 patients, 8066 normal CXR images and 5538 non-COVID-19 pneumonia. A procedure for splitting the dataset is provided in Wang's GitHub repository, even though it was not mentioned on the original article. [3] This was adopted in our preprocessing task for constituting the training and testing datasets and to assign their labels. In this work we do not account for a validation set, given the small size of data used. Regarding the small amount of COVID-19 image samples, we have used canonical data augmentation techniques combined with data balancing inside each batch, so as to favoritize COVID-19 samples in face of pneumonia and normal X-ray images. Furthermore, a cropping filter is applied in order to standardize the input size for the network. This goes along with the data augmentation technique described in the following section.

[2]The used databases are avaiable on www.kaggle.com/c/rsna-pneumonia-detection-challenge, github.com/ieee8023/covid-chestxray-dataset, nihcc.app.box.com/v/ChestXray-NIHCC and github.com/agchung/Figure1-COVID-chestxray-dataset.

[3]github.com/lindawangg/COVID-Net

# 4 Methods

## 4.1 Model architecture

Wang's COVID-Net is a customized CNN architecture. Its network specification consists of PEPX modules, flattening layers, fully connected layers and the activation layer. Furthermore, the connections follows a very particular structure with some forward and parallel feeds, where image representations combines itself in various ways.

The PEPX module stands for the abbreviation of projection-expansion-projection-extension. It works in the following way: a first-stage projection use a 1 x 1 convolution for projecting input features in a lower dimension; then, these features are expanded to a different higher dimension with 1x1 convolutions; after, spatial characteristics are learned fro depth-wise 3x3 convolutions; then there is another projection stage with 1x1 convolutions; and a final 1x1 convolutions channel extension.

Particularly, the depth-wise representation aims to minimize computational complexity and preserve representativeness. The key difference between the expansion and extension stages is that the first aims to expand features to a higher dimension diverse from the input, whereas the extension applies exclusively for the number of channels.

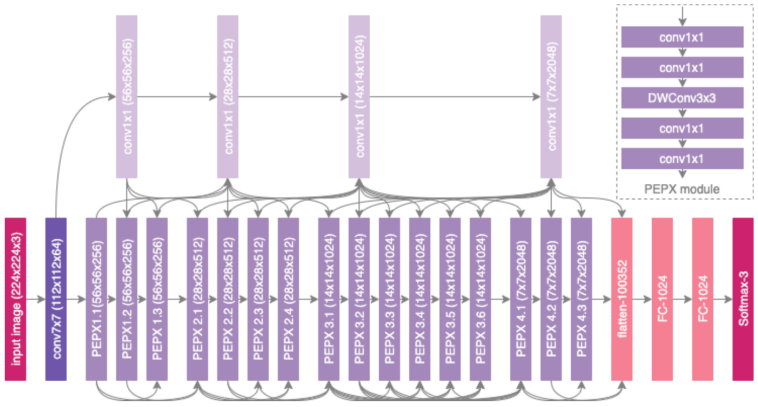


Figure 1: COVID-Net Architecture. High architectural diversity and selective long-range connectivity can be observed as it is tailored for COVID-19 case detection from CXR images. The heavy use of a projection-expansion-projection design pattern in the COVID-Net architecture can also be observed, which strikes a strong balance between computational efficiency and representational capacity. This image was extracted from **Wang (2020)**.

Likewise **Wang (2020)**, we provide a comparison of the COVID-Net with the Resnet50 that does not leverage PEPX modules and a selective long-range connectivity.

Some challenges regarding the reproduction of the COVID-Net architecture were the representation of the connections on Figure 1 that posed some questions regarding the correctness of those connections. Moreover, the expansion scheme inside each PEPX module also triggered some doubts given the lack of specification. To overcome these issues, we relied in the reported number of parameters on Wang's github and Papastratis' github implementation, given the chosen architecture version. Our network has 133.4M trainable parameters, whereas Wang's COVID-Net smaller version has 117.6M and the big version 134M.

## 4.2 Implementation details

Our implementation of COVID-Net was directly trained on COVIDx dataset (Section 3) using ADAM optimizer using a learning rate policy where the learning rate decreases when the sensitivity arives in a plateau region. If this is the case, after a pacience of 5, the learning rate is reduced by a factor of 0.7. Cross Entropy was used for the loss function. The following hyperparameters were used for training: learning rate=2e-5, number of epochs=16, batch size=8. Moreover, data augmentation techniques were used in order to improve performance: translation, rotation, horizontal flip, random crop and intensity shift. Finally, we implemented the balancing dataset and weight loss as mentioned in sections 4.4 and 4.5.

### 4.3 Data augmentation

The augmentation technique consisted of cropping the image to 224x224 size (along with center and random cropping), applying random color jitter to change the intensity of brightness, rotation, horizontal flip and translation. Those were applied using the torchvision package methods and aim to mimic the augmentation techniques applied by Linda Wang.

### 4.4 Balancing the dataset

In respect to sample proportion balancing, we have employed a configuration of 30% of the images of the batch to be from COVID19 and the remaining share containing a sample from pneumonia and normal labels. Training of the model was executed with a batch of 8 pictures. This setup mimics the approach chosen in the training of COVID-Net by Linda Wang.

### 4.5 Weighted loss

To increase the sensitivity of the model for prediction of the COVID-19, the underrepresented class in the given dataset, weights are applied to the loss values. Those emphasize learning to detect COVID-19 cases by multiplying those cases with a factor of 6 while treating normal and pneumonia with a factor of 1.

### 4.6 Analytics metrics

The indicators used to evaluate the model are Accuracy, Sensitivity, and Positive predictive value. Accuracy describes the percentage of correctly classified images. Sensitivity measures the proportion of actual positives that are correctly identified as such. The PPV describes the proportion of true positives out of all positives. Those measures are particularly relevant for this project since it operates on an unbalanced database and the accuracy of the model concerning individual classes must be checked individually. Besides, it aims to detect positive corona cases and should, therefore, be of high value especially in this class.

### 4.7 Auditing via model explainability

We extend our models (both COVID-Net and Resnet) to incorporate a method for explainability. This enables to show which parts of the image most influenced the predicted results. Similarly, **Wang (2020)** employs the GSInquire [6] method to highlight critical areas in lung pictures. In our approach, we make use of the GradCAM technique, supported by **Selvaraju et al. (2016)** [9].

This technique is readily applied to a range of CNN-based models on torchvision package, including Resnet50.[4] In this case, adaptions had to be made to fit our implementation. COVID-Net is not one of torchvisions standard models and is not supported by a package the way it exists for ResNet. Therefore, the GradCAM method was implemented for our COVID-Net model from scratch. The reason for this is that the activations, that are relevant for constructing the underlying heatmap have to be taken directly from the corresponding model .[5]

### 4.8 Calibration

Confidence calibration – the problem of predicting probability estimates representative of the true correctness likelihood – is important for classification models in many applications. In the study of medical diagnosis, the confidence of the model is essential for a proper examination. For this reason, once the models were trained, we applied temperature scaling – a single-parameter variant of Platt Scaling [8] – that was shown to be effective at calibrating predictions [3].

This method uses a single scalar parameter $T > 0$ for all classes. The new confidence prediction is:

$$\hat{q}_i = \max_k \sigma_{SM} \left( \frac{z_i}{T} \right)^{(k)} \tag{1}$$

[4] pypi.org/project/pytorch-gradcam/

[5] medium.com/@stepanulyanin/implementing-grad-cam-in-pytorch-ea0937c31e82

Where $z_i$ is the logit vector, $k$ is the number of classes and T is called the temperature. This modification "softens" the softmax. T is optimized with respect to Negative Log-Likelihood (NLL) on the validation set. Because the parameter T does not change the maximum of the softmax function, the class prediction $\hat{y}_i$ remains unchanged. In other words, temperature scaling does not affect the model's accuracy.

### 4.9 Extension trial with VAE

The uneasiness brought by the small amount of data puts under questioning whether a generative approach could support a better training of COVID-Net by increasing the representativeness of COVID-19 CXR images on the dataset. Fortunately, **Tarroni's (2018)** work supports the usage of VAE network for medical imaging. Our insight was to provide a probabilistic approach for data augmentation through a Resnet-VAE[6]. On the contrary of the proposed by this author, we focused on a simpler task consisting in using separate networks for classification and latent representation learning.

Briefly, the VAE is a encoder-decoder network that accounts for a parametric inner representation. Typically, the VAE seeks to optimize the Kullback-Leibler divergence between a true distribution of an image an its encoded-decoded representation. It can be shown that this corresponds on maximizing the ELBO [7], which could be broken into two parts: the MSE corresponding to the gaussian distribution of the reconstruction at the decoder side and the Kullback-Leibler divergence between a standard gaussian and the distribution of the latent representation from the encoder side. Images could then be reconstructed by propagating the mean value of the encoded input through the decoder.

## 5 Experiments

In order to evaluate the performance of the model and better understand its behavior during prediction, we perform both quantitative and qualitative analysis.

### 5.1 Quantitative analysis

To quantitatively investigate the performance of our implementation, we used the aforementioned COVIDx dataset. We started by analysing the complexity of each model in terms of accuracy and number of trainable parameters (see Table 1). We can notice that different implementations of COVID-Net have different number of trainable parameters, which reinforce that the paper did not clearly specified the architecture of the model. Our implementation got 86.4 accuracy, which is lower than suggested by Linda Wang. On the other hand, ResNet50 presented a very similar result, 91.3, even with less then one fifth of trainable parameters of COVID-Net.

To further interpret the results, we evaluated the confusion matrix (Fig 2), sentitivity for each class (Table 2) and Positive Predictive Value (PPV) (Table 3). The first point that stands out is the 0.00 sentisivity for the COVID-Net implementation done by Ilias Papastratis. This version does not achieve satisfying results on sensitivity and PPV for the COVID-19 class. Those deviations can be explained by the underlying dataset. Since this is unbalanced, with a dominating number of Normal and Non-COVID-19 cases, the model tends to predict those over the COVID-19 examples resulting in a sensitivity value of 0 and a PPV of 0 for the COVID-19 class.

Since this model was not trained using the rebalanced dataset nor the weighting for loss, it completely ignores the COVID-19 cases, because of the few examples, on the data. This leads to a good overall accuracy, but 0 sensitivity for this disease. Our implementation got almost 84% sensitivity to COVID-19, a better result than ResNet50, 74,2%. This reinforce the idea of a tailored model build to identify COVID-19.

With regards to PPV, COVID-Net with our rebalancing shown the best performance for "normal" class with PPV of 93.0%, but Linda's Wang implementation was superior for non-COVID19 and COVID-19 desiases with a PPV of 94.7% and 96.4% respectavly.

Once the model was trained, we applied temperature scaling in order to calibrate the predictions of our COVID-Net (see Figure 3). The optimal temperature found for the model was 1.954, which made

[6] github.com/LukeDitria/CNN-VAE

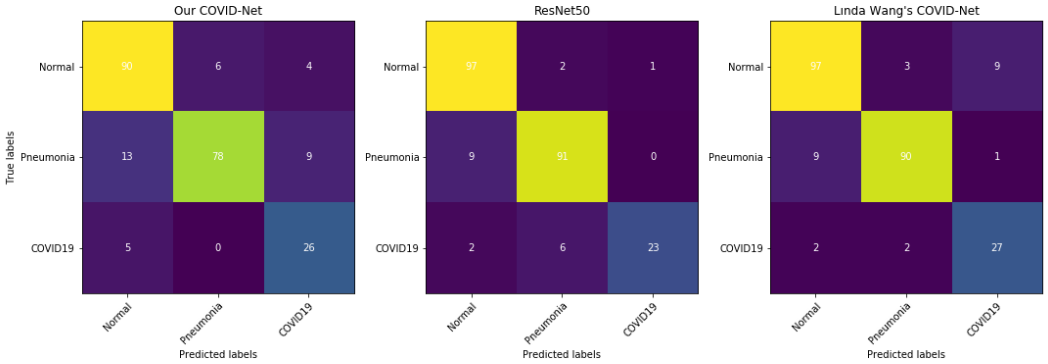


Figure 2: Confusion matrix for our COVID-Net, ResNet50 and Linda's Wang COVID-net on the COVIDx test dataset.

| Model | Params (M) | Accuracy |
|---|---|---|
| Our COVID-Net | **134.7** | 86.4 |
| ResNet50 | 23.2 | 91.3 |
| COVID-Net without rebalancing | 118.19 | 91.22 |
| Linda Wang's COVID-Net | 117.4 | **92.6** |

Table 1: Performance of COVID-Net on COVIDx test dataset

| **Sensitivity (%)** | | | |
|---|---|---|---|
| **Model** | **Normal** | **Non-COVID19** | **COVID-19** |
| Our COVID-Net | 90.0 | 78.0 | **83.9** |
| ResNet50 | **97.0** | **91.1** | 74.2 |
| COVID-Net without rebalancing | 92.9 | 85.9 | **0.00** |
| Linda Wang's COVID-Net | **97.0** | 90.0 | **87.1** |

Table 2: Sensitivity for each infection type.

| **Positive Predictive Value (%)** | | | |
|---|---|---|---|
| **Model** | **Normal** | **Non-COVID19** | **COVID-19** |
| Our COVID-Net | 90.0 | 78.0 | 83.9 |
| ResNet50 | **91.9** | 89.8 | 95.8 |
| COVID-Net without rebalancing | **93.0** | 84.2 | **0.00** |
| Linda's Wang COVID-Net | 89.8 | **94.7** | **96.4** |

Table 3: Positive predictive value (PPV) for each infection type.

the NLL drop from 0.694 to 0.620. We can notice that the calibrated model (orange line) is closer to the line completely calibrated line (blue curve) compared to the original version (green line). This outcome consolidates the idea of Chuan Guo [3] that temperature scaling is an effective at calibrating predictions.

In addition, we can notice that the original model was underconfident when predicting normal cases (Fig 3 left), since the green line was always bellow the blue curve. The temperature scaling ameliorate this behavior, as we can see the orange line neighboring the blue line.

### 5.2 Qualitative analysis

As mentioned in Section 4.7, we applied the GradCAM algorithm in order to gain insights into which critical factors of the image were used for classification. This is an important aspect of medical analysis in order to enhance transparency to verify that the model does not make decisions based on biased or irrelevant visual indicators. The critical factors identified by GradCAM of normal, pneumonia and COVID-19 cases are shown in Fig. 4.

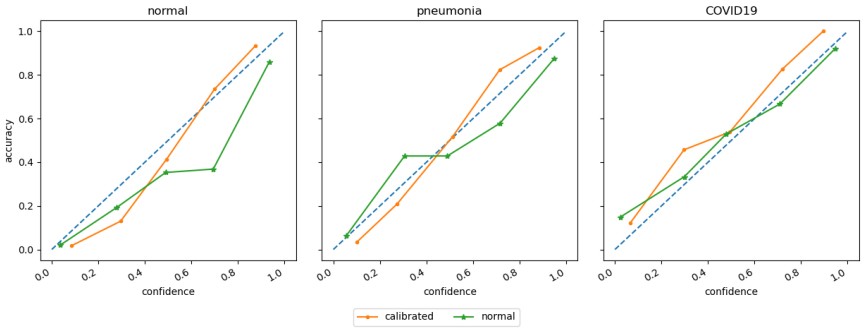


Figure 3: Reliability diagrams for the three classes with regards to predictions of a trained COVID-Net before and after temperature scaling.

We can notice that the heatmap highlight a region of the chest that COVID-Net used for its classification. It is a reasonable region to focus, since it points out a region of the chest and not some biased section of the x-ray.

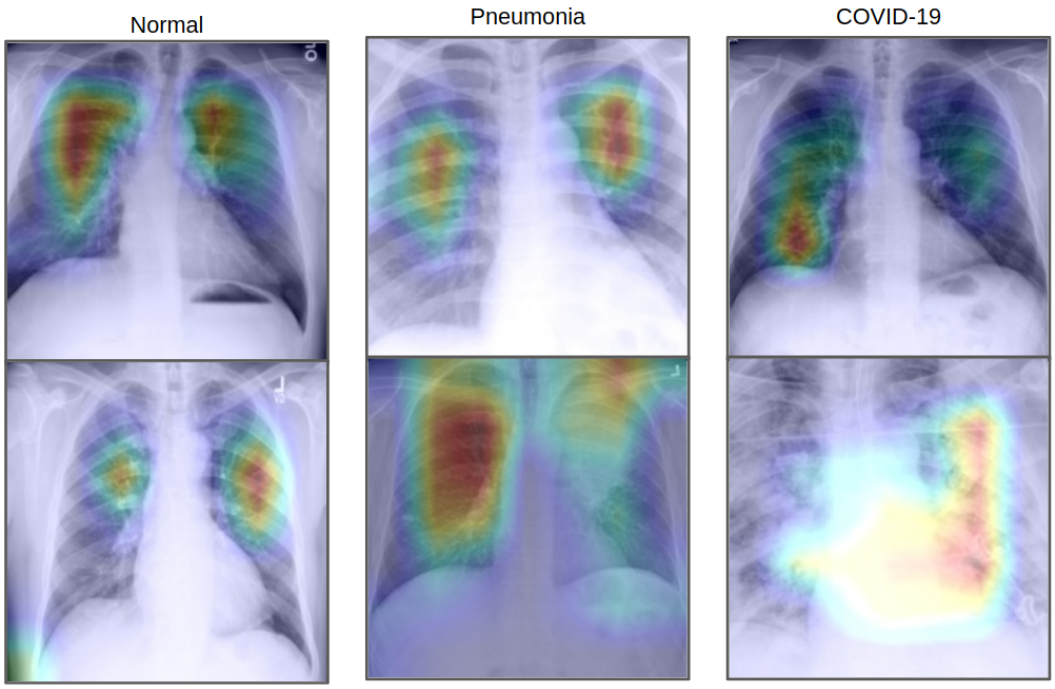


Figure 4: Result of CXR images of normal, pneumonia and COVID-19 cases with GradCAM[9] highlighting the their associated critical factors (displayed in red).

### 5.3 Extension: Generative Network

The Resnet-VAE application did not yield satisfactory results for image generation of COVID-19 latent distribution. We explored several configurations for the latent dimension (from 16 to 4000) as well as 2 types of losses: the standard loss that comes from the VAE-ELBO formulation and the InfoMaxVAE-loss[7]. The network topology to be used is something that deserves further investigation, since it was slightly modified so as to handle with 224x224 image dimension. Moreover, even though we had access to better networks designs investigated in this report, we chose to begin with a smaller

[7] github.com/AliLotfi92/InfoMaxVAE

architecture so as to provide faster code adaptation and results. In general, higher latent representation dimension yielded better results.

# 6 Conclusion

## 6.1 Summary of the key results

Our implementation provides a replicated implementation of Linda Wang's work using PyTorch. Metrics, covering accuracy, sensitivity, and PPV are captured for each network to provide a transparent comparison of the model's performance. The comparison of results between our implementation and Wangs COVID-Net shows a slight difference which might be explained by the lack of pretraining in our version. ResNet50, pre-trained on ImageNet, shows similar results compared to Wang's COVID-Net, yet uses less than 1/5 of the amount parameters. The analysis of similar approaches, i.e. the GitHub provided by Ilias Papastratis, underlines the need for data balancing for the specific task of predicting COVID-19 given an imbalanced dataset. Targeting the imbalanced dataset and taking measures towards sensitivity for COVID-19 pictures improves the sensitivity of this underrepresented class. Our implementation further applies the Grad-Cam algorithm - an alternative method to leverage explainability - to highlight relevant sections in the pictures. The application of temperature scaling calibrates the COVID-Net serves as an additional improvement. Furthermore, a platform for simplified usage of the model was added to the project.

## 6.2 Future Work

The work of **Wang (2020), Ilias Papastratis and Tarroni (2018)** openly indicates that there are still several configurations to be explored so as to leverage medical imaging analyses. Particularly, **Tarroni (2018)** approach for classifying while learning the latent representation is a promisable application for COVID-19 cases as well, even more under a adversarial fashion. Moreover, the continuous update of COVID-19 can shape the searching for specific architectures as the models are continuously tested. This could offer a better evaluation on the capacity of generative approaches in a stand-alone VAE fashion or in a combination of classification and generation tasks altogether.

# 7 Acknowledgements

We thank the initiative of Linda Wang and Ilias Papastratis on providing quick results. We also thank the opportunity gave by the DD2424 course of KTH to work on such a relevant theme.

# Appendix A: COVID-Net Platform

In addition to the analysis, we developed a web platform with Flask micro web framework (Fig 5). The goal was to show a use case for real scenarios where a doctor can import a new CXR image and the platform uses our version of COVID-Net to classify it among normal, pneumonia and COVID-19. Furthermore, it gives the predicted probability for each class and also uses Grad-CAM to indicate critical factors.

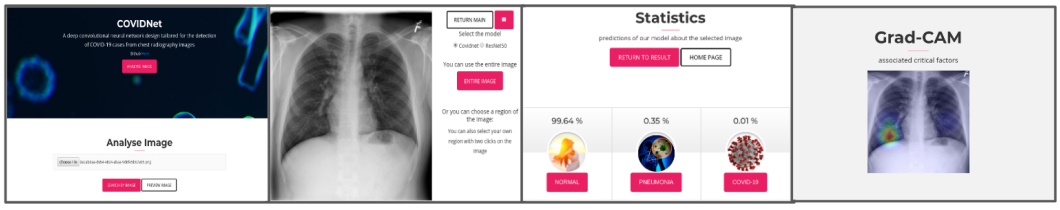


Figure 5: Use case of the COVID-Net platform created in order to facilitate the work of doctors to use COVID-Net as a part of the diagnoses.

## References


[1] Giacomo Tarroni Wenjia Bai Antonio De Marvao Georgia Dormou Martin Rajchl Reem Bedair Sanjay Prasad Stuart Cook Declan O'Regan Carlo Biffi, Ozan Oktay and Daniel Rueckert. Learning interpretable anatomical features through deep generative models: application to cardiac remodeling. 07 2018.

[2] D.-P. Fan, Tianjiang Zhou, G.-P. Ji, Youxi Zhou, Gwen guo Chen, H. Fu, Jia ji Shen, and Lanchun Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct scans. *ArXiv*, abs/2004.14133, 2020.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.

[4] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *ICML*, 2018.

[5] Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, 2020.

[6] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St. Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *ArXiv*, abs/1910.07387, 2019.

[7] Stephen G. Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function. *Department of Machine learning Research, RETINA-AI Health, Inc.*, 07 2019.

[8] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

[9] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[10] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. 2020.