POLITECNICO
MILANO 1863

PROJECT

FOR THE COURSE IN

APPLIED AI IN BIOMEDICINE

Federico Carrara

A.A.2021/2022

1 INTRODUCTION

Cardiovascular diseases (CVDs) are the top cause of mortality worldwide (World Health Organization, 2017). According to the World Health Organization (WHO), approximately 17.7 million people died from CVDs in 2015. During this project, we focused on a particular group of CVDs, namely the arrhythmias. Arrhythmias consist in electrical disorders of the heart which alters its normal rhythm and rate. The diagnosis of arrhythmia is usually based on the analysis of electrocardiogram (ECG) recordings. Indeed, by inspecting the shape of individual heartbeats and their position with respect to contiguous heartbeats, it is possible to detect abnormalities and malfunctions of this electrical system. The visual and manual interpretation of ECG performed by clinicians is a time-demanding and expensive task. However, thanks to the latest discovery in the field of Artificial Intelligence (AI), many researchers have started to develop computer-aided diagnosis (CAD) systems to automatically analyze ECG (Martis, et al., 2014). Therefore, introducing such tools into clinical settings as an adjunct tool to aid the cardiologists would certainly reduce the patient waiting time, lessen the workload of cardiologists and reduce the cost of ECG signal processing in the hospitals, without affecting the quality of the diagnostic process

The objective of this project is to provide a computer-aided method to classify ECG signals of heartbeats in the following 3 classes (Fig.1):

- Normal Sinus Rhythm (NSR),
- Premature Atrial Complex (PAC),
- Premature Ventricular Complex (PVC).

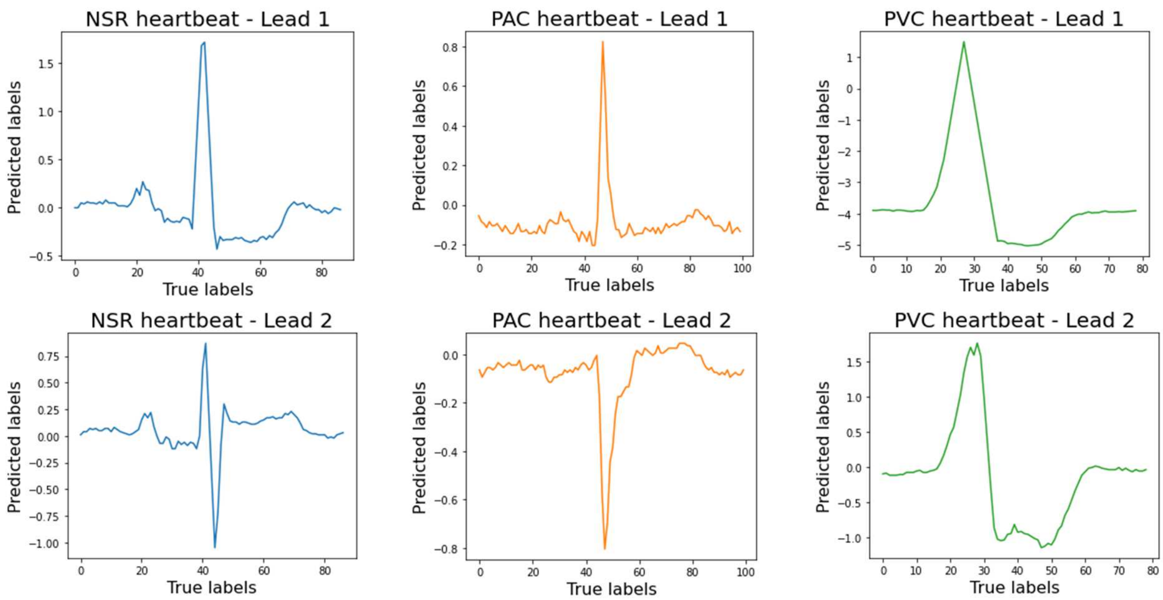

Figure 1: Examples of ECG heartbeat signals divided by arrhythmia class and lead

With this goal in mind, after a careful preprocessing of the raw ECG signals, a Convolutional Neural Network (CNN) classifier is employed to solve the heartbeats classification task. The choice of a deep learning paradigm is because, differently from statistical or machine learning methods, the feature selection phase is implicitly performed inside the network. This is a great advantage, as extracting meaningful features from the raw ECG signals is a complex task that requires deep knowledge in the biomedical field.

2 MATERIALS AND METHODS

2.1 ECG DATASET

The ECG heartbeat signals are extracted from a database consisting of 30-minutes long, II-leads ECG recordings from 105 subjects. Some of these signals are sampled at 128Hz, while others at 250Hz. Moreover, two lists containing the positions of R-peaks and the labels corresponding to each heartbeat are provided for each subject. From this dataset a total number of 244336 ECG heartbeat signals are extracted.

2.2 DATA PREPROCESSING

The extraction of ECG heartbeat signals from the original dataset is performed by computing the midpoints of the R-R peak intervals and splitting the sequences of beats in these points. In this way, the extent of each heartbeat, which varies with the current heart rate of the subject, together with its peculiar morphology is preserved. The presence of heartbeats whose extent is beyond extreme physiological limits (i.e., less than 40 BPMs) is due to faulty imputation of the R-peak positions. Therefore, such signals are excluded from the analysis.

The result of this extraction approach is a list of heartbeat signals of different length, also due to the different sampling frequency. Hence, resampling based on Fast Fourier Transform (FFT) is applied to map each heartbeat to the final length of 128 sampling points, without modifying the morphology. Then, a bandpass filter is applied to each signal between 0.5 and 47 Hz to remove motion artifacts and minimizing possible effects of noise on model classification (Fig.2). Finally, Z-score standardization is performed on each heartbeat signal.

In addition to ECG data, some features concerning the relative position of each heartbeat with respect to the others are extracted. This approach is particularly useful to distinguish NSR from PAC beats. Indeed, from a morphological point of view, the difference between the heartbeats of these classes is not always remarkable (Fig.3). However, PAC heartbeats tend to be anticipated, resulting in a higher heart rate. One approach to capture this behavior is to measure the intervals between R-peaks in sequences of heartbeats (Sannino & De Pietro, 2018). Therefore, the following features are extracted:

- *Pre-RR interval*, defined as the RR-interval between a given heartbeat and the previous heartbeat,
- *Post-RR interval*, defined as the RR-interval between a given heartbeat and the following heartbeat,
- *Local average RR interval*, defined as the average of the RR-intervals within a sliding window covering the past 10 s,
- *Global average RR interval*, defined as the average of the RR-intervals within a sliding window covering the previous 5 min.

We refer to these quantities as RR-Features (RRFs). To guarantee that all the RRFs are properly defined and given the data abundance, we drop the first 5 minutes of each subject's ECG recording. In order to assess the information gain brought by considering these additional features, in the following we compare the performances of classification algorithms using and not using RRFs.

After preprocessing, the final number of II-lead ECG heartbeat signals is 199225.

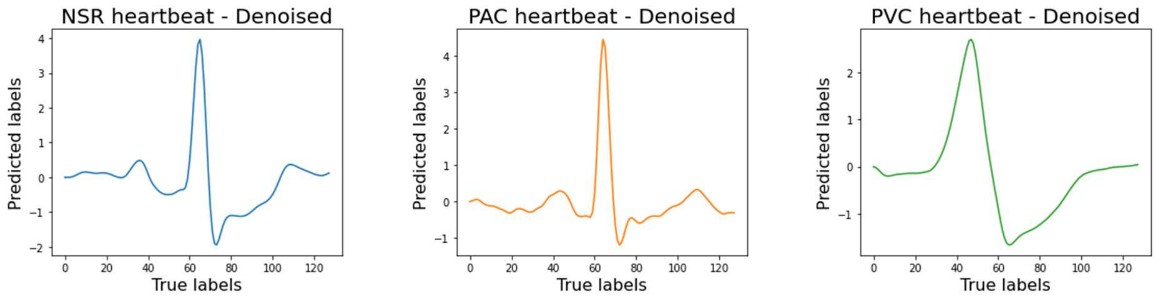

Figure 2: Example of ECG heartbeat signals after FFT resampling and bandpass denoising.

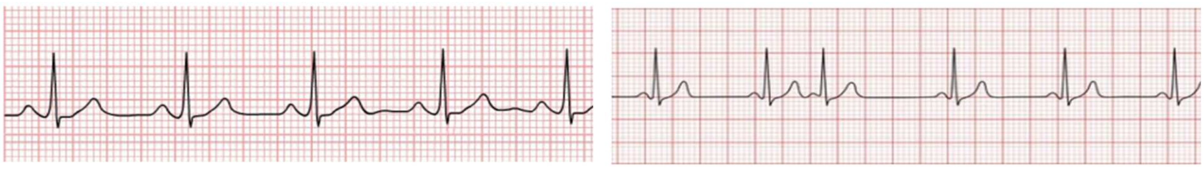

Figure 3: Comparison of a Normal Sinus Rhythm (left) and Premature Atrial Complex rhythm (right).

2.3 CLASS IMBALANCE PROBLEM

Success of ECG analysis based on deep learning mainly relies on the richness and diversity of the data. Although large amount of ECG recordings is available in our dataset, the incidence of abnormal cardiac events is much lower than that of normal beats, resulting in a highly imbalanced dataset. In Fig.4 the numerosity of heartbeats belonging to NSR, PAC and PVC classes is reported. We can observe that normal heartbeats are more than 20 times more frequent than abnormal ones. The class imbalance problem needs to be carefully treated when training the classification algorithm, otherwise it can lead to highly biased results. In this project we try to solve this problem using a twofold approach:

- On the one hand, we consider the original dataset, and we apply *Weighting.* This approach consists of computing class weights which are inversely proportional to the class frequencies and then using these weights during training for increasing the importance of samples belonging to the less represented classes.
- On the other hand, we perform *Data Augmentation* (DA) on the original data to increase the numerosity of the samples in the minority classes as to match the numerosity of the most represented class (Pan, et al., 2020). In particular, we follow the approach of (Acharya, 2017). Namely, the samples of synthetic data are generated by adding a random noise on the standard deviation and mean values used for Z-score calculation on the original ECG heartbeat signals. In addition, we also perform cropping of the external segments of the heartbeats with subsequent

resampling to 128 data points. These operations combined allow to modify the amplitude and time scale of the signals, without changing their physiological structure and consistency.

An important aspect to notice is that when *Data Augmentation* is employed to train the network, better results in terms of classification are obtained by performing *Test Time Data Augmentation* (TTDA). This procedure is applied differently when evaluating the model on the validation and on the test set. Indeed, since in the first case labels are available, we apply the same DA schema as for the training set. On the contrary, since for the training set labels are not available, we cannot know a-priori which samples actually belong to the minority classes and, thus, need oversampling. Therefore, we must modify the DA scheme such that from each test sample a fixed number of synthetic instances are generated. Then, for each fold of synthetic samples relative to the same parent signal, we first predict a label for each sample and, finally, we attribute the class given by majority voting.

In the following we present and compare the results obtained by using *Weighting* and *Data Augmentation* approaches.

2.4 Classification Algorithm: Architecture

Deep learning algorithms are a powerful tool for analyzing non-tabular data, such as images, texts or signals, since they can directly process these types of unstructured data. ECG recordings are a particular case of non-tabular data that can be successfully analyzed using an Artificial Neural Network (ANN). In this work we decided to employ a peculiar kind of ANN, which is the Convolutional Neural Network (CNN). The advantage of using a CNN is that, thanks to the convolution operation, it keeps into account the spatial structure of the input signal, namely the relative position of the signal's data points. Moreover, it allows to analyze the different leads of the input signal at once, mapping the content of the different channels in higher dimensional latent representations.

For this project, we developed a CNN based on the *Res-Net* architecture [He] with 1-D convolutions. The network is composed by 8 *identity blocks* (Fig.5). At the beginning of the network and between every couple of stacked identity blocks there is an additional block provided with one convolution followed by a *batch normalization* layer. Finally, the last identity block is connected to a *global max poling* layer that brings the flattened hidden volume to the *softmax* classifier. This deep architecture contains almost 11 million of trainable weights. Although this number is considerably high, the number of available instances to train the network is enough to avoid overfitting.

When considering also the features relative to the R-R peak intervals, a 2-layer *Feed-Forward Neural Network* is added to the architecture. In particular, the FFNN receives as input the 4-elements long feature vector, and produces as output its latent representation, which is concatenated to the output of the *global max poling* layer of the CNN. We refer to this architecture as *Wide Res-Net*.

2.5 Classification Algorithm: Training

Due to the dimensions of the dataset and the computational requirements imposed by the CNN, we use a hold-out approach to train and evaluate the classifier performance. Therefore, the original dataset is split in training, validation and test sets using stratification, and with, respectively, 65%, 20% and 15% proportions.

Then, in order to solve the class imbalance problem, we either feed the CNN with the original training and validation sets and the class weights computed on the labels of the training set, or the augmented training and validation sets. Notice that data augmentation is performed separately on the two sets. In order to get comparable results, in both cases we use the same training schedule. Namely, training is performed using the Adam optimizer with an initial learning rate of 0.01 on batches of 64 signals for a total of 40 epochs. Moreover, to avoid overfitting an early stopping mechanism is implemented. Finally, to enhance the optimization, a learning rate scheduler is triggered when validation loss does not decrease for some consecutive epochs.

For what concerns *Wide Res-Net*, the training schedule is extended with 2 additional steps. Indeed, after the initial training of the whole network, the convolutional part of the network gets frozen for some epochs, during which only the FFNN is fine-tuned. Then, we do the opposite, by freezing the FF part and fine-tuning the rest. This multistep training schema is implemented to avoid that the convolutional part, which brings the most information to the output, completely covers the content brought by the FFNN.

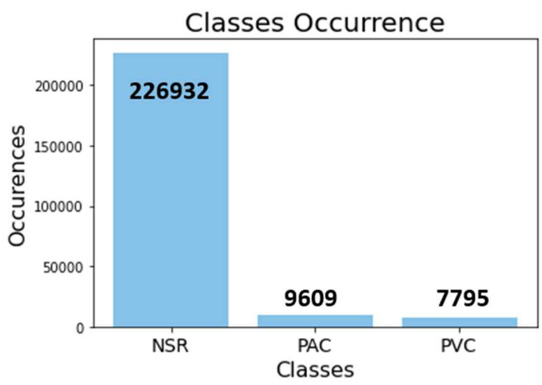

Figure 4: Heartbeats numerosity for each class.

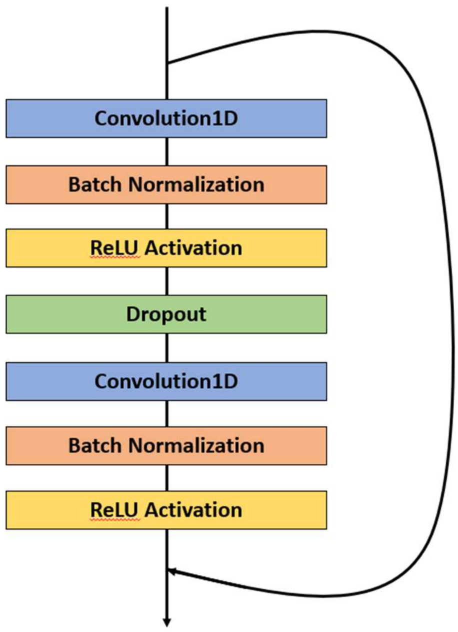

Figure 5: Layer structure of the Identity block employed in Res-Net architecture for 1D signals classification.

3 RESULTS

In this chapter we report and discuss the results of classification of ECG heartbeat signals. The outcomes are evaluated using the *accuracy, precision, recall* metrics and the *confusion matrix*. that *recall,* which is defined as the number of instances which the model correctly identified as *relevant* out of the total relevant instances (Wood, 2021), is particularly important for this kind of classification. Indeed, when dealing with ECG classification the most desirable characteristic for an algorithm is being able to find the most of the arrythmias. An exhaustive summary of metrics for all the algorithms and datasets considered is reported in Tab.1.

We start presenting the classification performance of the CNN trained on the original dataset with *weighting*. In this case, we compare the results using the base *Res-Net* and the *Wide Res-Net.* The related confusion matrices and metrics for the 2 cases are reported, respectively, in Fig.6 and in the two upper block of Tab.1. As we can see, the *Wide Res-Net* is better at classifying the PAC heartbeats. This is probably due to the fact that PAC heartbeats are usually matched with an alteration of the heart rate, which is successfully captured by the RRFs associated to those heartbeat. In addition, *Wide Res-Net* has a slightly higher *recall* both on the classification of the single classes and globally. Therefore, we conclude that *Wide Res-Net* is a more powerful algorithm overall.

The second comparison we present is about the methods employed to deal with the class imbalance problem. On the one hand we use *weighting*, to give more importance to the samples of the less represented classes, while on the other hand we resort to a *Data Augmentation* technique. In both cases we use the *Wide Res-Net* architecture. Thus, the results for what concerns results of *weighting* one can look at the previous paragraph. On the contrary, the confusion matrices for DA are in Fig.7 and the metric in the two lower blocks of Tab.1. It is worth to mention that for what concerns the *Data Augmentation* approach two different evaluations are reported. In the first case we consider the test set as if the labels were known, hence we can use the standard DA technique. On the opposite in the second case, we consider labels to be unknown, exactly as in a real scenario, and we perform the *test time* DA approach described in the previous paragraphs. This twofold evaluation helps to have a deeper understanding of the results. With a first glimpse at the confusion matrices, we immediately realize that while the results of the first case are extremely positive, the one for the other case shows a terrible accuracy for instances of NSR class. This preliminary impression is confirmed by the poor performance metrics for the *test time* DA approach. Anyway, there are few exceptions, as, for instance, the recall for PAC and PVC classes. An explanation of this strange behavior can be found in overfitting. Notice that in this case overfitting does not concern the value of data themselves. Indeed, if this was the case, we would have had bad performances also for the case of test set with known labels. Therefore, in this case overfitting is more about the structure of the datasets obtained after applying DA. This means that during training the model learns with data that are extremely balanced in terms of class distributions. This implies that also at test time the model expects to be fed with a dataset of the same structure. Hence, a model trained in this way is not able to classify correctly a dataset with highly imbalanced design, such as the one provided by test time DA. Therefore, we can conclude that:

- on the one hand, by looking at the results on the augmented test set with labels, DA technique is a powerful and promising technique, since it allows to obtain extremely consistent results.
- On the other hand, it is evident that there is still some more research to be done in order to properly design a DA methodology which is effective in real scenarios of ECG signals processing.

Scenario	Class	Accuracy	Precision	Recall
Weighting & Res-Net	NSR	-	0.9967	0.9838
	PAC	-	0.7103	0.9120
	PVC	-	0.9128	0.9474
	Global	0.9799	0.8732	0.9477
Weighting & Wide Res-Net	NSR	-	0.9978	0.9829
	PAC	-	0.7039	0.9323
	PVC	-	0.9116	0.9558
	Global	0.9802	0.8711	0.9570
Data Augmentation with Known labels & Wide Res-Net	NSR	-	0.9891	0.9919
	PAC	-	0.9629	0.9544
	PVC	-	0.9596	0.9655
	Global	0.9708	0.9706	0.9706
Test Time Data Augmentation & Wide Res-Net	NSR	-	0.9997	0.3851
	PAC	-	0.0879	0.9666
	PVC	-	0.1376	0.9674
	Global	0.4257	0.4084	0.7730

Table 1: Classification metrics for the different CNN architectures and methodologies to deal with class imbalance.

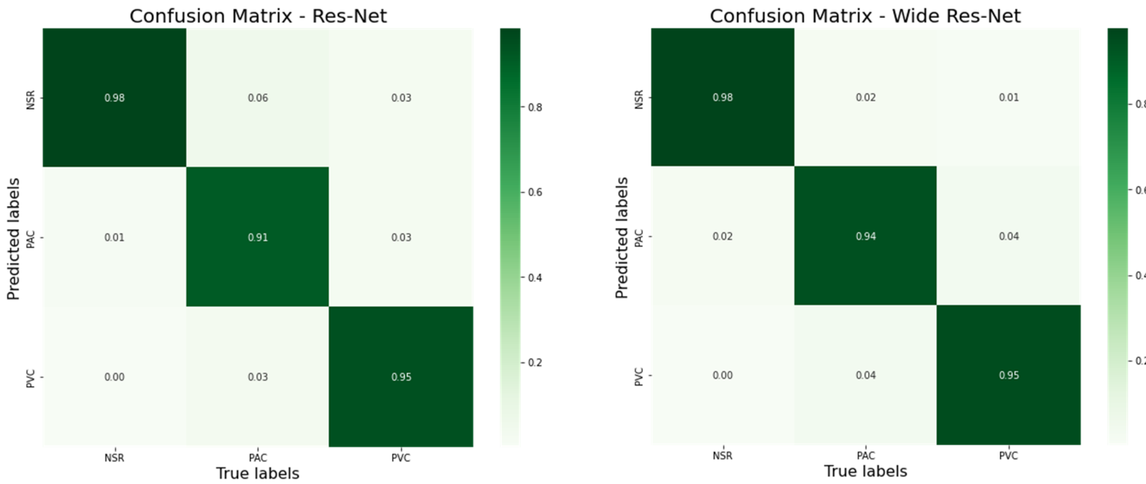

Figure 6: Confusion matrices for the CNN architecture comparison.

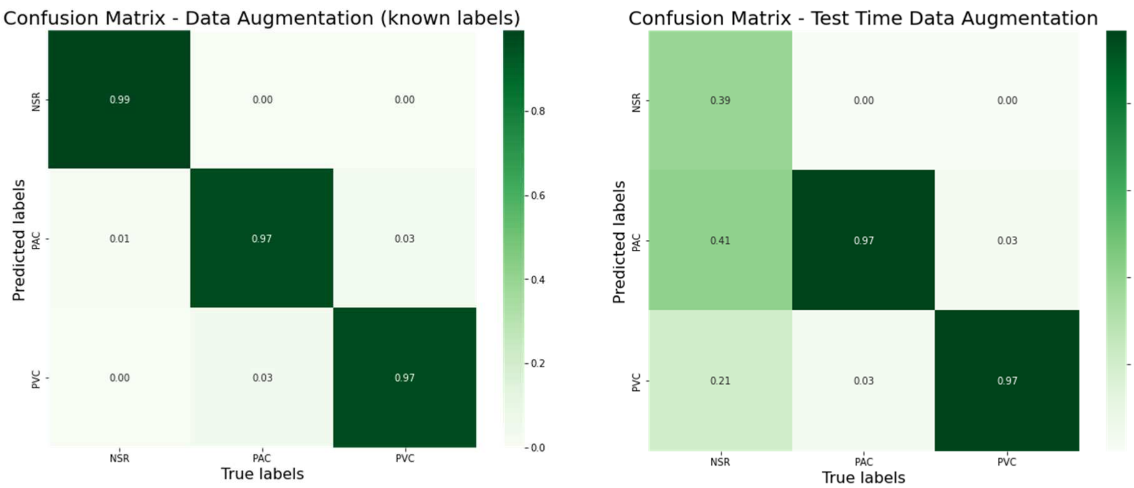

Figure 6: Confusion matrices obtained after training with Data Augmentation.

4 CONCLUSION

In this study an ECG signals processing pipeline and a deep learning approach are presented to automatically identify and classify the different types of ECG heartbeats, which are crucial for diagnosis of cardiac arrhythmia. The CNN that we developed can automatically classify 3 different ECG heartbeat classes, NSR, PAC and PVC. Thus, it can be implemented into a CAD ECG system to perform a quick and reliable diagnosis in clinical settings. The best performing CNN architecture is the *Wide Res-Net*, which combines in a single algorithm the ECG heartbeat signals and some numerical features regarding the R-R peak intervals. The problem of class imbalance is extremely evident in heartbeats classification, as the number of normal heartbeats is way greater than the number of abnormal ones. In order to overcome this problem, the most effective way, at the current state of the art, is to apply a *weighting* approach during model training. Anyway, the approach of *data augmentation* is an undoubtedly a very promising technique for dealing with class imbalance problem. However, the application of this approach in the field of signals analysis is still not supported by enough high-level studies and, thus, to assess the consistency and to determine best practices for this approach some proper validation studies are needed.

5 REFERENCES

Acharya, U. R., 2017. A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine,* Volume 89, pp. 389-396.

Martis, Acharya & Adeli, 2014. Current methods in electrocardiogram characterization. *Computers in Biology and Medicine,* Volume 48, pp. 133-146.

Pan, Q., Li, X. & Fang, L., 2020. Data Augmentation for Deep Learning-Based ECG Analysis. In: *Feature Engineering and Computational Intelligence.* s.l.:Springer, pp. 91-100.

Sannino, G. & De Pietro, G., 2018. A deep learning approach for ECG-based heartbeat classification for. *Future Generation Computer Systems.*

Wood, T., 2021. *Deep AI.* [Online]
Available at: https://deepai.org/machine-learning-glossary-and-terms/precision-and-recall

World Health Organization, 2017. *Cardiovascular Diseases (CVDs),* s.l.: s.n.