Deep Learning for Medical Time Series Classification applied to ECG and Lung Transplantation

Final Project Report

Delanoue Pierre, Floyrac Aymeric, and Gueneau Clément

ENS Paris-Saclay, MVA Master*

Abstract. In this work we develop from scratch neural networks models applied to the classification of cardiac problems from electrocardiograms (ECG) and to decision making in lung transplantation. The innovation of our work is that we have adapted methods that have proven to be successful in other applications such as medical imaging. We are implementing, for the first time to our knowledge, a Bayesian approach to neural networks applied to time series classification. We show that this Bayesian approach is relevant given the needs of our medical setting. All our codes can be found on the associated GitHub repository.

Keywords: Deep Learning · Time Series Classification · ECG · Lung Transplant.

1 Introduction

Time series are ubiquitous in the medical context. Efficient and relevant machine learning methods for classification of time series would allow an innumerable number of applications in pathologies detection, decision making or creation of vital indicators. For all these reasons, the study of time series classification in health is the subject of intense research.

However, there are many aspects that make this task complex. First of all, a time series is a high dimensional object which complicates all the tasks of the algorithms. For example, depending on the case studied, only a small variation can be found to be the discriminating element of a disease. As a result we are dealing with performance, execution time, and memory problems. Secondly, the acquisition of vast and reliable databases has a high cost. In addition to the current nonexistence of a uniform format (or even unit of measurement) between the different hospitals, the labelling of these databases can often only be done by few experts. Finally, the interpretability and transparency of results is essential in medicine. New forms of information need to be invented for time series

* Course of Deep learning for medical imaging taught by O.COLLIOT and M.VAKALOPOULOU

to allow us to understand the results of our classification algorithms. All these difficulties and issues remind us of many difficulties found in medical imaging (see A in Appendix).

The goal of this project is to develop deep learning architectures to perform different time series classification tasks in a medical context.

After a brief presentation of the state of the art, we first present the models that we have tested, then we present methods for the interpretability of these models, after that we present our dataset and problems choices as well as our results. Finally we discuss these results.

2 Related Work

Concerning deep learning applied to time series classification, Fawaz et al. 2019 [4] is a great starting point. We use this extensive review of models to select our deep learning architecture baselines. As they well explain, Deep Learning has surprisingly been considered for time series classification only by a few papers. Important approaches include the paper of Wang et al. [19], Zheng et al. [20] and the really recent paper of Tang et al. [18] to create strong baselines.

Multiple time series classification papers applied to medicine can be listed. The huge majority of these papers do not use deep learning. Deep learning has only really recently started to be investigated: In particular for ECG ([17], [6]), surgical skills evaluation ([12], [8]), early detection of Parkinsons disease ([15], [5]). As far as we know, no study has yet investigated Bayesian neural networks for time series classification.

3 Methodology

In this section we present the deep learning models as well as the interpretation methods we have chosen to study.

3.1 Models

We decided to focus our research on 3 types of models:

Multi Layer Perceptron Even being the simplest architecture of deep learning, MLP shown good performances for time series classification (see Wang et al. [19]). Note that MLP looses the temporal information of the time series, each time stamp having its own weight.

Convolutional Neural Networks CNN for time series classification has been studied by Wang et al. [19] from which we took some of the architectural choices. We decided to innovate by using Stochastic Weights Averaging (SWA), introduced by Izmailov et al. [9]. SWA has been shown to significantly improve results in computer vision tasks. In a nutshell, the idea is to average the weights of the neural network at different points in the learning process, using a cyclic learning rate scheduler.

ResNet ResNet is a convolutional network which implements shortcuts across layers, performing an *identity mapping*. This idea was first introduced in 2016 in [7] to tackle the vanishing/exploding gradients issue in deep networks. We use this architecture adapted to time series, ie by replacing 2D convolutional layers by 1D convolutional layers, to determine if it improves the results.

3.2 Interpretation

In medicine more than in many fields, it is essential to understand the results provided by our algorithms. So when making decisions, we are going to avoid using a *black box* model. A criticism often made to deep learning models is that it is difficult to understand which element has been decisive in the model's decision making. Moreover, we have no indicators of confidence in the prediction that has been provided with neural networks.

We present here two approaches that we have put into practice to alleviate these problems. The first approach takes advantages of the CNN structure and the second approach consists in using a Bayesian framework.

Class activation maps Our first visualization tool is heavily inspired from imaging methods. A classical explanatory tool for CNNs is the so called "class activation map": given an image, we can retrieve which zones induced which level of activation and build a heatmap from these activation levels. We can proceed similarly with our CNN, except that they are 1D-CNN instead of 2D-CNN as in imaging. The process consists in a forward pass, retrieving the activations and averaging them along filters of the considered layer. The most important layer is generally the last one.

The idea is thus to highlight some parts of the input time series. It can be particularly relevant in the context of ECG: since it is a very structured signal, we expect the anomaly detectors to highlight some discriminative patterns of the signal. Taking a step back, it can enable physicians to check the CNN prediction.

Bayesian Neural Networks The Bayesian approach to machine learning is a hot topic of research. The main reason for this interest is the fact that for similar performances, the Bayesian framework allows us to get more information about

the confidence of our model on its prediction. This is mainly due to the fact that we may have access to an a posteriori variance on our parameters or even on our prediction.

A Bayesian approach to neural networks would combine the predictive power of deep learning with Bayesian interpretive tools. An exciting paper by Lee et al. [10] shows how to do Bayesian inference with a Multi Layer Perceptron. In a nutshell, Lee et al. prove that an MLP with infinite width behaves like a Gaussian Process. The deep fully-connected neural network with an i.i.d. prior over its parameters then has an equivalent Gaussian process in the limit of infinite network width. They call this limit a Neural Network Gaussian Process (NNGP). Then, Lee et al. give a procedure to build the NNGP equivalent to any MLP architecture. This is exactly what we do to our MLP model.

The main benefit of this method is to explicitly obtain the kernel of the Gaussian Process. When fitted, our model will give us a variance on our predictive mean. This enables us to construct confidence intervals which gives us a strong and easy way to interpret the trustworthiness of a prediction.

4 Evaluation

We are now studying the performance of these models. First we present the datasets and the problems we will use. After presenting our working tools, we will analyze our results. Finally we study the relevance of our interpretation tools.

4.1 Datasets

ECG5000 This dataset contains 5000 electrocardiograms recording labeled in 5 classes. Divided into a training set and a test set, this dataset has an acceptable size for the application of deep learning. ECG5000 is heavily unbalanced: two of the five classes represent 92.7% of the total dataset (see Table 1). This is a classical dataset used in time series classification since Bagnall et al. [1] (see more in appendix C.1).

Class	1	2	3	4	5
Proportion (%)	58.4	35.3	1.9	3.9	0.5

Table 1. Distribution of classes in the ECG5000 database

MIT-BIH The MIT-BIH arrhythmia database is a large database of electrocardiograms, initially composed of half-hours long recordings. It has been made

public in 1980 in order to develop automated arrhythmia detectors. We use here a cleaned version of the database with isolated heartbeats, so that we have 110 000 samples.
The database is highly unbalanced, which is standard in the medical context. There are 5 different classes, which distribution is given in table 2 (see more in appendix C.2).

Class	Normal	1	2	3	4
Proportion (%)	82.8	2.5	6.6	0.7	7.3

Table 2. Distribution of classes in the MIT-BIH database

Lung Transplant We use a dataset created by the Foch Hospital (Suresnes, France) and DataForGood France on lung transplantation [3]. The aim of the dataset is to optimize the decision making process at the end of an operation for patients who have received a lung transplant. The data set contains, among other things, the evolution of vital indicators (heart rate, expired oxygen concentration, etc.). The target is binary and indicates whether the patient should be extubated (1) or not (0). The data set contains 330 observations. Each time series is 1082 in length. The distribution of classes is 42% of 1 (need to be extubated) 58% of 0. See more details about this dataset and lung transplantation in the appendix C.3.

4.2 Setup

Our neural networks have been implemented using the Tensorflow framework with add-ons for recent techniques such as Stochastic Weight Averaging, and the keract library to retrieve activations. To construct our Bayesian Neural Network we use the recently released *Neural Tangents* [13] package developed by Google.

4.3 Results

To be able to assess the performance of our models, we compare our results with those of a K Nearest Neighbors model (KNN) using Dynamic Time Wrapping (DTW) as distance (see B.1 in Appendix for more details). We run our models on the introduced datasets and we obtain the following accuracy in Table 1 (see Additional Tables for the Recall 9, Precision 10 and F1-Score 11 in appendix D).

The results are mixed. For the transplant dataset , the results are globally weak. The neural networks are unable to capture the information without doubt because of the small amount of data. For ECG5000, the best network is the 1D-CNN with SWA. This simple and very scalable network performs slightly better

than ResNet. 1D-CNN and ResNet are ahead of KNN by capturing almost perfectly the behavior of the two largest classes while getting good results on the minority classes. However, let's notice that KNN performs slightly better on the minority classes of ECG5000 compared to neural networks. On the other hand, MLP fails to capture information from the minority classes and simply predicts the majority classes well on this dataset. Concerning MIT-BIH, ResNet and 1D-CNN with SWA show very close performance (about 99% of accuracy), and capture information on every class, even the tiniest one. The MLP totally misses a class, which is an unforgiving weakness in a medical context, and is outperformed by the convolutional architectures. This confirms that convolutional layers are able to detect some kind of structure in the signal, which the MLP cannot. KNN also show weaker results than the convolutional architectures. However, it was trained on a subset of the whole dataset for computational reasons: it cannot handle more than a few thousands samples. For the same reason, NNGP could not be trained on the whole dataset.

Accuracy					
	KNN	MLP	1D CNN	ResNet	NNGP
Transplant	0.48	0.58	0.58	0.58	0.60
ECG5000	0.95	0.92	0.96	0.95	0.95
MIT-BIH	0.95	0.93	0.99	0.99	0.97

Fig. 1. Accuracy of each dataset for each model.

4.4 Visualisation

Activation Maps Figure 2 shows the class activation maps of the 10 last layers of a ResNet trained on the MIT-BIH dataset. It looks like, along layers, the model examined some meaningful patterns of the signal, like the QRS complex.

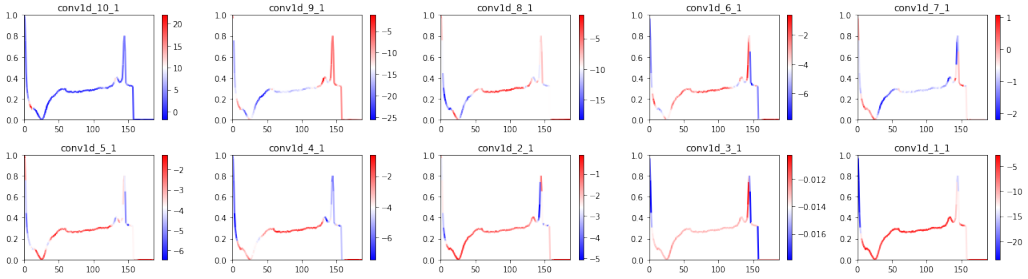

Fig. 2. Class activation maps for the ResNet

Bayesian Inference We intended to exploit the NNGP Bayesian framework to derive information from model predictions. For this we use the predictive mean (equivalent of the MLP output) and its variance (obtained from the Kernel computation). The performances of the NNGP on Transplant being bad, we concentrate our observations on the two ECG datasets. We want first of all to verify the relevance of the information given to us by the variance. We would like the variance to be small when the model gives a good prediction and to be large when the model is wrong. By observing the confusion matrix 3 and 13 (see in Annex E) we can see that we obtain what we want.

We can now imagine the use that would be made of this variance in a concrete case. For example with the MIT-BIH dataset we propose to display a figure such as Figure 4 on a case-by-case basis. On this exemple where the model gave a wrong prediction, the high variance obtained, as well as the derived confidence interval, encourages us to reconsider the conclusions of the model.

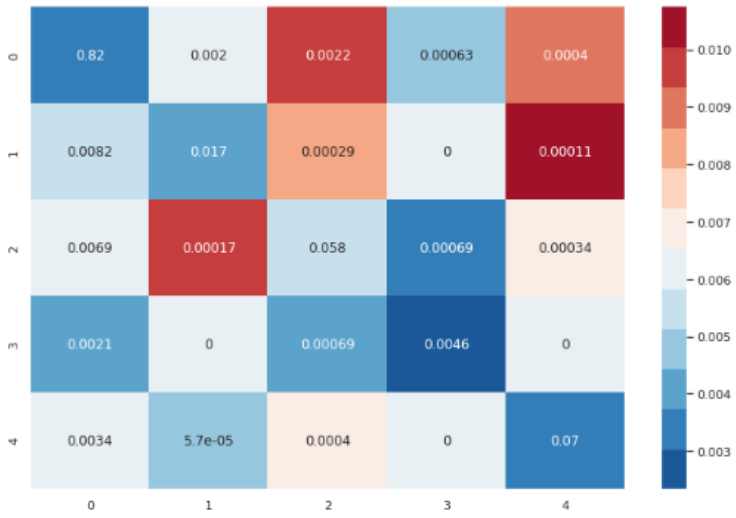

Fig. 3. Confusion matrix (displayed number) for the NNGP with the mean predictive variance (color) on the MIT-BIH dataset: Each cell color corresponds to the mean of the predictive variances according to the predicted label (x axis) and true label (y axis). The number display is the one of the confusion matrix of the NNGP model. The posterior variance seems lower on the diagonal (correct predictions). High variances (red) correspond to hesitation of the model which is why having high variances on errors is good.

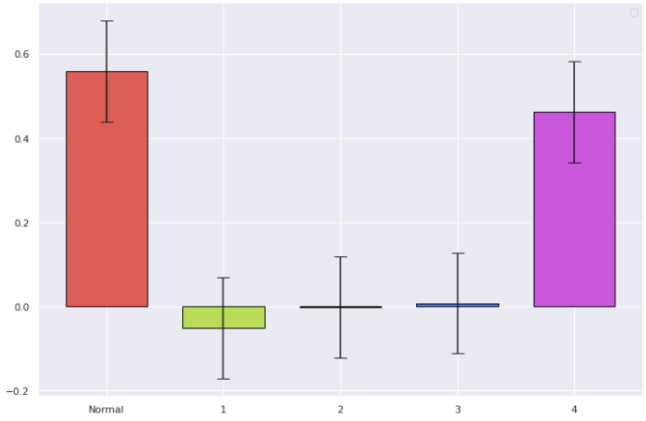

Fig. 4. Predictive mean of the NNGP with its 95% confidence interval for an element of the test of the MIT-BIH dataset. In this example, the time series was wrongly classified (Predicted label: Normal, True label: 4). We observe that the variance of the predictive mean is high for this prediction. The model hesitated with 4. In a real situation, this might require a doctor's close look at pathology 4 for this patient.

5 Discussion

Although not exhaustive, our study already highlights the many difficulties encountered when performing time series classification in a medical setting. A realistic dataset such as Transplant is characteristic of these difficulties: a low number of observations, unbalanced classes, time series of different lengths, measurement problems, etc. It is important to point out that the execution time of the algorithms is also crucial. Depending on the problem and the dataset available, we see how important it is to collaborate with experts in the field to choose a suitable approach.

On the other hand, we confirm that the advances made in medical imaging and deep learning are a source of relevant inspiration for the classification of time series. Architectures that have proven themselves on images seem to be a good starting points for an adaptation for time series. The results of the NNGP are also very promising and the use of the package Neural Tangents makes it really easy to compute. Following the work of Novak et al. [14], we know that convolutional layers can also be used in a Bayesian framework. This invites us to reproduce the Bayesian equivalents of other successful architectures such as the 1D CNN and the ResNet.

6 Conclusion

Starting from proven basic architectures, we investigated innovative approaches such as Stochastic Weights Averaging, Class Activation Map and Bayesian neural networks for the classification of time series in a medical setting. Because of their predictive power and interpretability, NNGPs have proven to be relevant in a medical setting. Deepening experimentation with Bayesian Convolutional architectures or even Neural Tangent approaches seem to be a natural continuation of this study. Exploring the possibilities of transfer learning in our context also seems to be an interesting direction.

Medical Time Series Classification proves to be in all cases a vast and challenging research subject which presents specific needs and problems that are very similar to that of medical imaging, and might lead to the development of a new research field.

References

1. A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31:606–660, 2017.
2. O. Colliot and M. Vakalopoulou. Deep leaning for medical imaging. *ENS Paris-Saclay MVA Master*, 2020.
3. DataForGood. Transplant : Prédire l'autonomie respiratoire après transplantation pulmonaire., Feb 2019.
4. H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, Feb 2019.
5. J. Goschenhofer, F. M. J. Pfister, K. A. Yuksel, B. Bischl, U. Fietzek, and J. Thomas. Wearable-based parkinson's disease severity monitoring using deep learning. *CoRR*, abs/1904.10829, 2019.
6. A. Gupta, E. A. Huerta, Z. Zhao, and I. Moussa. Deep learning for cardiologist-level myocardial infarction detection in electrocardiograms, 2019.
7. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
8. H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14(9):1611–1617, Sept. 2019.
9. P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization, 2018.
10. J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. *ArXiv*, abs/1711.00165, 2017.
11. M. Meinard. *Information retrieval for music and motion*. Springer, 2010.
12. X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer Methods and Programs in Biomedicine*, 177:1–8, 2019.

13. R. Novak, L. Xiao, J. Hron, J. Lee, A. A. Alemi, J. Sohl-Dickstein, and S. S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020.
14. R. Novak, L. Xiao, J. Lee, Y. Bahri, G. Yang, J. Hron, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes, 2018.
15. T. D. Pham, K. Wardell, A. Eklund, and G. Salerud. Classification of short time series in early parkinson's disease with deep learning of fuzzy recurrence plots, 2019.
16. S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, Oct 2007.
17. N. Strodthoff and C. Strodthoff. Detecting and interpreting myocardial infarction using fully convolutional neural networks. *Physiological Measurement*, 40(1):015001, jan 2019.
18. W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein. Rethinking 1d-cnn for time series classification: A stronger baseline, 2020.
19. Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585, 2017.
20. Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In F. Li, G. Li, S.-w. Hwang, B. Yao, and Z. Zhang, editors, *Web-Age Information Management*, pages 298–310, Cham, 2014. Springer International Publishing.

Appendix : In this appendix we provide additional information and resources about our work.

A Similarities between imaging and time series in a medical environment.

From a methodological and practical point of view, it is interesting to note the many similarities between imaging and medical time series. Indeed, images and time series are omnipresent high-dimensional objects in the medical environment. The study of these objects represents many perspectives of interactions between medicine and machine learning. By taking up some key points from the introductory session of our course "Deep learning for medical imaging" of the MVA Master (See [2]), we can underline the specificities and similarities that the study of time series in a medical setting shares with medical imaging:

- **Challenges**: The study of time series makes it possible to analyze many phenomena. Coupled with statistical and machine learning methods, we seek to detect sensitive variations inaccessible to the human eye. We can count a multitude of applications: detection of abnormalities, screening, pathologies follow-up, medical decision making, surgical follow-up, research, assisting medical evaluations, predicting the future state of a patient etc.
- **Cross-Road Subject**: At the cross-road between Machine Learning, Statistics and Geometry, the study of time series in a medical framework has specificities that make it a difficult task. Among those specificities : 1) Presence of numerous *multivariate data*: to the high dimension of time series as an object, we add the fact that we often have to follow the evolution of several different indices. 2) *Different lengths*: Whether in the context of a surgical operation or the follow-up of a patient, we will very often find ourselves having to analyse time series of different sizes. 3) *Interpretation*: In medicine, understanding the decisions of our algorithm is essential. The performance is not sufficient to envisage a use in real situations.
- **Data Acquisition**: As with medical imaging, the acquisition of time series in a medical setting relies on the understanding of different *monitoring machines*. In addition to the presence in *real-life situations* of numerous noises (faulty or badly placed sensors, bad contact etc.), the harmonization of the recording *format* across different hospitals (unit of measurement, start of recording etc.) is an important issue. *Labelling* of data sets is also a major issue. Indeed, time series are complex objects that only few experts are able to label. The lack of rich databases correctly labelled is one of the explanations for the difficulty of developing methods that achieve the same performances as in a non-medical context.
- **Invariances**: the theoretical motivation for convolutional networks in imaging dwells in a certain number of invariances and symmetry properties. These properties happen to be quite similar for medical time series (eg invariance by translation: the ECG of an ill heart, when shifted of a few seconds, still represents an ill heart).

B Learning Models

B.1 KNN with Dynamic Time Wrapping

The K Nearest Neighborhood is a well-known algorithm which, combined with an adapted distance, constitutes a strong baseline for the classification of time series [1].

We need to find a distance adapted to time series. The distance that comes naturally first to mind is the Euclidean distance. However, this one has several disadvantages. Firstly, this distance is not robust to time shifts or dilatation. For example, if we take two identical series but slightly shift one, then the Euclidean distance may consider that these series are very different from each other. Secondly, the Euclidean distance only works with series of the same length, which is not realistic in a large number of applications (duration of a surgical operation, etc.).

A correct approach will be to use Dynamic Time Warping (DTW). (More details in Chapter 4 of [11])

Definition 1. *[Warping Path] With $(N, M) \in N^2$, a (N, M)-wraping path is a sequence $w = (w_1, ..., w_L)$ with $w_l = (n_l, m_l) \in [\![1 : N]\!] \times [\![1 : M]\!]$ for $l \in [\![1 : L]\!]$ satisfying:*

- *Boundary condition: $w_1 = (1, 1)$ and $w_L = (N, M)$*
- *Step size condition: $w_{l+1} - w_l \in \{(1, 0), (1, 1), (0, 1)\}$*

We remark that $max(N, M) \leq L \leq N + M$

Given a distance d on R, which we will choose to be the euclidean distance here, we then define the distance between two times series with respect to a wraping path.

Definition 2. *For two univariate time series, $X = (x_1, ..., x_{|X|})$ and $Z = (z_1, ..., z_{|Z|})$, we note $D_{X,Z}$ the function defined on the space of all possible wrap path between X and Z such that*

$$D_{X,Z}(w) = \sum_{l=1}^{L} d(x_{w_{l,1}}, z_{w_{l,2}}) = \sum_{l=1}^{L} |x_{w_{l,1}} - z_{w_{l,2}}|$$

where w is a given wrapping path between X and Y and L is the length of w.

We then define the optimal wrap path.

Definition 3. *[Optimal Warp Path] The optimal warp w^* path between two univariate time series $X = (x_1, ..., x_{|X|})$ and $Z = (z_1, ..., z_{|Z|})$ is*

$$w^*_{X,Z} = argmin D_{X,Z}(w)$$

Definition 4. *[DTW distance] The Dynamic Time Wrapping distance between X and Z is then now defined by*

$$DTW(X, Z) = D_{X,Z}(w^*_{X,Z})$$

The dynamic time warping algorithm has an $O(N^2)$ time and space complexity, we will use the FastDTW algorithm [16] for implementation, which has a linear time and space complexity.

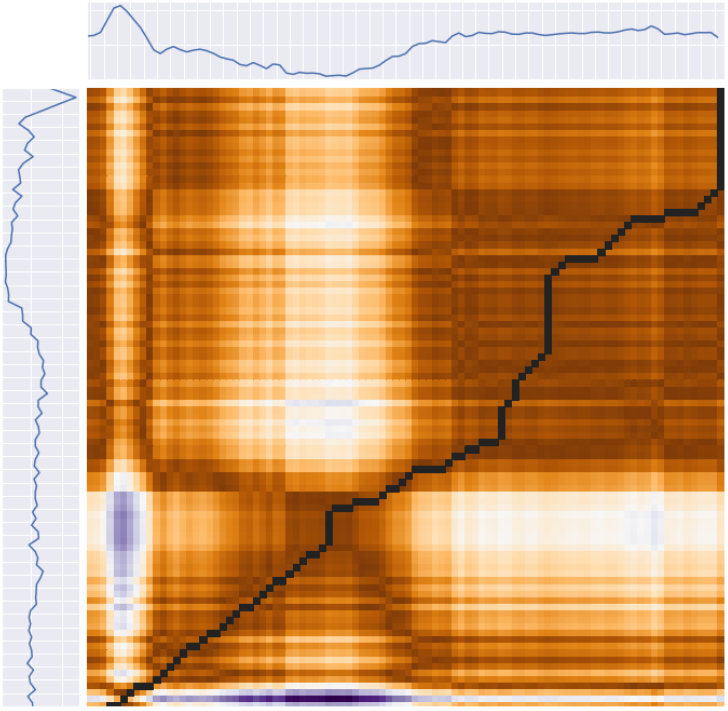

Fig. 5. DTW path (in black) for first and second training examples of the ECG200 dataset.

B.2 A word about Multivariate Time Series Classification

C More Details about the Datasets

C.1 ECG5000

You can see how elements in the ECG5000 dataset look like in Table 6.

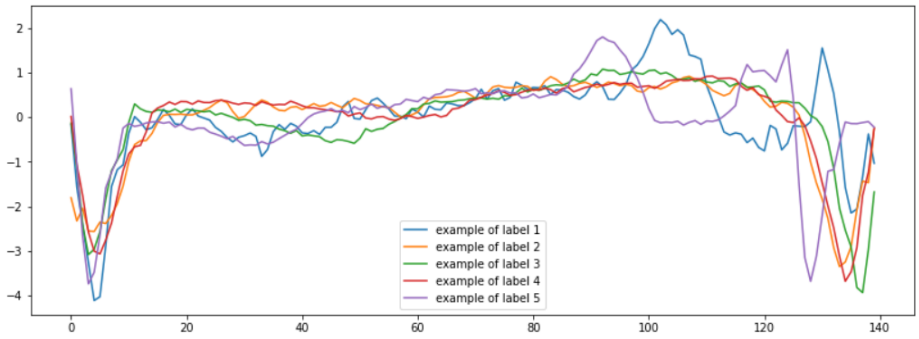

Fig. 6. An example of each label for the ECG5000 dataset.

C.2 MIT-BIH

Figure 7 shows one example of ECG for each class in the MIT-BIH dataset.

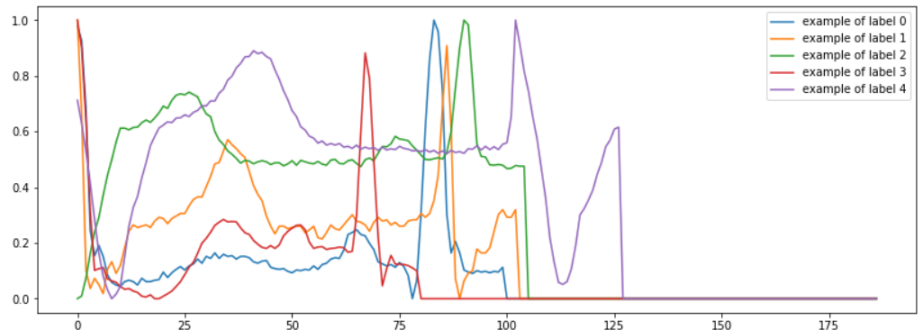

Fig. 7. An example of each label for the MIT-BIH dataset.

C.3 Lung Transplant

About the Transplantation: Lung transplantation is a surgical procedure that involves replacing one or two lungs. These operations are carried out in particular when the patient's lungs are affected by degenerative diseases such as lung fibrose or cystic fibrosis. This operation can give a second chance to patients, often young, condemned by a chronic respiratory insufficiency disease. Lung transplantation is one of the most delicate to perform and one of the most difficult surgical transplantation. The realization of a lung transplant requires

a great technicality because of the fragility of this organ. The success of the surgery depends on factors related to the recipient, transplant and different per operative events.A very important decision at the end of a lung transplant is whether the patient should be extubated or not. Whether or not to keep the artificial breathing system to the patient can have important consequences on his or her state of health.

About the data acquisition: The Foch Hospital is an establishment located in Suresnes in the Hauts-de-Seine. The Foch Hospital specializes in the field of pulmonary and respiratory pathologies. It is the hospital that carries out the most lung transplant operations in France.

The Foch Hospital provided a history of 412 lung transplant patients who received a lung transplant between January 2012 and 2018 (approximately 280k raw rows of information). DataForGood brought a team of data scientist to work on this dataset by creating the database of 330 patients that we use.

About the dataset: The dataset contains :

- Variables characterizing the recipient before the operation
- Variables characterizing the graft before implantation
- Numerous time series extracted from measuring instruments in the operating theatre.
- Time markers collected manually during transplantation.

In this project we have investigated 28 columns of the dataset which were time series. We decided to have a focus on this 10 time series columns :

- BIS SR: The Bispectral Index (BIS) is a scale derived from the measurement of brain electrical activity in anesthetized patients that is used to optimize the depth of anesthesia and drug delivery.
- ETCO2: End-of-Exhalation Oxygen Concentration
- FC : Heart rate
- FR : Respiratory Rate
- FiO2: Inspired fraction in oxygen
- PAPsys: Maximum pulse blood pressure
- PASm: Mean systolic blood pressure
- PEEPtotal: Positive end-expiratory pressure, i.e. the positive pressure that will remain in the airways at the end of the respiratory cycle (end of exhalation) that is greater than the atmospheric pressure in mechanically ventilated patients.
- SpO2: Blood Oxygen Saturation
- SpO2 by FiO2: Blood oxygen saturation by fraction inspired in oxygen

Choices:

A lung transplant surgery lasts on average 9 hours. Our dataset contains operations ranging from 5 to 18 hours of recording. So our time series do not have

the same length in the same dataset. For this reason we decide to calibrate to the largest series (1082 points/minutes of recording) by adding zeros at the end of time series smaller than 1082. Now we have series of same length. As we decided to do univariate time series classification, we have put together the time series of the 10 columns we are studying in order to obtain longer time series.

All in all, this dataset is really challenging. It never has been investigated in a published paper. The data are pretty raw and messy. But this dataset has the great advantage to illustrate how real medical data look like.

You can see how elements in the Transplant dataset look like in Figure 8.

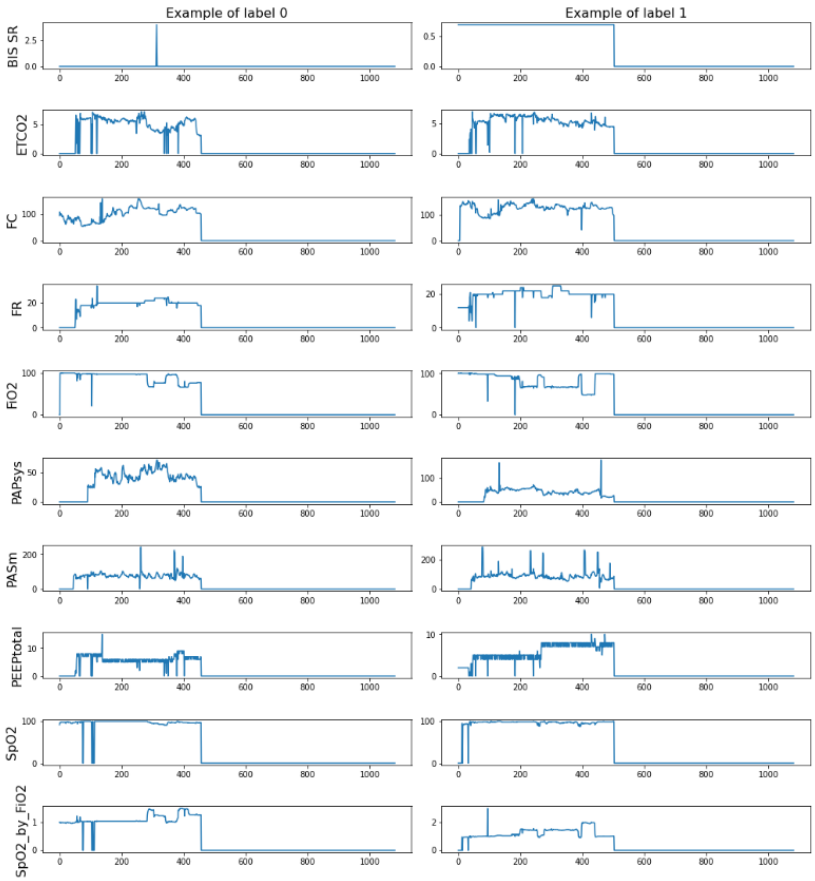

Fig. 8. An example of each label for the Transplant dataset.

D Additional Tables

Recall					
	KNN	MLP	1D CNN	ResNet	NNGP
Transplant	0.48	0.5	0.5	0.5	0.63
ECG5000	0.83	0.39	0.72	0.62	0.60
MIT-BIH	0.82	0.54	0.92	0.92	0.82

Fig. 9. Recall of each dataset for each model.

Precision					
	KNN	MLP	1D CNN	ResNet	NNGP
Transplant	0.48	0.28	0.28	0.28	0.64
ECG5000	0.78	0.36	0.78	0.64	0.67
MIT-BIH	0.90	0.69	0.94	0.92	0.91

Fig. 10. Precision of each dataset for each model.

F1-Score					
	KNN	MLP	1D CNN	ResNet	NNGP
Transplant	0.48	0.37	0.37	0.37	0.60
ECG5000	0.79	0.38	0.74	0.61	0.63
MIT-BIH	0.85	0.58	0.93	0.92	0.86

Fig. 11. F1-Score of each dataset for each model.

E Additional Plots

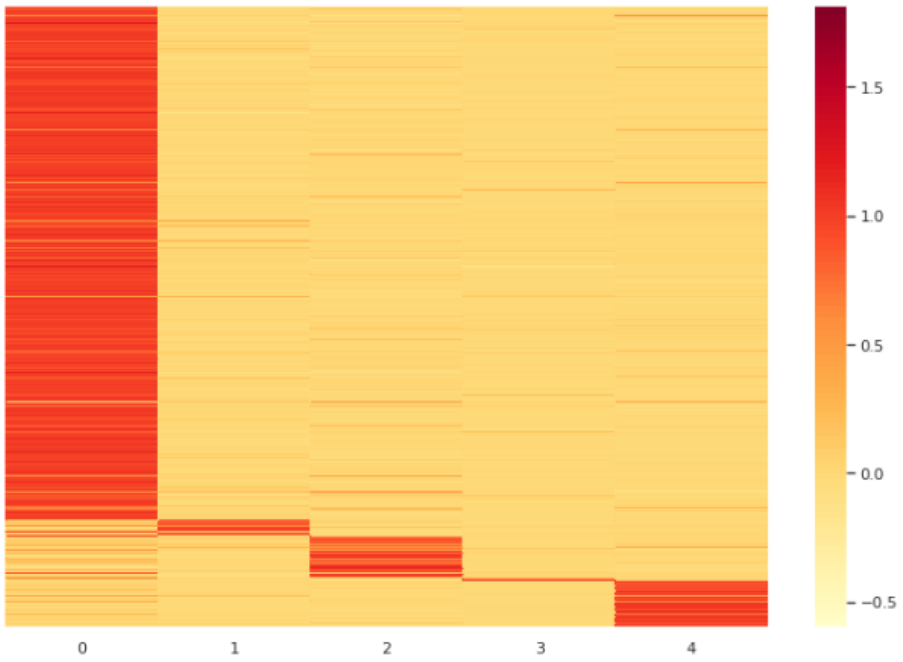

Fig. 12. Posterior Mean of the Neural Network Gaussian Process for the test set of the Mit-Bih: Each line represents one of the 17511 time series of the test set, each column is the value of it's posterior mean for the corresponding label. We have ordered the time series by classes (from 0 to 4, from top to bottom) to observe that high posterior mean for one class compared to the other classes corresponds to correctly classified images.

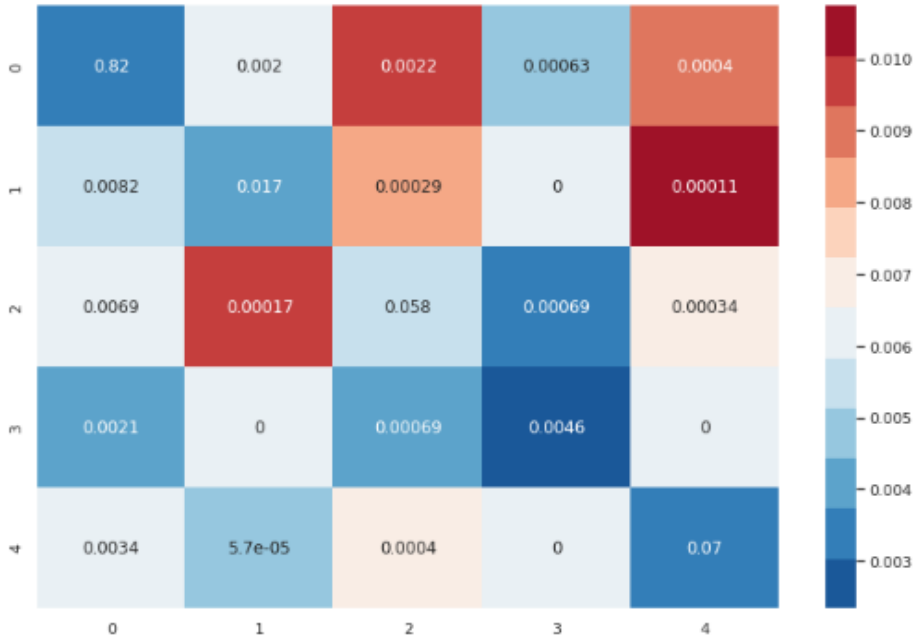

Fig. 13. Confusion matrix (displayed number) for the NNGP with the mean predictive variance (color) on the ECG5000 dataset: Each cell color corresponds to the mean of the predictive variances according to the predicted label (x axis) and true label (y axis). The number display is the one of the confusion matrix of the NNGP model. The posterior variance seems lower on the diagonal (correct predictions). High variances (red) correspond to hesitation of the model which is why having high variances on errors is good.