POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

# Parametric machines: generalization and explainability in deep learning

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

**Author:** Martina Garavaglia

**Advisor:** Prof. Piercesare Secchi

**Co-advisors:** Dr. Mattia G. Bergomi, Dr. Pietro Vertechi

**Academic year:** 2022-2023

## 1. Introduction

Deep learning is becoming progressively more challenging to navigate and comprehend: deep neural architectures are increasing in complexity to deal with novel and more complex problems. We believe that attaining a formal definition of neural architecture and thus being able to represent deep learning models as points in a well-defined mathematical space would accelerate the design and implementation of neural networks, make models more easily shareable, and possibly provide guarantees of convergence and stability of such models.

We aim to compare classical deep-learning networks and novel models drawn from the parametric machines framework [4] on tasks such as classification and forecast of time-varying signals. Moreover, we investigate the foundational parametric-machine framework to discuss the generalization power (i.e., the ability to forecast or classify data with different features than the ones belonging to the training set) of parametric machines and probe their internal dynamics to provide a notion of explainability (i.e., the capacity to give an explanation to the internal process of the model) for such models.

## 2. Mathematical framework

The intuition behind machines is that neural networks can be considered as an endofunction $f : X \to X$ on a space of global functions $X$ (defined on all neurons on all layers). Instead, in a classical deep learning framework, different layers in a network are combined using composition. However, this framework brings with it some disadvantages: shortcut connections are not supported and non-sequential architectures fail to be created.

In [4], the authors consider a global space $X = \bigoplus_{i=0}^{d} X_i$ and the global endofunction

$$f = \sum_{i=1}^{d} l_i \in C^1(X, X)$$

In this framework, classical and novel neural architectures are described as points of a function space. Among these points, we find novel architectures with a rich shortcut structure that could enable us to overcome issues such as vanishing and exploding gradient or instability.

In particular, in a feedforward machine (dense and convolutional), all layers of higher depth take knowledge from all layers of shallower depth, building a network with complete shortcuts.

In the time machine, an additional knowledge dimension is added: the timestamp. All layers, as well as learning from the previous layers, also evolve with past knowledge. In practice, a time machine is a hybrid of recurrent and convolutional architectures, based on parameters choice. The implementation of these architectures can be found in [1].

## 3. Classification

We aim to test parametric machines on time series classification and provide strategies to regularize and interpret the parameters learned by the machine during training. With this aim in mind, we consider the ECG200 dataset, a benchmark data set for time series classification [5]. Each series traces the electrical activity recorded during one heartbeat. Time series are labelled as normal or abnormal heartbeats (myocardial ischemia). Importantly, alterations of the hearbeat signal due to ischemia can be extremely varied. This variability and the complexity of the mechanics underlying heart dynamics make the ECG200 dataset suitable for testing novel techniques and architectures such as parametric machines.
We consider two types of architectures: the dense and time machines (see [2]). Using both, the results surpass the state of the art, reaching an accuracy of 0.9 and 0.91 respectively.

**Regularization** During training, we decide to add a regularization term to the loss function encouraging the model to learn patterns that are consistent over time and reducing noise impact in data. In our scenario, the smoothing process is only applied to the temporal dimension, which involves squaring the difference between model weights in two consecutive data points in the time series:

$$\tau = \lambda \cdot \sum_{t=1}^{T} (w(t) - w(t-1))^2 \qquad (1)$$

In practice, the smoother the weights in time, the smaller the regularization term will be. The strength of time smoothness regularization is controlled by a hyperparameter $\lambda$, which determines the tradeoff between fitting the training data well and keeping the weights smooth.

## 4. Explainability

The word explainability refers to the ability to understand and interpret the decision-making process of a machine-learning model in an input-dependent fashion. In this sense, our aim is to devise a technique that could effectively communicate the workings of deep learning models to individuals who lack expertise in the field.
We devise and implement an explainability module—sensitivity maps—that takes advantage of the formal definition of parametric machines. Since machines are endofunctions on a global space, we compute the derivative of the nonlinear activation function on the linearized machine. On the one hand, sensitivity maps allow us to highlight parts of the signal that are relevant for the model. Hence, sensitivity maps could be useful to the end-user to gain intuition on the model's decisions (see fig. 1, panel a) and b)). In symbols, we express the sensitivity $\rho$ as

$$\begin{aligned} y &= W * z + x_0 \\ z &= \sigma(y) \\ \rho &= \sigma'(y) \end{aligned}$$

where $y$ is the machine's output before the nonlinearity $\sigma$, $W$ is the weights matrix, $x_0$ is the input vector and $z$ the machine's output after the nonlinearity.
On the other hand, via a dimensionality (UMAP) and cardinality (Mapper) reduction algorithms, we provide a measure of classification uncertainty on test data. The two observed clusters (see fig. 1, panel d)) divide the observations based on the degree of loss, providing an indication of their reliability in terms of classification. Specifically, the cluster with lower loss is considered more reliable, while the cluster with greater loss is less reliable. The next step is to identify, given a new observation, which of these two clusters it will belong to, so as to be able to say something about the degree of uncertainty of the classification and to provide to the medical doctor examples of uncertain cases.

## 5. Forecast

The second task we try to solve is a forecasting problem regarding energy consumption time series. Data consists of 255 time series about energy consumption, one per user, sampled every 15 minutes. The differences in distribution

among various users are significant, as the areas and energy usage can vary greatly. These differences can have important implications for understanding patterns of energy usage and identifying areas where energy efficiency improvements may be needed. Furthermore, the ultimate objective is to predict real power demand for week 44.

In this framework, there is not a pre-existing state of the art, so we implement three different models: a time machine and two classical convolutional neural networks (CNNs). We build the two CNNs models to have a fairly comparison with the time machine. The first CNN (CNN-1) has the same layers structure of the time machine. The second model (CNN-2) has the same number of parameters of the time machine (see [3]).

Moreover, we train the models on a single user and test them to all the other. The results proof that the time machine is able to predict more accurately then the sequential models and the train pipeline allows us to show the generalization power of the time machine.

## 6. Generalization

The study led us to observe a fundamental characteristic of parametric machines: their capacity to generalize to unseen data. Initially, the study focused on individual users due to computational issues, but it ultimately led to a surprising result. By analyzing individual users, the model demonstrate very strong generalization abilities that extended to all other users, resulting in highly accurate predictions (see fig. 2).

## 7. Conclusions

We consider a novel deep-learning framework—parametric machines—that generalizes deep neural architectures and provides a formal mathematical definition of operators and models commonly used in deep learning. We apply parametric machines to two case studies: a classification and forecasting problem regarding time series. The two proposed case studies are, by their very own nature, challenging for classical deep-learning models. In both cases, we show that parametric machines outperform the classical models. Moreover, this article has directly addressed two open problems in deep learning: explainability and generalization. The formal mathematical framework shows that parametric machines are good models to delve into these topics. In a forthcoming paper, we plan to further explore the potential of these models in different fields and applications, which could lead to the development of more accurate, reliable, and explainable deep learning solutions.

## 8. Acknowledgements

First and foremost, I would like to express my gratitude to professor Secchi for giving me the opportunity to work on a project in a new and unfamiliar research field for him. I would also like to extend my thanks to Mattia and Pietro, my co-advisors, who have been instrumental in this thesis with their unwavering patience and eagerness to teach.

## References

[1] Mattia G. Bergomi and Pietro Vertechi. `https://github.com/LimenResearch/ParametricMachinesDemos.jl`, 2022.

[2] Martina Garavaglia. `https://github.com/martina-garavaglia-sdg/ischemia_classification_analysis.jl`, 2023.

[3] Martina Garavaglia and Paola Serra. `https://github.com/paola-serra-sdg/Energy_forecast.jl`, 2023.

[4] Pietro Vertechi and Mattia G. Bergomi. Machines of finite depth: towards a formalization of neural networks. 2022.

[5] Xujing Yao, Xinyue Wang, Shui-Hua Wang, and Yu-Dong Zhang. A comprehensive survey on convolutional neural network in medical image analysis. *Multimedia Tools and Applications*, pages 1–45, 2020.

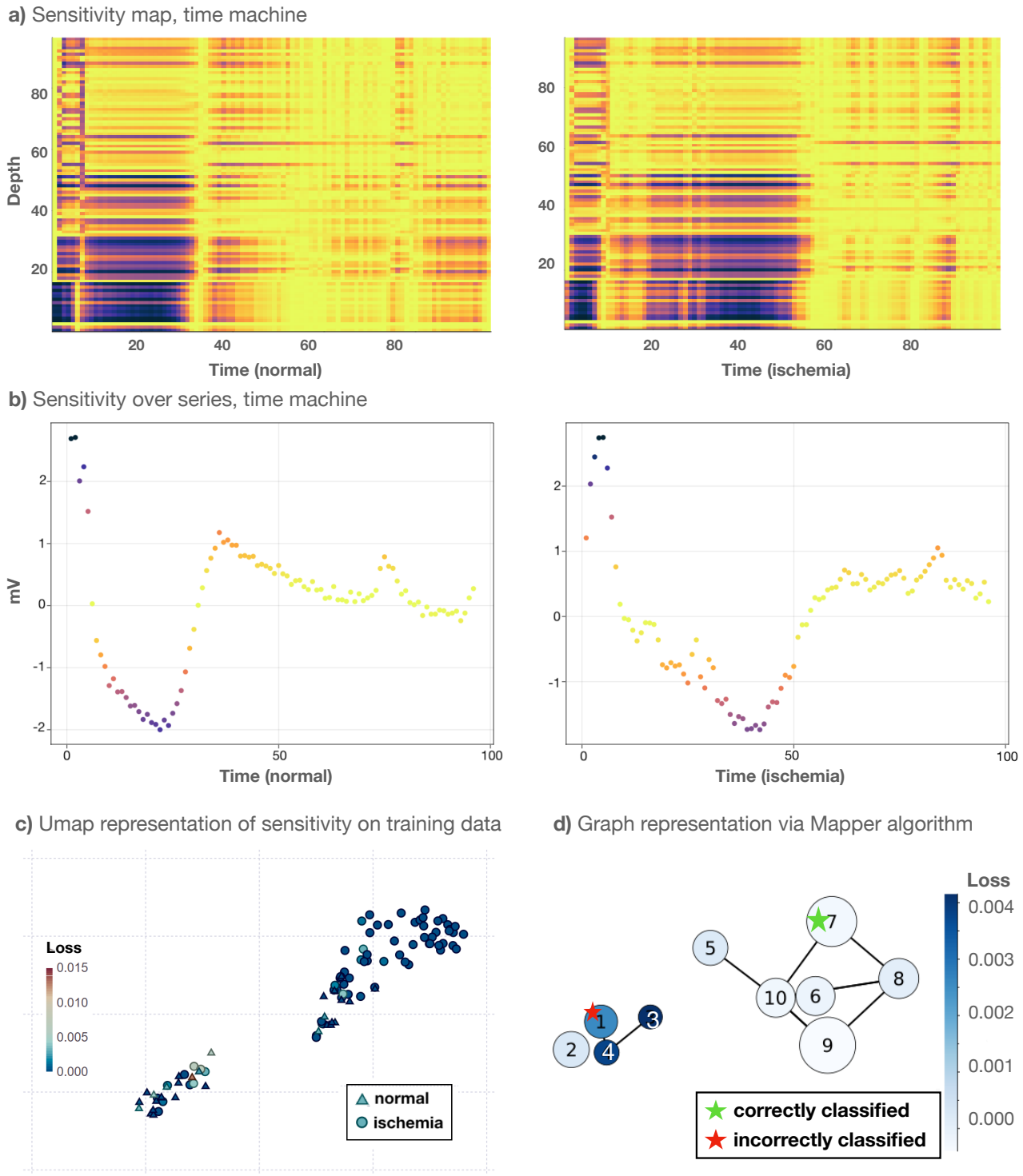


Figure 1: Sensitivity maps, time-series visualization, UMAP and Mapper algorithm representation. **a)** Sensitivity map for a time machine considering a normal and an ischemia time series. **b)** Normal and abnormal heartbeat series colored by sensitivity. We can observe that the most sensitive part of the signal are the peaks in the normal heartbeat that correspond to a ST depression in the ischemia sample and the initial slope. **c)** Sensitivity map visualization after dimensionality reduction (UMAP). **d)** Mapper graph on reduced sensitivity maps. The sensitivity maps obtained from the training set are first vectorized and reduced to 40-dimensional points through UMAP. Then, the projected points are clustered via Mapper. The graph presents connected components organized according to the loss realized by the samples associated with their nodes. The green star represents the mapping of a correctly classified test sample according to its sensitivity map. Symmetrically, the red star corresponds to a misclassified sample.

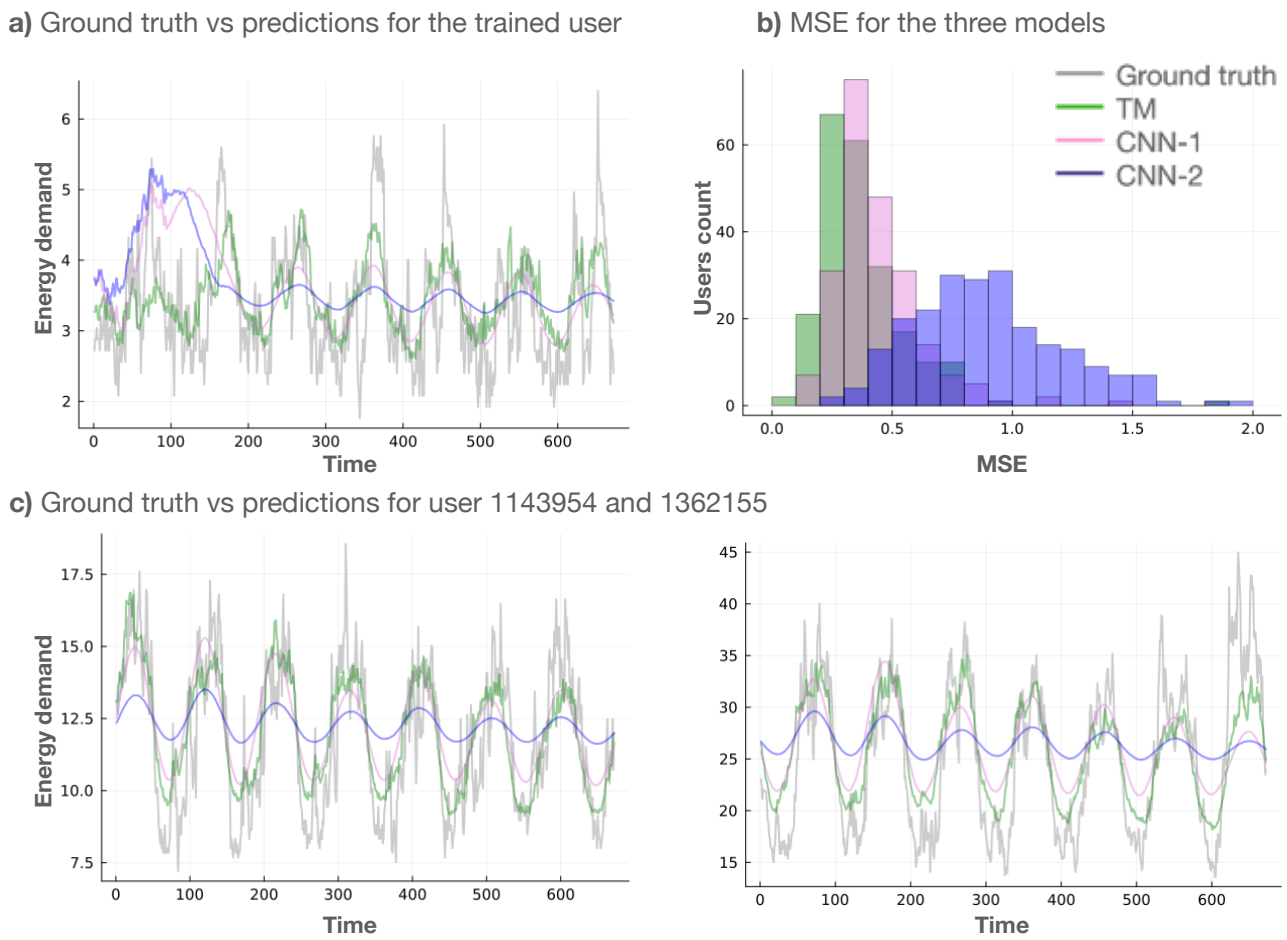


Figure 2: Predictions and performance evaluation. **a)** Ground thruth for the 44th week and predictions for the three models. Parametric machine is able to predict accurately the behaviour of the energy demand, the two CNNs tend to predict a sinusoidal trend over time. **b)** MSE barplot showing that the parametric machine model performs better then the other models in the majority of the observations. **c)** Two new user's predictions example using models trained on a single different user. Parametric machines perform well also on new users showing their generalization power.