

**Data Science
Bootcamp**

Classification du signal d'électrocardiogramme (ECG)



**DAVID MOUGEY
HENRY DOSSOT
MAXENCE DI MARCANTONIO**

Contexte & Problématique

L'épilepsie c'est ...



- Plus de 50 millions de personnes atteintes dans le monde
- 500 000 personnes en France (la moitié a - 20 ans)

Conséquences des crises et états convulsifs :



Lésions/dégradation de l'état intellectuel et moteur du patient

Quels moyens pour lutter ?



Electrocardiogramme (ECG) : enregistrement de l'activité électrique du cœur

Contexte & Problématique

AURA conçoit un dispositif connecté qui détecte les crises et alerte le patient pour qu'il puisse se mettre à l'abri.



ONG française



Fédère ingénieurs et médecins



Politique de science ouverte

Objectif :

"Classifier la qualité du signal d'un ECG selon le bruit chez des patients épileptiques"

Sommaire

I. Exploration des données (EDA)

II. Caractérisation du signal
(Fourier, Ondelettes)

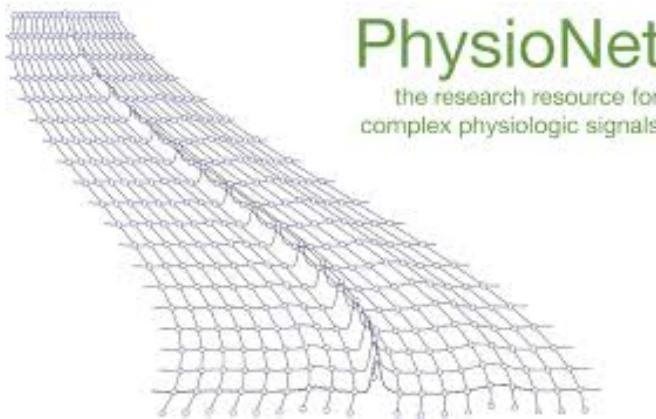
III. Modèle d'apprentissage

IV. Pistes d'amélioration

I. Exploration des données (EDA).

- Présentation des données
- Traitement des données
- Dataframe concaténé

Présentation des données



- 18 patients
- Données d'ECG brutes ".dat" (90M d'enregistrements)
- Données annotées ".csv" (qualité signaux noté 1, 2 & 3)
- 1 observation = 24h d'enregistrement

Base Physionet

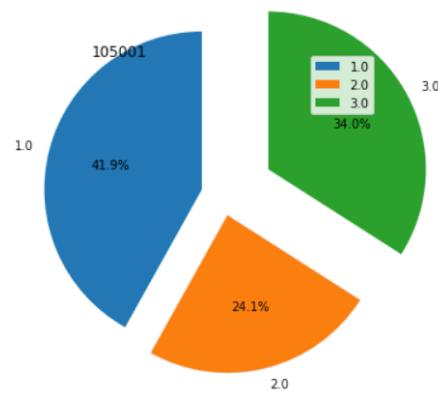
Retraitemet avec nomenclatures des colonnes en header :

	start1	end1	class1	start2	end2	class2	start3	end3	class3	start	end	class	record	signal_length	:
0	1.0	198867.0	2	1.0	19525.0	1	1	7047	2	1.0	7047.0	2	100001	7046.0	
1	198868.0	320282.0	1	19526.0	28694.0	2	7048	17209	1	7048.0	17209.0	1	100001	10161.0	
2	320283.0	373109.0	2	28695.0	32739.0	1	17210	28390	2	17210.0	28694.0	2	100001	11484.0	
3	373110.0	2197974.0	1	32740.0	96699.0	2	28391	32653	1	28695.0	32653.0	1	100001	3958.0	
4	2197975.0	2582746.0	2	96700.0	110564.0	1	32654	71061	2	32654.0	112474.0	2	100001	79820.0	

Traitement des données

étape 1 : exploration des fichiers

- ⚠️ Attention à la durée de la première observation ...13h d'enregistrement (patient 1)

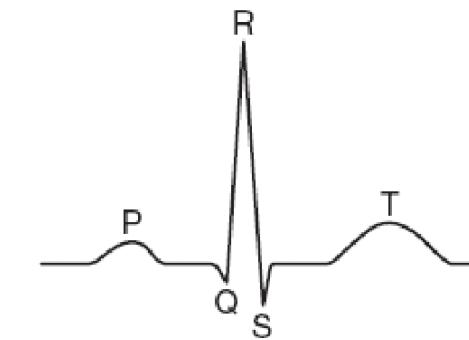


étape 2 : choix des données à utiliser

Visualisations des fichiers par répartition des 3 classes

étape 3 : découpage en durée 2000 ms

Les données brutes en 2000 features pour capter le signal complet.



étape 4 : rééchantillonnage à 250 Hz

réduction fréquence du signal (1000Hz à 250Hz)

1/8 points avec période de 2 secondes)

étape 5: concatenation matrice/array label

Création d'un dataframe pandas

Dataframe concaténé

	0	1	2	3	4	5	6	...	244	245	246	247	248	249	label
0	-187.0	-202.0	-188.0	-187.0	-204.0	-210.0	-214.0	...	331.0	375.0	395.0	454.0	487.0	518.0	2.0
1	-342.0	-336.0	-335.0	-329.0	-319.0	-316.0	-302.0	...	-102.0	-105.0	-107.0	-108.0	-114.0	-111.0	3.0
2	548.0	521.0	503.0	533.0	498.0	476.0	462.0	...	479.0	485.0	491.0	478.0	477.0	451.0	2.0
3	-417.0	-437.0	-459.0	-453.0	-465.0	-448.0	-420.0	...	87.0	68.0	18.0	-33.0	-69.0	-115.0	1.0
4	318.0	365.0	533.0	785.0	918.0	789.0	760.0	...	902.0	905.0	901.0	914.0	928.0	929.0	3.0

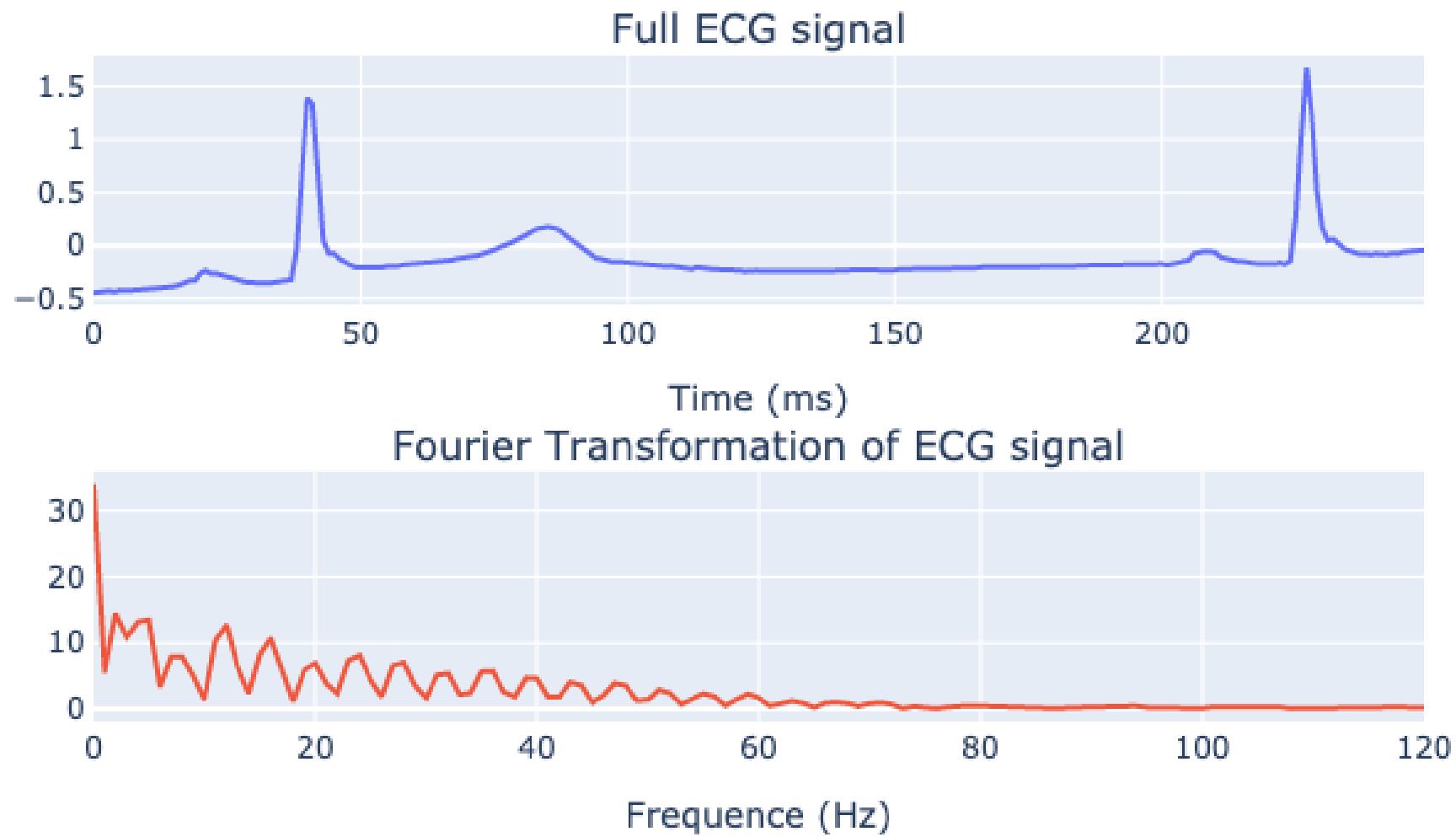
Dataset concaténé avec le signal de 250 features dont "label" en dernière colonne.

III. Caractérisation du signal (Fourier, Ondelettes).

- Transformée de Fourier
- Représentations visuelles des signaux ECG (Transformée des Ondelettes)
- Scalogrammes

Caractérisation du signal ECG : Transformée de Fourier

Son objectif : décomposer le signal d'origine en une somme de ses signaux plus simples.



Point négatif :

Fonctionne lorsqu'aucune variation dans le temps ne se produit.

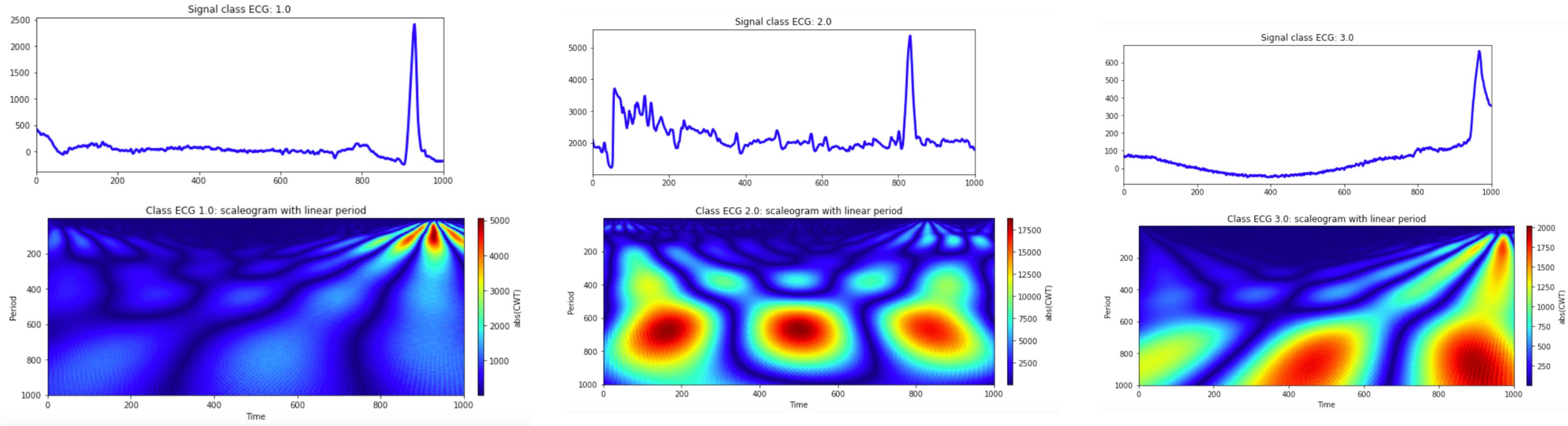


A retenir : Elle ne prend pas en compte la notion de variation dans le temps.



Or ECG = temps ET fréquence

Représentations visuelles des signaux ECG



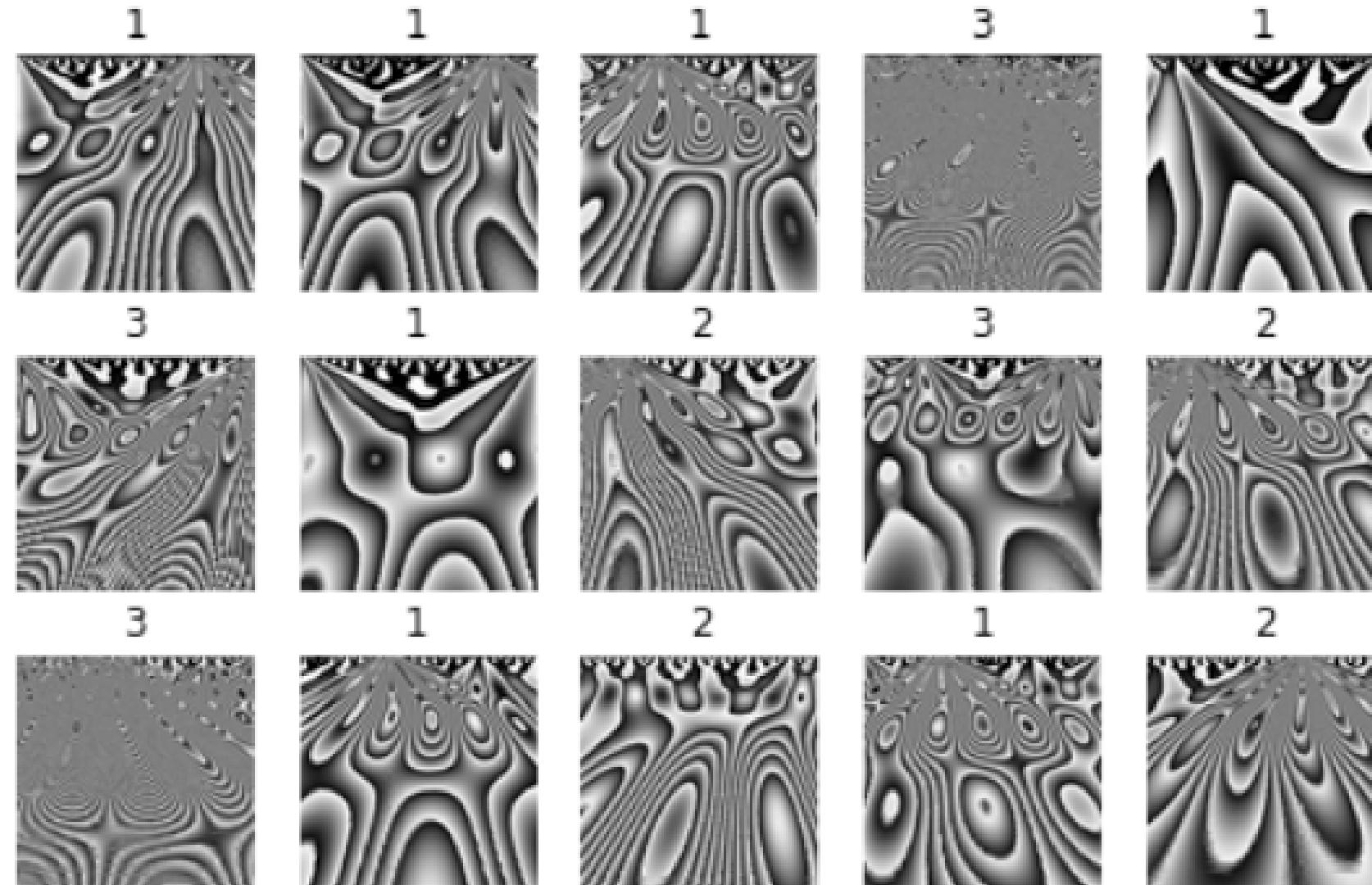
classe 1

classe 2

classe 3

Les résultats sont l'application d'une transformée en ondelettes appelés "**scalogrammes**" avec temps (en abscisse) et fréquence (en ordonnée).

Scalogrammes



La méthode des Ondelettes sert à caractériser le signal en 2 dimensions (fréquence & temps).



A retenir : ces images vont être entraînées dans un modèle de réseau de neurones.

Représentations des signaux à partir de la méthode des Ondelettes.

III. Modèle d'apprentissage

- Avec un réseau de neurones simple
- Avec un réseau de neurones de convolution (CNN)

Avec un réseau de neurones simple

Rappel : prédire la classification des signaux ECG en fonction du bruit observé.

Architecture du modèle :

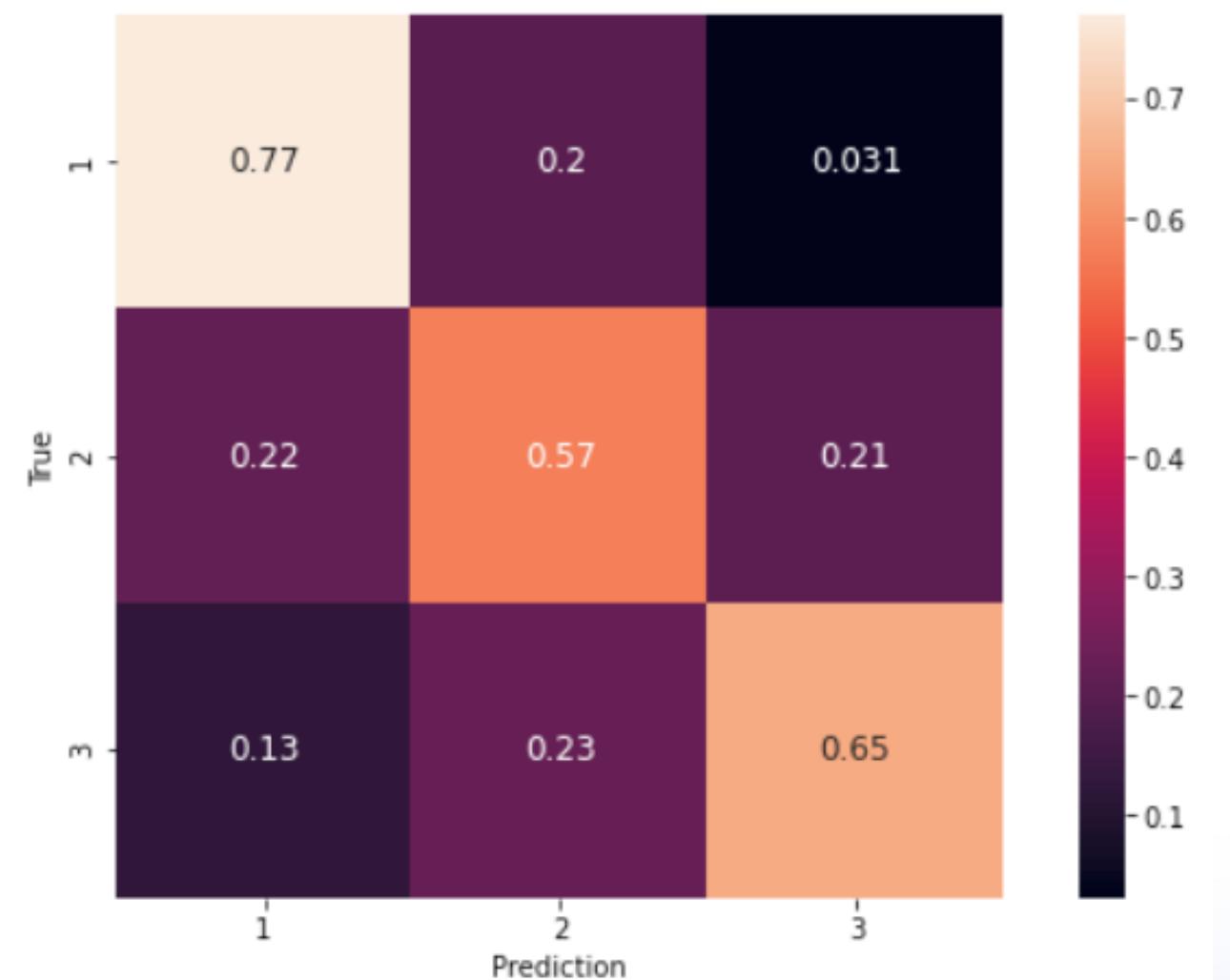
Model: "sequential_1"

Layer (type)	Output Shape	Param #
=====		
flatten_1 (Flatten)	(None, 186003)	0
dense_3 (Dense)	(None, 1028)	191212112
dense_4 (Dense)	(None, 512)	526848
dense_5 (Dense)	(None, 300)	153900
dense_6 (Dense)	(None, 100)	30100
dense_7 (Dense)	(None, 3)	303
=====		
Total params: 191,923,263		
Trainable params: 191,923,263		
Non-trainable params: 0		

3000 images en entrée, 30 Epochs

loss: 0.8112 - accuracy: 0.6667

Résultats du modèle avec une matrice de confusion :



Avec un réseau de neurones de convolution (CNN).

Architecture du modèle :

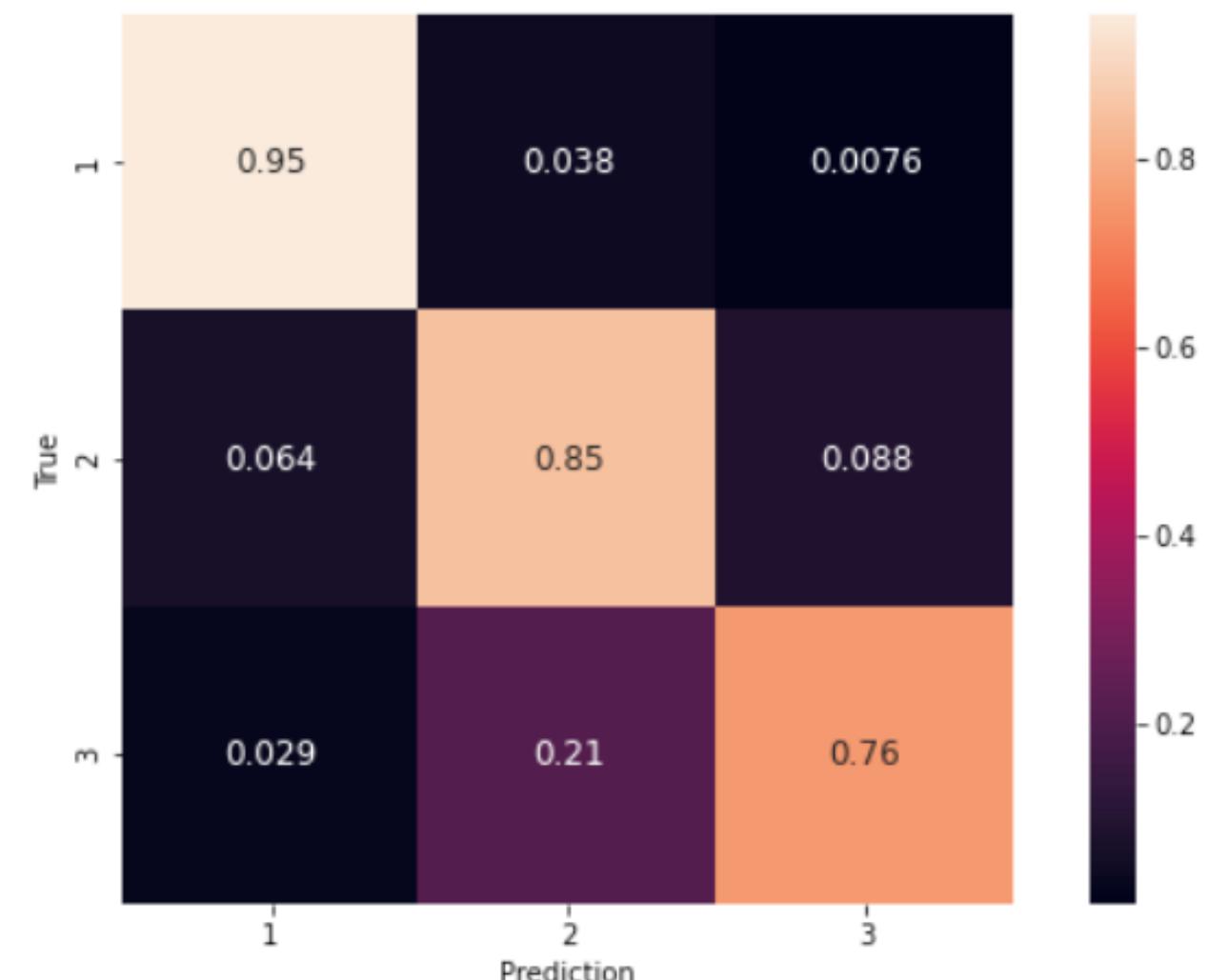
```
Layer (type)          Output Shape         Param #  
=====          ======         ======-----  
conv2d (Conv2D)      (None, 249, 249, 32)    896  
max_pooling2d (MaxPooling2D) (None, 124, 124, 32)    0  
conv2d_1 (Conv2D)      (None, 124, 124, 32)    9248  
max_pooling2d_1 (MaxPooling2 (None, 62, 62, 32)    0  
conv2d_2 (Conv2D)      (None, 62, 62, 64)    18496  
max_pooling2d_2 (MaxPooling2 (None, 31, 31, 64)    0  
flatten_2 (Flatten)    (None, 61504)        0  
dense_8 (Dense)       (None, 64)           3936320  
dropout (Dropout)     (None, 64)           0  
dense_9 (Dense)       (None, 32)           2080  
dense_10 (Dense)      (None, 16)           528  
dense_11 (Dense)      (None, 3)            51  
=====-----  
Total params: 3,967,619  
Trainable params: 3,967,619  
Non-trainable params: 0
```

3000 images en entrée, 30 Epochs

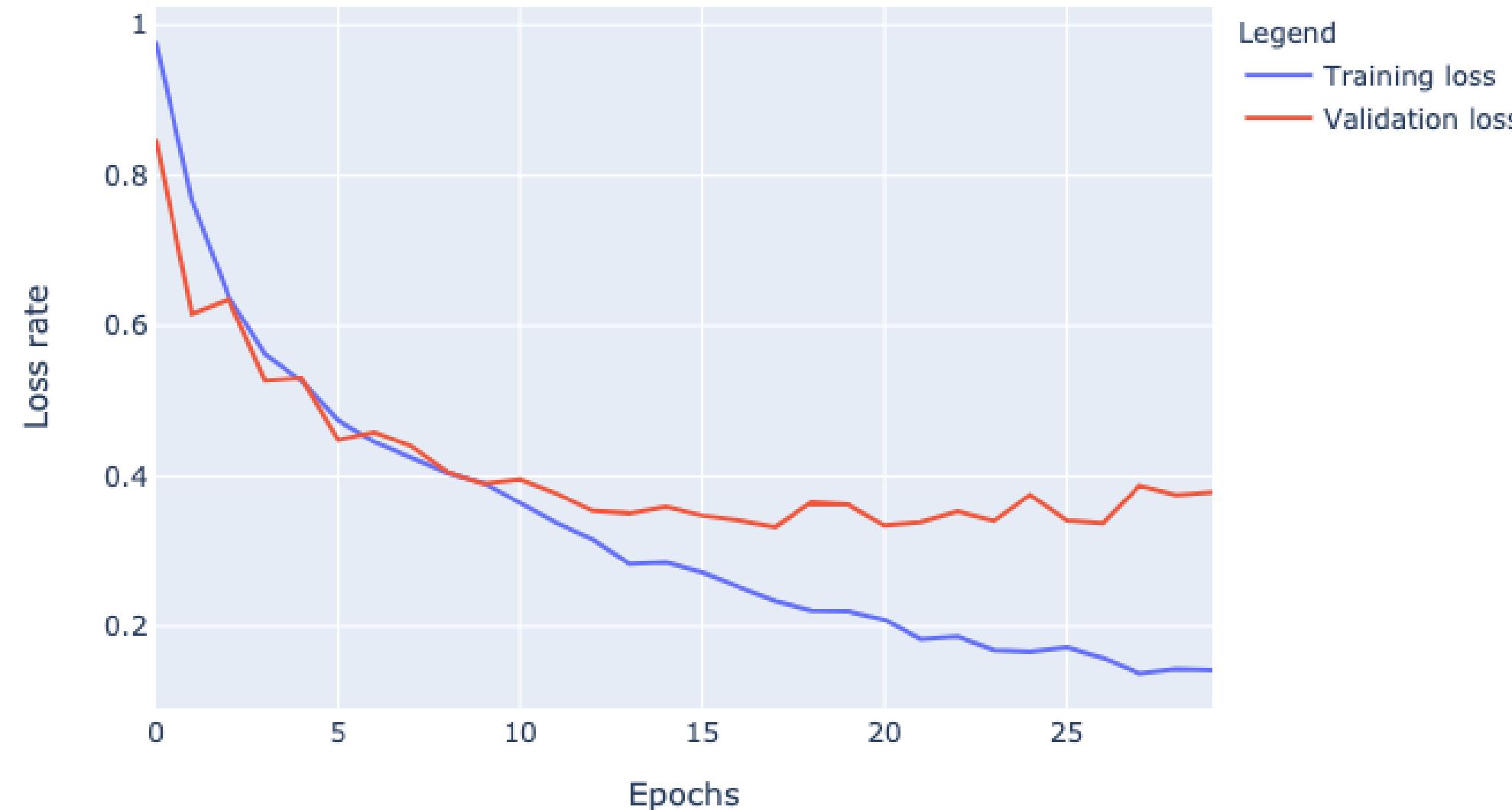


loss: 0.3596 - sparse_categorical_accuracy: 0.8560

Résultats du modèle avec une matrice de confusion :



Observation des courbes de validation & training loss



A retenir : divergence après 10 epochs

Pistes d'améliorations

- Améliorer le rééchantillonnage avec la mise en place d'un filtre passe-bas (diminue la fréquence).
- Augmenter le volume d'images (nécessite + de puissance de calcul).
- Mesurer le niveau de confiance des annotations sur les signaux.
- Moduler les hyperparamètres du CNN (epochs/learning rate/callbacks).
- Augmenter la variété des données (types de patients).
- Limiter à 2 classes pour générer de meilleurs résultats.

Merci pour votre attention

En résumé :

- 1 Exploration et appropriation des données d'un signal ECG.
- 2 Caractérisation du signal avec transformée d'ondelettes.
- 3 Crédit d'un dataset avec images des scalogrammes / entraînement des images dans un réseau de neurones.
- 4 Résultats probants grâce au CNN avec 85% d'accuracy - voire 95% si on se limite à 2 classes.