Projet : Classification du rythme cardiaque de l'ECG

Omar SABRI - Reda EL JAI

6 février 2022

Table des matières

Table des figures

1 Intoduction

L'objectif de ce projet est de réaliser une étude comparative des différents algortihmes de classification d'électrodiagrammes, afin de prédire la nature du battement cardiaque. Nous distinguons 5 classes de battements à détecter : battements normaux, battements inconnus, battements ectopiques ventriculaires, battements ectopiques supraventriculaires et battements de fusion. Dans ce qui suit nous allons détailler les résultats fournis par les classifications se basant sur DTW, ACP et classification par kppv, les forêts aléatoires, les réseaux de neuronnes avant de réaliser une étude personnelle. Nous comparons d'abord les algorithmes sur des bases d'apprentissage et de test réduite et avec 3 classes, avant d'augmenter taille des bases de données et le nombre de classe à 5 dans l'étude personnelle. Cette étude aboutit à un choix de classificateur à entraîner sur l'ensemble de la base de données fournie.

2 Système de reconnaissance d'activité physique avec la DTW

Le système de reconnaissance d'activité physique avec la DTW fournit des résultats encourageants. Avec une précision de 50% sur la base de tets et une matrice de confusion à diagonale dominante.

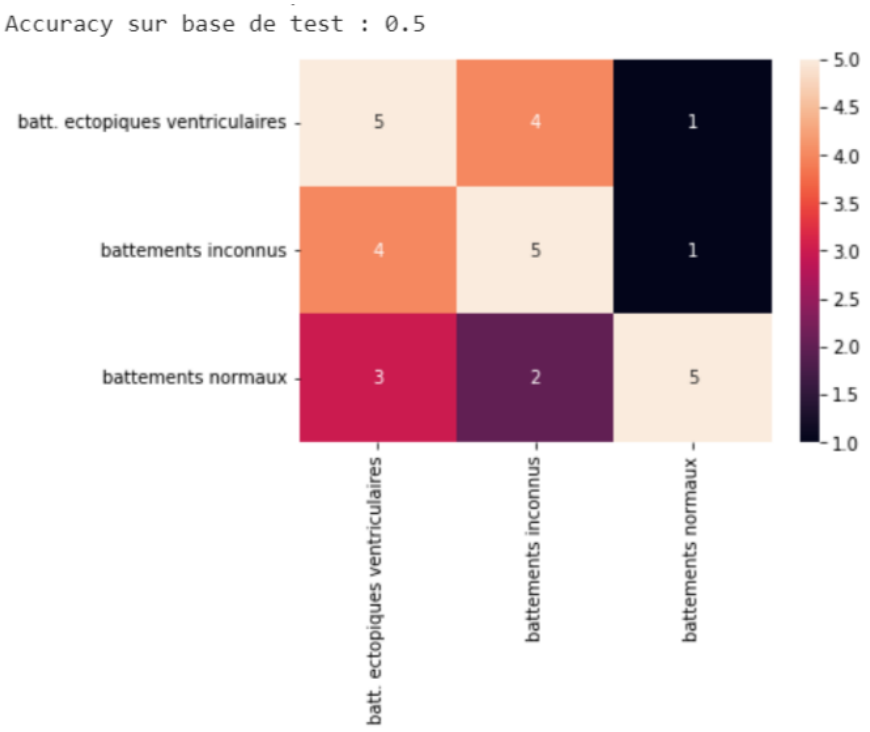


FIGURE 1 – Matrice de confusion dtw

3 comparaison de la programmation dynamique avec une méthode de classification après réduction de dimension par ACP

Une classification à base de k plus proches voisins à la suite d'une analyse en composante principale réalisés en ne gardant que les 3 premières composantes principales, fournit une précision de 0.36 sur la base de test. Le classificateur basé sur DTW est plus adapté aux séries temporelles ce qui explique ses meilleurs performances par rapport à kppv qui est utilisable sur tout type de données.

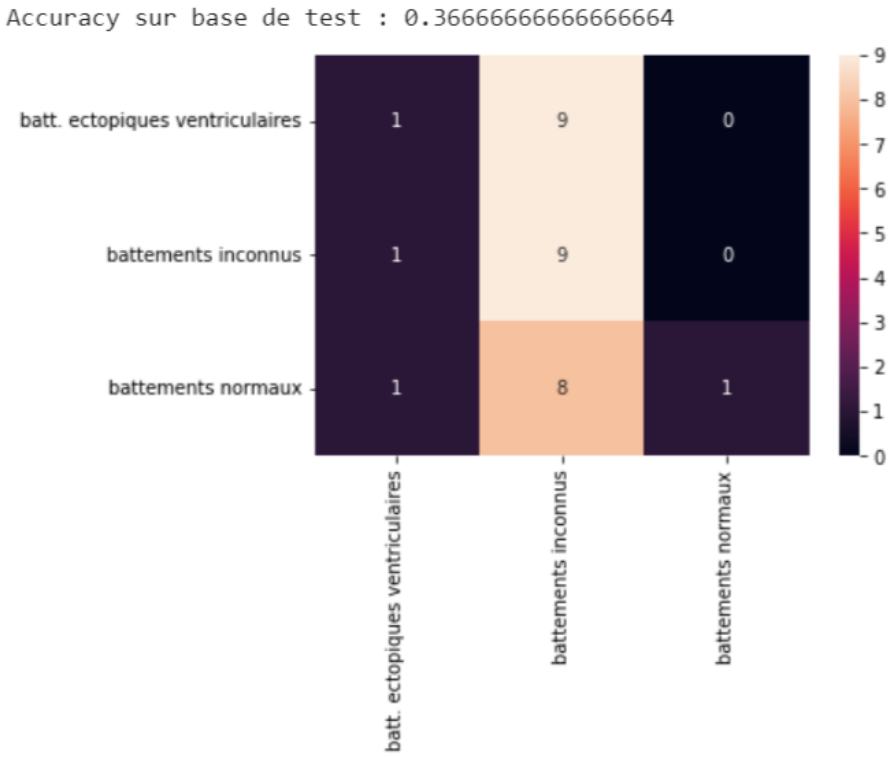


FIGURE 2 – Matrice de confusion k plus proches voisins

4 Comparaison avec d'autres méthodes d'apprentissages supervisés

4.1 Classification par Forêts aléatoires :

La classification à base de forêt aléatoire donne une précision de 40% sur la base de test.

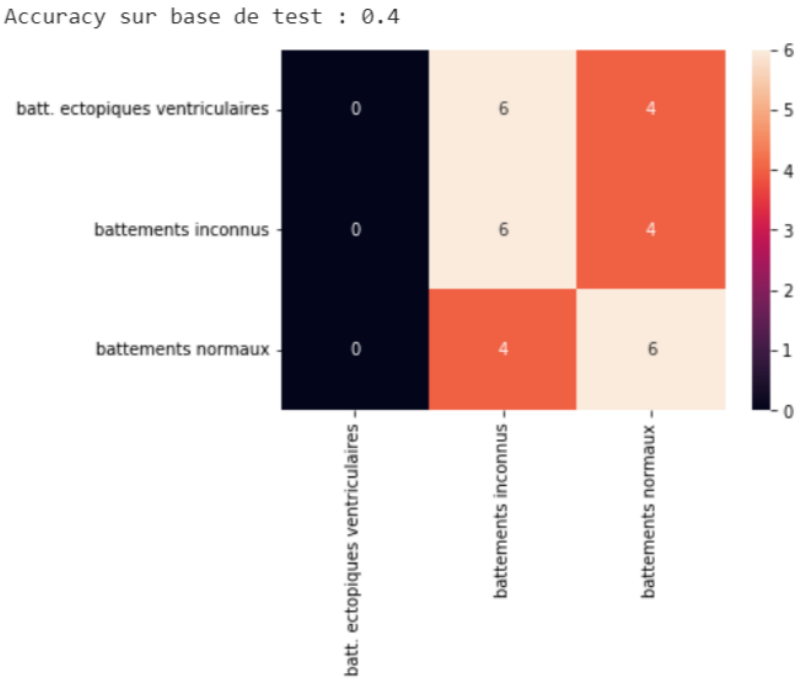
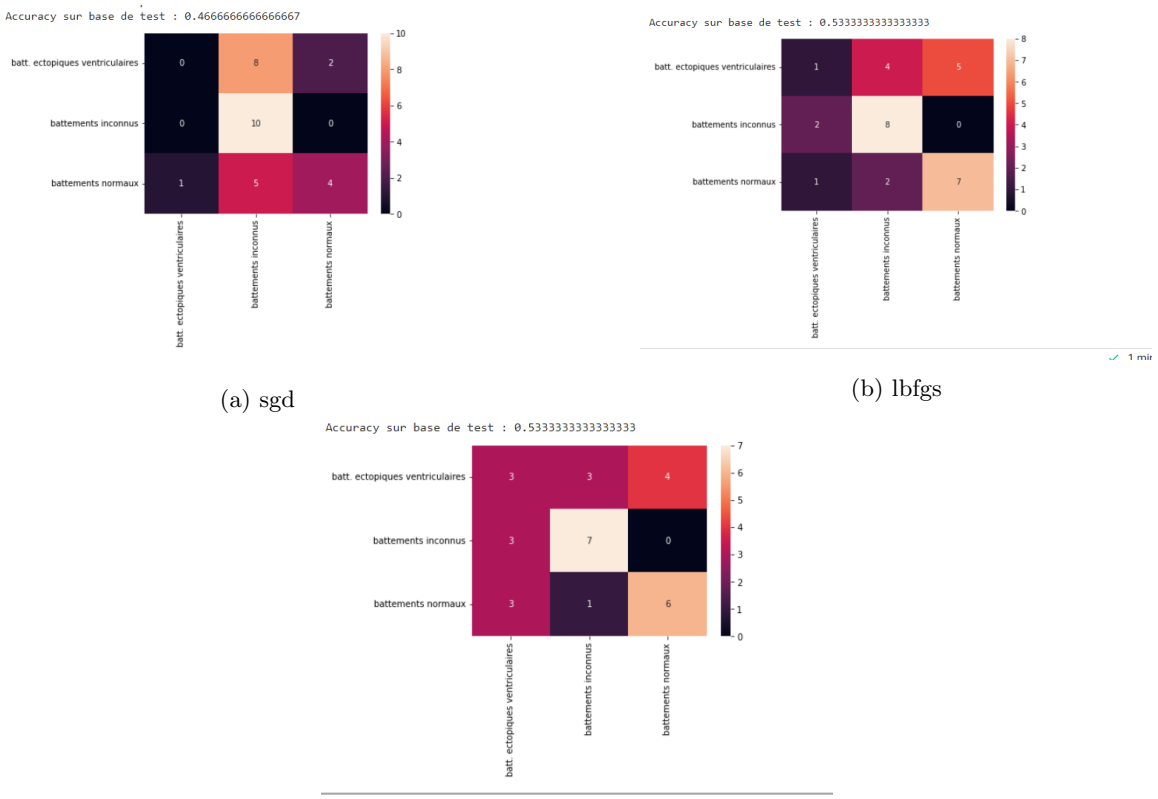


FIGURE 3 – Matrice de confusion random forest

4.2 Classification par réseaux de neuronnes

Nous avons choisi un réseau de neuronnes à base de fonction d'activation relu pour classifier les séries temporelles avec deux couches de 400 et 100 neuronnes. Nous comparons les précisions et les matrices de confusions obtenus pour 3 optimiseurs différents : adam, lbfgs,et sgd. Adam et lbfgs donnent la même précision de 53% alors que sgd est à 46.67%. On devrait s'attendre à de meilleurs résultats quand les bases d'entraînement seront plus importantes. Ils sont sous entraîné dans le cas échéant.



(a) sgd

(b) lbfgs

(c) adam

FIGURE 4 – Matrice de confusion perceptron multicouche pour différents optimiseurs

5 Notre étude :

Dans ce qui suit, nous allons étudier l'impact de l'augmentation de données, et de nombre de classes sur la qualité de la classfication. Pour ce faire, nous allons d'abord passer de 10 et 20 à 150 et 3000 pour les tailles de base de test et d'apprentissage respectivement, en gardant le nombre de classes à 3, puis en l'augmentant à 5 classes. Nous utilisons le même perceptron mulitouche utilisée dans la section d'avant. Par la suite, nous allons comparer ces performances avec celles de trois autres classificateurs : un régresseur linéaire , kppv avec acp et kmeans. A la suite de cette comparaison, nous allons choisir le classificateur avec les meilleurs performances pour l'entraîner sur l'ensemble de données fournies.

5.1 Augmentation de la taille des données

5.1.1 Réseau de neuronnes

L'augmentation de la taille de la base de données d'apprentissage et de test conduit à une amélioration de la précision qui passe à 90.66% sur la base de test.

```
Accuracy sur base de test : 0.9066666666666666
```

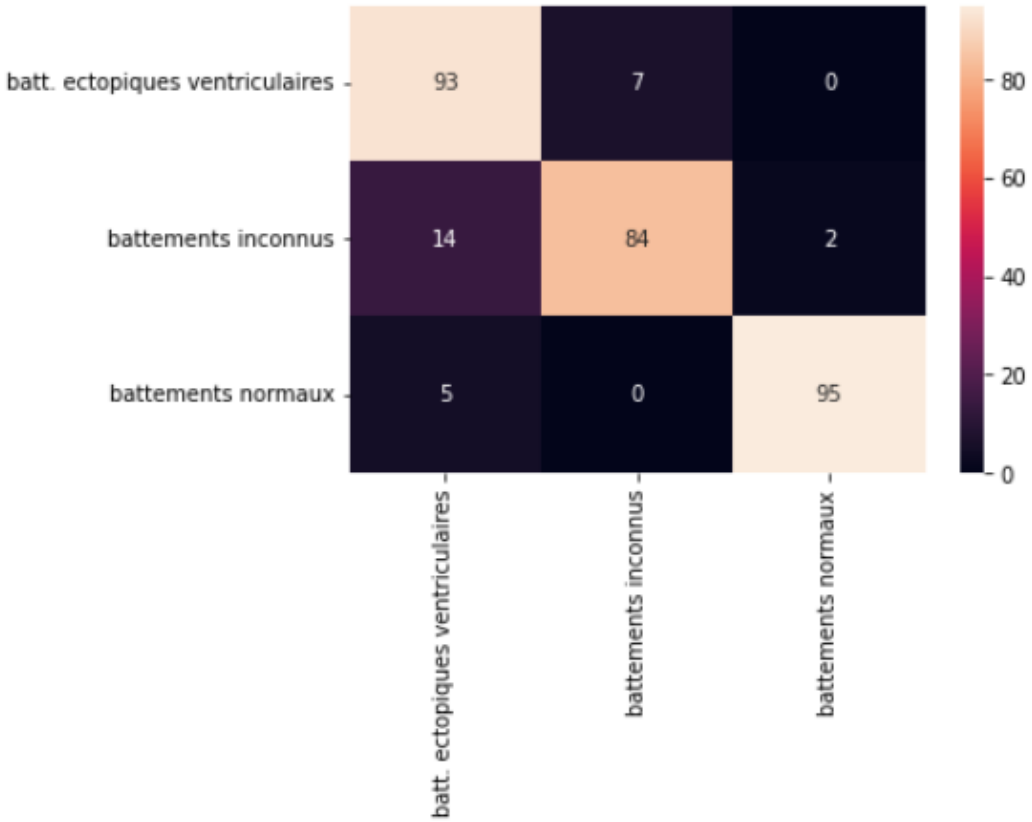


FIGURE 5 – Matrice de confusion réseau de neuronnes avec augmentation de données

5.2 Augmentation du nombre de classes

5.2.1 Réseau de neuronnes

L'augmentation de nombre de classes de 3 à 5 classes pour des tailles de test et d'apprentissage fixées à 100 et 3000 respectivement baisse la précision de prédiction à 88.2% .

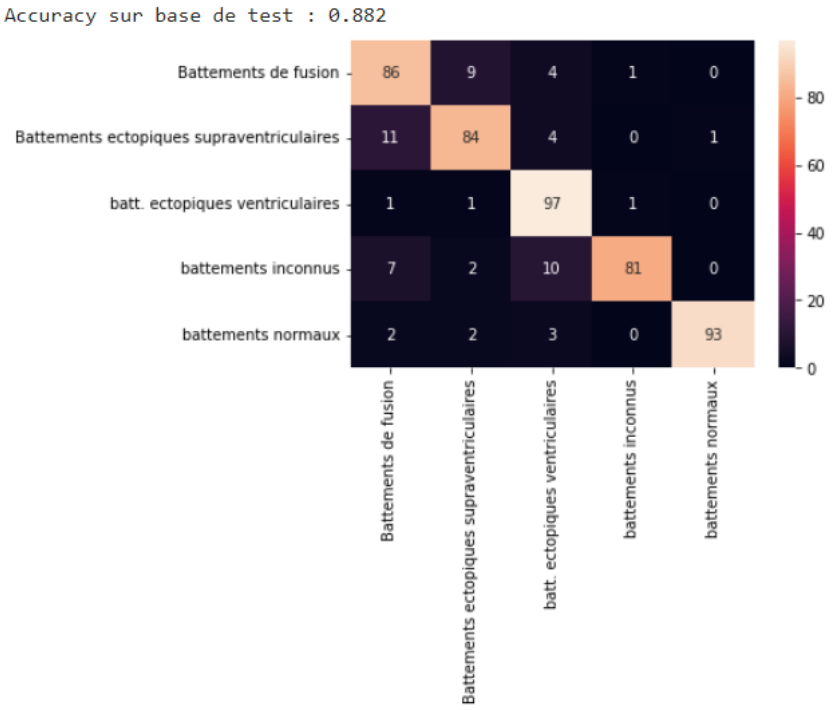


FIGURE 6 – Matrice de confusion réseau de neuronnes avec augmentation de classes

5.3 Comparaison avec un autre algorithme supervisé : régresseur linéaire

Le random forest ne fournissant pas des résultats satisfaisants pour 5 classes, nous avons décidé d'étudier un autre algorithme supervisé qui est le régresseur linéaire. Il permet une classification avec une bonne précision de 76,13%. Sa matrice de confusion est la suivante. Il est moins précis que le réseau de neuronne mais bien plus rapide. (2s contre 47 seconde pour le réseau de neurones).

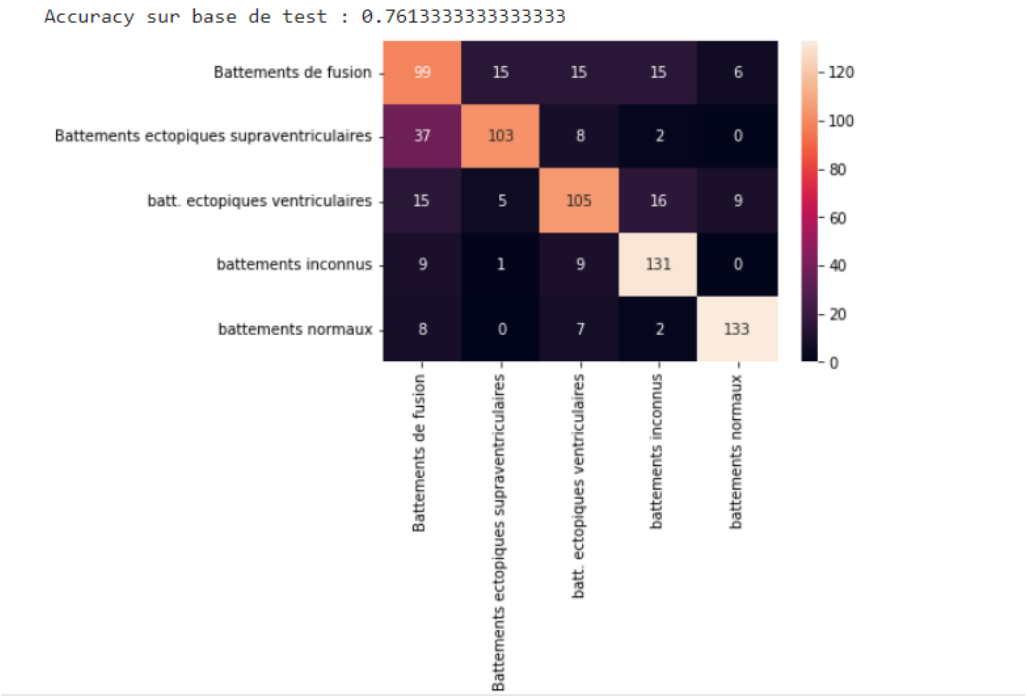


FIGURE 7 – Matrice de confusion du régresseur linéaire

5.4 Comparaison avec d'autres algorithmes non supervisés

5.4.1 kmeans

Le classificateur à base de DTW étant trop coûteux en terme de temps d'éxecution pour 5 classes et 3000 dans la base d'apprentissage, nous avons d'abord pensé à faire un prétraitement grâce à une PCA. Mais le temps d'éxécution restait suffisamment grand pour ne plus être candidat face au réseau de neuronnes en vue de l'implémentation finale (on a arrêté l'éxécution au bout de 30 minutes sans pca et de 15 minutes avec pca). On s'est donc penché sur kmeans en tant qu'autre classificateur non supervisé à sa place. kmeans classifie les signaux avec une précision de 45,46% en initialisant le nombre de clusters à 5 et avec 300 comme nombre d'itérations. De plus La matrice est à diagonale dominante.

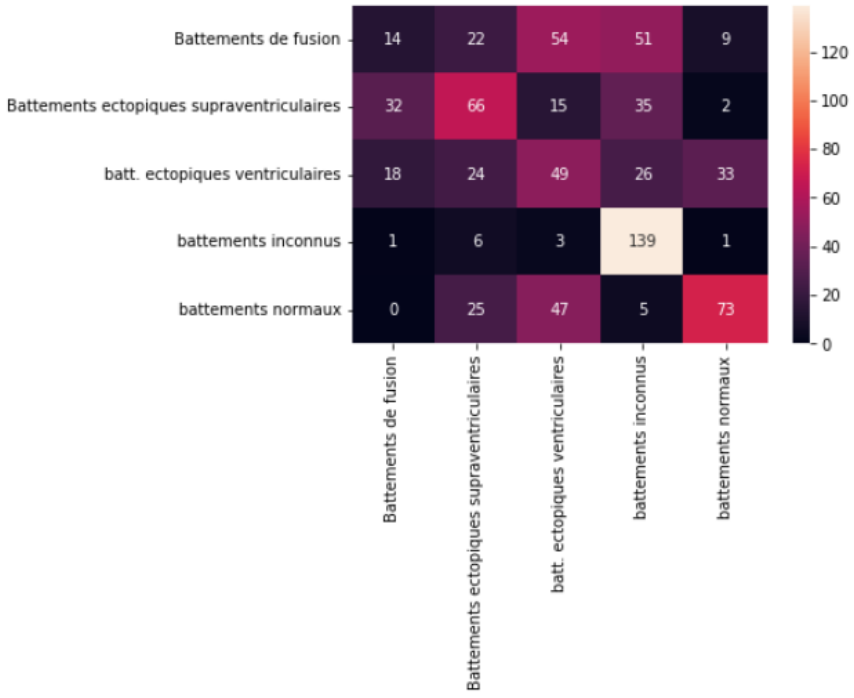


FIGURE 8 – Matrice de confusion de kmeans

5.4.2 kppv avec ACP

kppv précédé d'une ACP présente des résultats mitigés pour 5 classes et des données augmentées : pour des valeurs de k entre 1 et 10, la précision est en moyenne de 23%. La figure suivante représente la matrice de confusion dans le cas k = 3 . Ceci est peut-être dû au fait que la distance entre les classes est trop petite et que les données ne sont pas suffisamment dispersées.

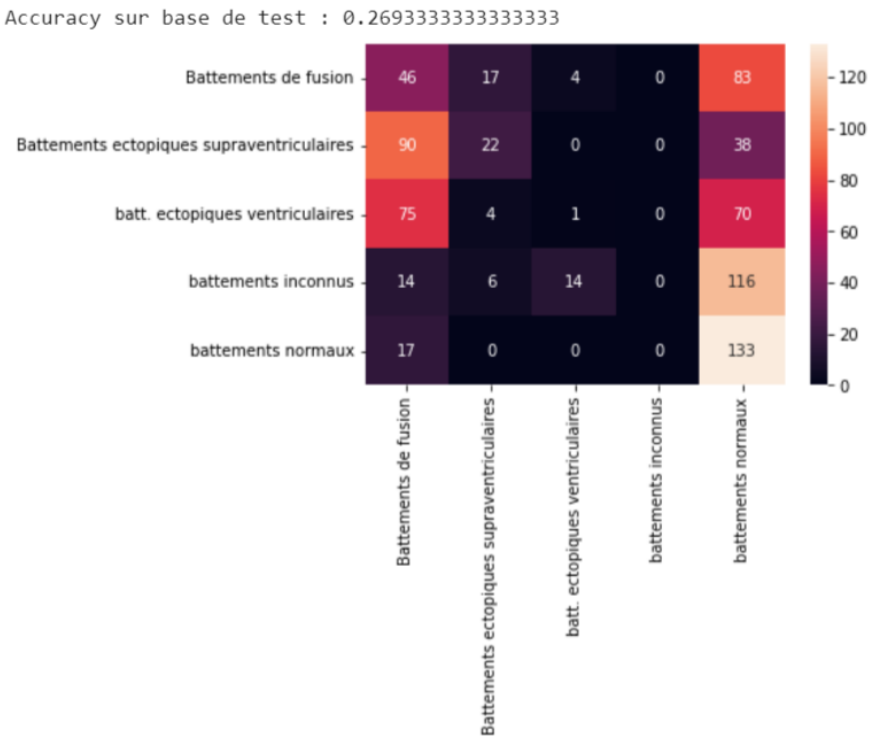


FIGURE 9 – Matrice de confusion perceptron multicouche sur toutes les données

5.5 Utilisation de l'ensemble de la base de données

En s'appuyant sur les résultats précédents, il est clair tout d'abord que les classificateurs supervisés étudié sont plus efficaces que les non supervisés.Et que le perceptron multicouche fournit les meilleurs résultats de prédiction pour 5 classes, sur les bases de données constituées précédemment. Nous avons donc choisi de l'utiliser pour classifier les signaux de l'ensemble de la base de données : Nous obtenons une bonne précision de 79% et la matrice de confusion à diagonale dominante qui suit :

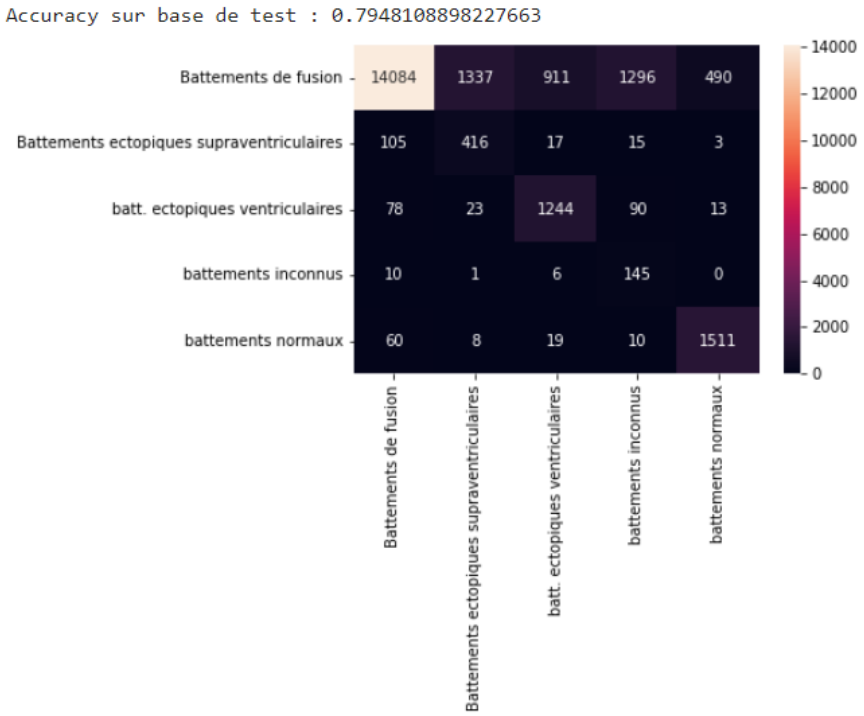


FIGURE 10 – Matrice de confusion perceptron multicouche sur toutes les données

6 Conclusion :

Pour conclure, ce projet nous a permis de comparer différentes techniques de classification des séries temporelles entre K plus proches voisins, k-means, forêts aléatoires, classificateur basé sur Dynamic time Wrapping propre aux séries temporelles, les réseaux de neuronnes et le régresseur linéaire. Nous avons constaté que ces derniers fournissait les résultats les plus satisfaisants. Les performances ce sont d'autant plus amélioré grâce à l'augmentation de taille des bases d'apprentissage et de test.