HEIG-VD — ARN

Laboratoire 3 – Rapport

[redacted]

30 octobre 2023

1 Introduction

Ce travail pratique, basé sur la modélisation guidée par les données, travaille sur la reconnaissance d'un locuteur (homme, femme ou enfant) à partir de sons pré-enregistrés. L'approche de validation croisée a été utilisée pour évaluer les performances des neurones entrainés et sélectionner un modèle final. Finalement, les performances finales du modèle ont été évaluées à l'aide, entre autres, de matrice de confusion.

2 Homme – Femme

Cette première expérience utilise uniquement des voix *naturelles* d'hommes et de femmes pour entrainer le réseau de neurones dans le but de reconnaitre le genre de l'interlocuteur.

L'échantillon de données contient 36 fichiers sonores (.wav) de voyelles prononcées par des femmes et le même nombre d'enregistrements pour les hommes.

Comme cela a été fait dans le premier laboratoire, des valeurs ont été extraites à l'aide du MFCC puis la médiane de ces valeurs a été effectuée afin de ne conserver que 13 valeurs représentatives, et ce, pour chaque enregistrement. Ces valeurs ont ensuite été normalisées avec l'algorithme *MinMax*.

Afin de pouvoir différencier les deux classes utilisées, une $14^{\text{ème}}$ valeur a été ajoutée à chaque entrée du jeu de données :

- 1 pour les hommes
- -1 pour les femmes

Ces valeurs permettent d'utiliser la fonction d'activation tanh, retournant une valeur comprise dans l'intervalle $[-1, 1]$. Un threshold à 0.0 permet une séparation claire de l'intervalle en son centre.

2.1 Exploration du nombre d'epochs

Nous commençons notre étude par la recherche du nombre d'epochs nécessaire pour traiter les données. Ceci est fait au moyen de plusieurs itérations de construction du modèle avec un nombre variable de neurones cachés.

Après plusieurs observations, nous avons déterminé que le learning rate 0.001 et l'élan à 0.8 amènent à une convergence du modèle sans trop d'oscillations.

Les résultats de nos observations se trouvent dans la figure 1. On remarque qu'il y a très peu d'amélioration après 200 epochs et que le modèle converge facilement peu importe le nombre de neurones cachés.

Nous avons donc choisi de garder 200 epochs pour la suite de l'étude du modèle.

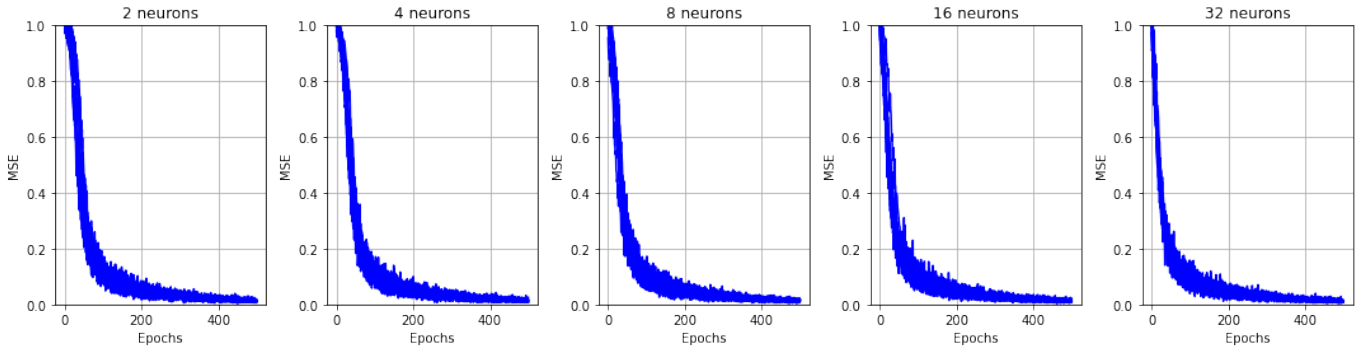

Figure 1 – MSE des données d'entrainement de la première expérience

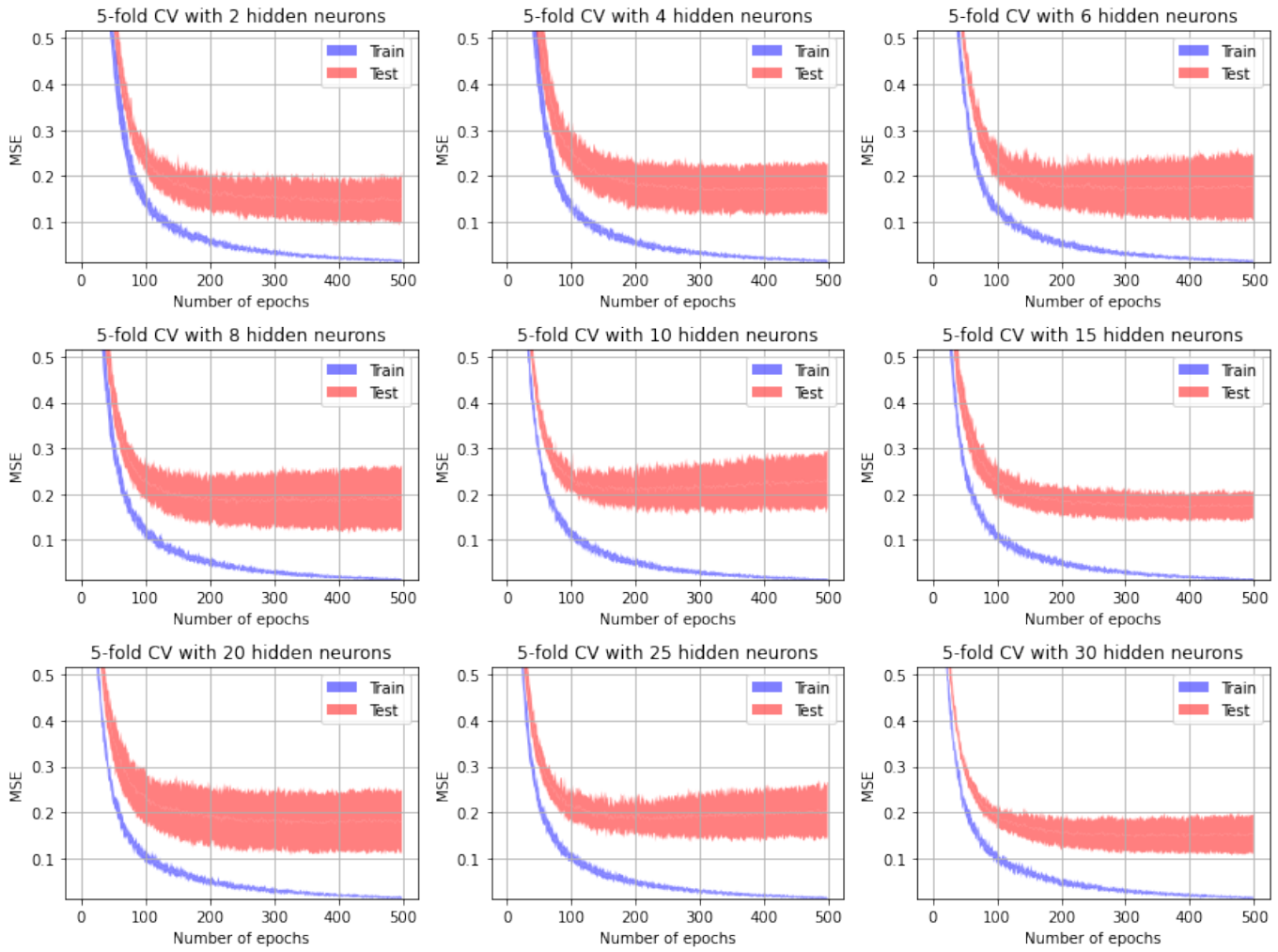

FIGURE 2 – MSE des données d'entrainement et de validation de la première expérience

2.2 Exploration du nombre de neurones cachés

Maintenant que nous avons déterminé le nombre d'epochs, nous pouvons étudier plus en détail le nombre de neurones cachés à utiliser.

On conserve ici le learning rate de 0.001 et l'élan à 0.8.

Pour pouvoir mieux observer les résultats, visibles dans la figure 2, nous avons conservé un nombre d'epochs supérieur à celui estimé précédemment. Cela permet dans notre cas d'observer très explicitement un overfitting du modèle dans la plupart des observations à partir de 250 à 300 epochs.

Nous avons sélectionné ici le modèle avec 4 neurones cachés. En effet, il offrait un bon compromis entre les gains de performance et les courbes MSE qui sont pratiquement équivalentes aux modèles avec un nombre de neurones plus élevés.

2.3 Modèle final et analyse

Pour calculer le modèle final, nous avons donc sélectionné les hyper-paramètres présentés dans le listing 1.

On observe dans la matrice de confusion présentée dans la figure 3 des résultats très bons. Le f1-score de 0.94 est très bon – puisque très proche de 1 – et correspond à nos attentes.

L'erreur MSE d'entrainement est de 0.053 et celle de test est de 0.175, une erreur environ 3 fois plus grande mais qui reste raisonnable. Nous sommes globalement satisfaits de ce résultat.

Listing 1 – Hyper-paramètres utilisés pour la création du premier modèle

```
K = 5
EPOCHS = 200
LEARNING_RATE = 0.001
MOMENTUM = 0.8
THRESHOLD = 0.0
LAYERS = [13, 4, 1]
ACTIVATION_FN = tanh
```

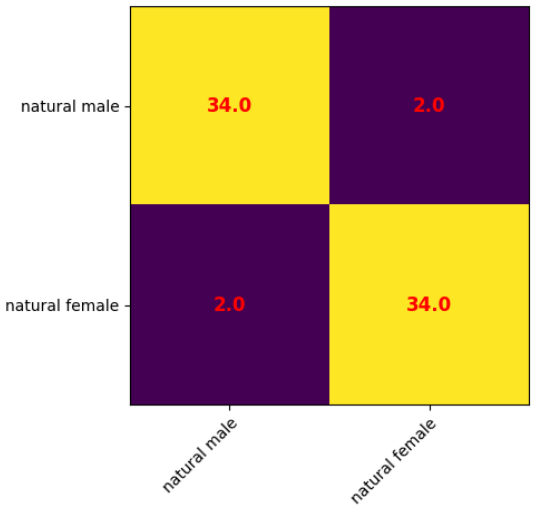

FIGURE 3 – Matrice de confusion des résultats du premier modèle

3 Homme – Femme – Enfant

Cette seconde expérience utilise uniquement des voix *naturelles* d'hommes, de femmes et d'enfants pour entrainer le réseau de neurones, dans le but de reconnaitre le type d'interlocuteur.

L'échantillon de données contient 36 fichiers sonores (.wav) de voyelles prononcées par des femmes et le même nombre d'enregistrements pour les hommes. L'échantillon de données pour les enfants possède cependant 108 fichiers sonores pour les enfants, répartis en trois classes d'âge (3, 5 et 7 ans), il faut donc effectuer un sous-échantillonnage pour n'en conserver que 36. La meilleure solution dans ce cas est de prendre aléatoirement 36 des enregistrements. Cependant, pour garder nos résultats uniforme entre les expériences et par soucis de temps, il a été choisi de simplement prendre une valeur sur trois, soit 12 enregistrements par tranche d'âge.

Les enregistrements sonores ont subi le même traitement que dans le premier test de ce laboratoire, à savoir l'extraction des caractéristiques avec MFCC, les calculs de leurs médianes et la normalisation de l'ensemble des données.

Afin de pouvoir séparer les différentes classes, le codage suivant a été utilisé, ajoutant ainsi 3 nouvelles valeurs à chaque entrée :

- $(1, -1, -1)$ pour les hommes
- $(-1, 1, -1)$ pour les femmes
- $(-1, -1, 1)$ pour les enfants

3.1 Exploration du nombre d'epochs

En continuant sur les observations de la partie précédente, nous avons gardé le learning rate de 0.001 et l'élan de 0.8 qui ont ici aussi amené à une convergence du modèle sans trop d'oscillations.

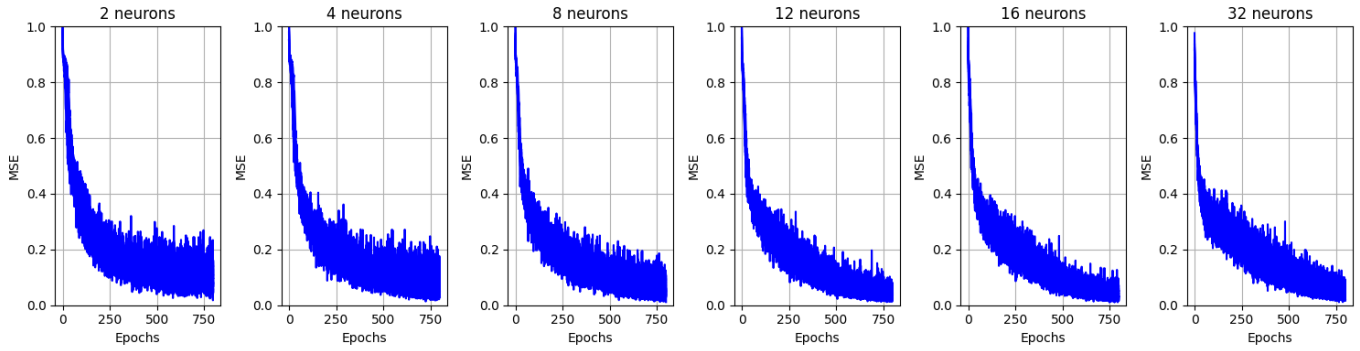

FIGURE 4 – MSE des données d'entrainement de la seconde expérience

Listing 2 – Hyper-paramètres utilisés pour créer le second modèle

```
K = 5
EPOCHS = 400
LEARNING_RATE = 0.001
MOMENTUM = 0.8
THRESHOLD = 0.0
LAYERS = [13, 8, 3]
ACTIVATION_FN = tanh
```

Les résultats de nos observations se trouvent dans la figure 4. On remarque qu'il y a de moins en moins d'améliorations après 400 epochs et que le modèle converge facilement peu importe le nombre de neurones cachés, mais la convergence est parfois plus lente avec un nombre de neurones élevé.

Nous avons donc choisi de garder 400 epochs pour la suite de l'étude du modèle.

3.2 Exploration du nombre de neurones cachés

Maintenant que nous avons déterminé le nombre d'epochs, nous pouvons étudier plus en détail le nombre de neurones cachés à utiliser.

On conserve ici le learning rate de 0.001 et l'élan à 0.8.

Pour pouvoir mieux observer les résultats, visibles dans la figure 5, nous avons conservé un nombre d'epochs supérieur à celui estimé précédemment. On constate moins d'overfitting que dans l'expérience précédente.

Nous avons sélectionné ici le modèle avec 8 neurones cachés. Notre réflexion est essentiellement la même que pour l'expérience précédente.

3.3 Modèle final et analyse

Pour calculer le modèle final, nous avons donc sélectionné les hyper-paramètres présentés dans le listing 2.

On observe dans la matrice de confusion présentée dans la figure 6 des résultats corrects pour la classe « natural male ». Le f1-score correspondant de 0.90 est très bon et correspond à nos attentes.

Cela dit, on peut constater que le modèle a plus de difficultés avec les classes « natural female » et « natural kid ». Les f1-score étant, respectivement, 0.76 et 0.8. Toutefois, cette observation ne nous choquent pas, car il est possible d'imaginer plusieurs similarités entre les voix d'enfants jeunes et des voix féminines.

En faisant l'addition des observations par classe, on peut aussi s'apercevoir que le modèle associe une même valeur à plusieurs classes simultanément. Nous avons essayé de modifier le threshold, mais cela n'a pas donné de résultat.

L'erreur MSE d'entrainement est de 0.12 et celle de test est de 0.33, une erreur étant également environ 3 fois plus grande. Elle montre également que le modèle a plus de difficulté à converger vers un résultat idéal, mais n'effectue quand même pas de l'overfitting.

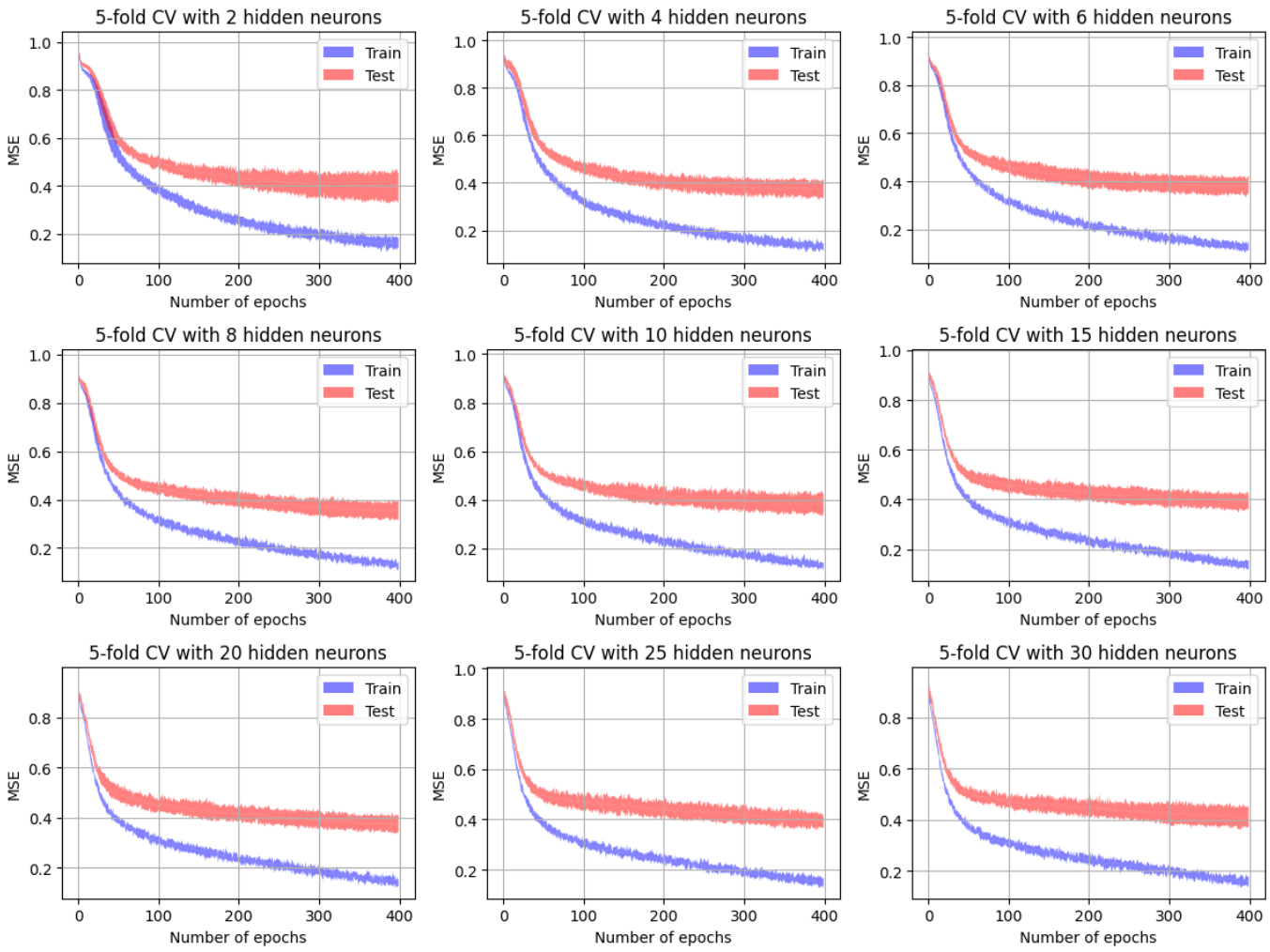

FIGURE 5 – MSE des données d'entrainement et de validation de la seconde expérience

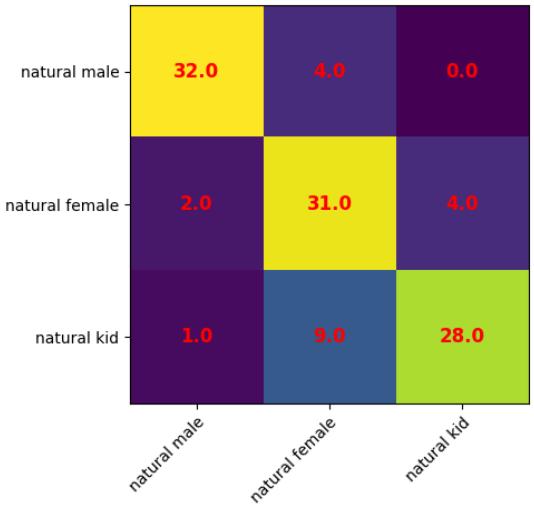

FIGURE 6 – Matrice de confusion des résultats du second modèle

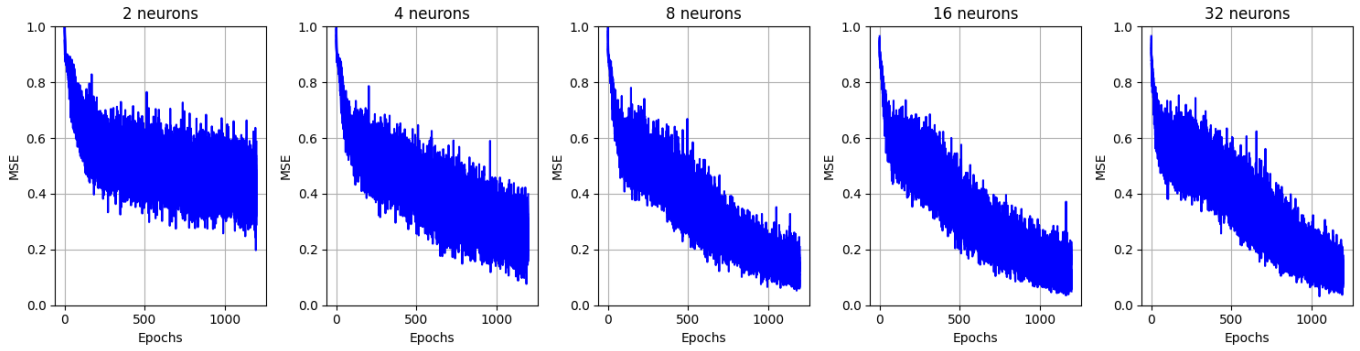

FIGURE 7 – MSE des données d'entrainement de la troisième expérience

4 Enfant de 3 ans – Enfant de 5 ans – Enfant de 7 ans

Cette dernière expérience tentera de différencier les différentes classes d'âges disponibles dans les enregistrements sonores enfants. Pour cela, seules les voix *naturelles* seront utilisées pour entrainer le réseau de neurones, dans le but de reconnaître l'âge de l'interlocuteur.

L'échantillon de données contient 108 fichiers sonores (.wav) de voyelles prononcées par des enfants, divisés en groupe de 36 fichiers par âge.

Les enregistrements sonores ont subi le même traitement que dans les deux autres tests de ce laboratoire, à savoir l'extraction des caractéristiques avec MFCC, le calcul de leurs médianes et la normalisation de l'ensemble des données.

Afin de pouvoir séparer les différentes classes, le codage suivant a été utilisé, ajoutant ainsi 3 nouvelles valeurs à chaque entrée :

- $(1, -1, -1)$ pour les enfants de 3 ans
- $(-1, 1, -1)$ pour les enfants de 5 ans
- $(-1, -1, 1)$ pour les enfants de 7 ans

4.1 Exploration du nombre d'epochs

En continuant sur les observations des deux autres expériences, nous avons gardé le learning rate de 0.001 et pris un élan de 0.8 qui ont ici aussi amené à une convergence du modèle sans trop d'oscillations.

Les résultats de nos observations se trouvent dans la figure 7. On remarque qu'il y a une amélioration constante, plus le nombre de neurones est élevé. On pourrait donc penser qu'il faudra plus de 1200 epochs pour que le modèle converge, mais l'exploration suivante montrera que ce n'est pas un bon choix.

Nous avons donc choisi de garder 100 epochs pour la suite de l'étude du modèle.

4.2 Exploration du nombre de neurones cachés

Maintenant que nous avons déterminé le nombre d'epochs, nous pouvons étudier plus en détail le nombre de neurones cachés à utiliser.

On conserve ici le learning rate de 0.001 et l'élan de 0.8.

Pour pouvoir mieux observer les résultats, visibles dans la figure 8, nous avons conservé un nombre d'epochs supérieur à celui choisi précédemment. On constate un clair overfitting peu importe le nombre de neurones, confirmant qu'un nombre plus élevé d'epochs n'aurait servi à rien.

Nous avons sélectionné ici le modèle avec 6 neurones cachés. Notre réflexion restant dans la même optique que pour les deux expériences précédentes, à savoir un compromis entre performances et MSE.

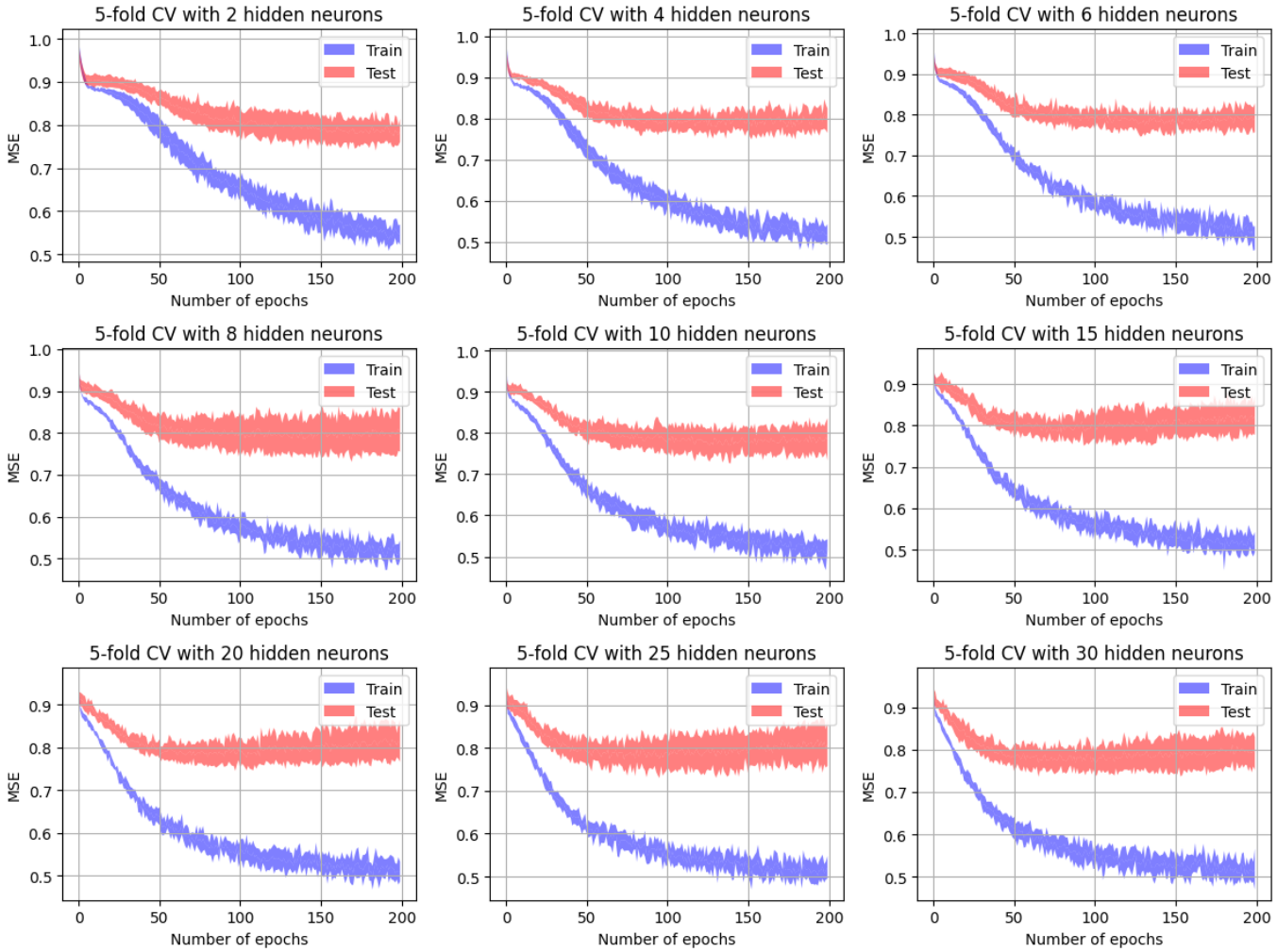

FIGURE 8 – MSE des données d'entrainement et de validation de la troisième expérience

Listing 3 – Hyper-paramètres utilisés pour créer le troisième modèle

```
K = 5
EPOCHS = 100
LEARNING_RATE = 0.001
MOMENTUM = 0.8
THRESHOLD = 0.0
LAYERS = [13, 6, 3]
ACTIVATION_FN = tanh
```

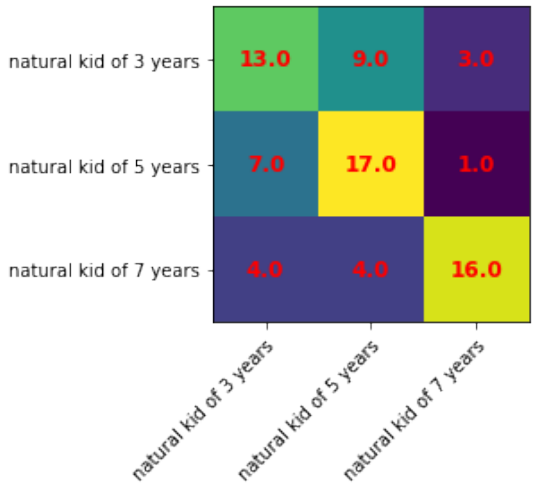

FIGURE 9 – Matrice de confusion des résultats du troisième modèle

4.3 Modèle final et analyse

Pour calculer le modèle final, nous avons donc sélectionné les hyper-paramètres présentés dans le listing 3.

On observe dans la matrice de confusion présentée dans la figure 9 que le modèle a de la difficulté à différencier les différentes classes. On observe, en effet, des f1-score de 0.53 pour la classe « natural kid of 3 years », de 0.62 pour la classe « natural kid of 5 years » et de 0.73 pour la classe « natural kid of 7 years ». Toutefois, ces observations ne sont pas surprenantes, car il est facilement imaginable que des enfants, avec aussi peu d'écart en termes d'âge, aient des voix similaires.

Comme pour l'expérience précédente, en faisant l'addition des observations par classe, on observe que le modèle associe une même valeur à plusieurs classes simultanément. Nous avons essayé de modifier le threshold, mais cela n'a pas donné de résultat concluant.

L'erreur MSE d'entrainement est de 0.58 et celle de test est de 0.76, une erreur étant environ 1.5 fois plus grande.

5 Conclusion

Ce laboratoire, mise en pratique concrète du cours, permet de vraiment faire l'expérience de l'apprentissage par réseau de neurones et du temps que cela peut prendre. Étant donné le nombre restreint de données pour entrainer les différents modèles, les résultats obtenus sont satisfaisants et correspondent aux attentes. L'overfitting observé ainsi que les différences marquées entre courbes MSE d'entrainement et de test sont sans doute dû à ce manque de données.