

Projet 8 : Soutenance

Participez à une compétition Kaggle
SETI Breakthrough Listen – E.T. Signal Search

Gaëtan PELLETIER

Sommaire

- Présentation du projet
- Analyse du jeu de données
- Pré-traitement des données
- Choix d'une architecture de CNN
- Modélisations proposées
- Kernel Kaggle : AutoML
- Limitations du projet
- Synthèse

Projet 8 : Soutenance

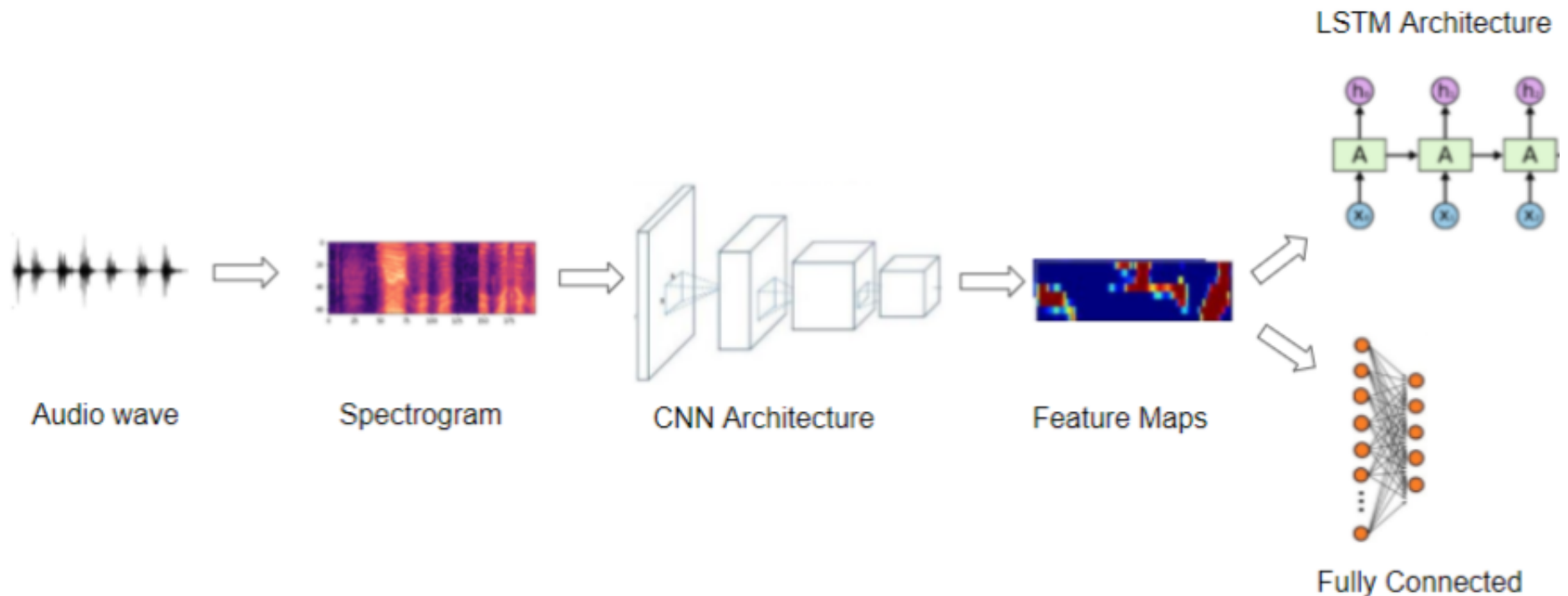
Présentation du projet

Présentation du projet

- Participation à une **compétition** de Data Science
- Plateforme :
 - **Kaggle**
- Compétition choisie :
 - **SETI Breakthrough Listen – E.T. Signal Search**
- Objectif supplémentaire :
 - fournir un **noyau Kaggle** sur un élément intéressant avec la communauté

Présentation du projet

- Les signaux étudiés sont des **ondes**
- Transformée de Fourier → **Spectrogramme**

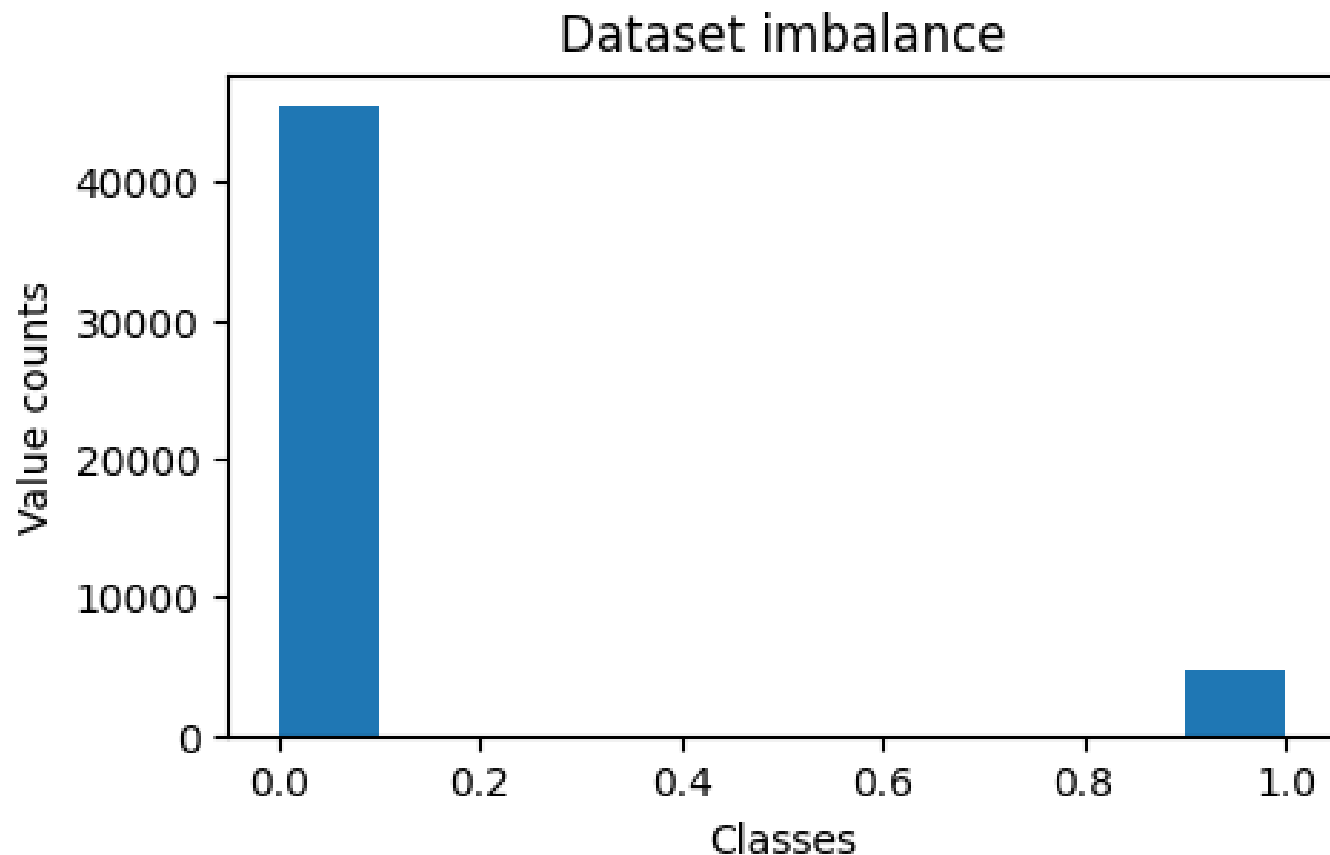


Projet 8 : Soutenance

Analyse du jeu de données

Analyse du jeu de données

Étude du fichier csv contenant id images et **classes**



Analyse du jeu de données

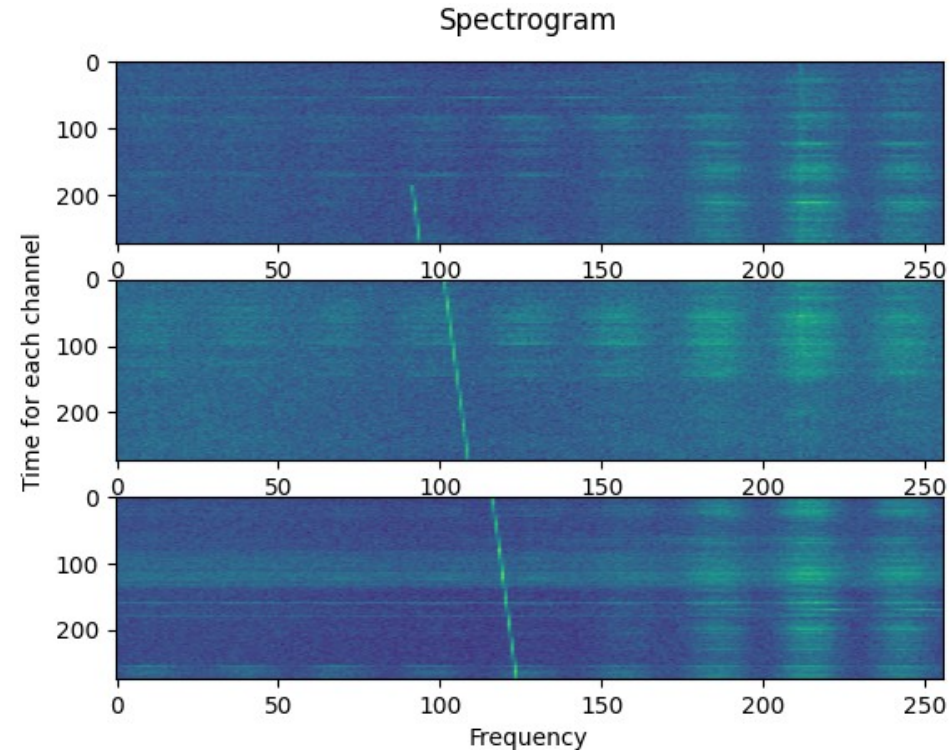
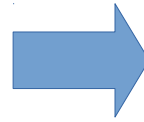
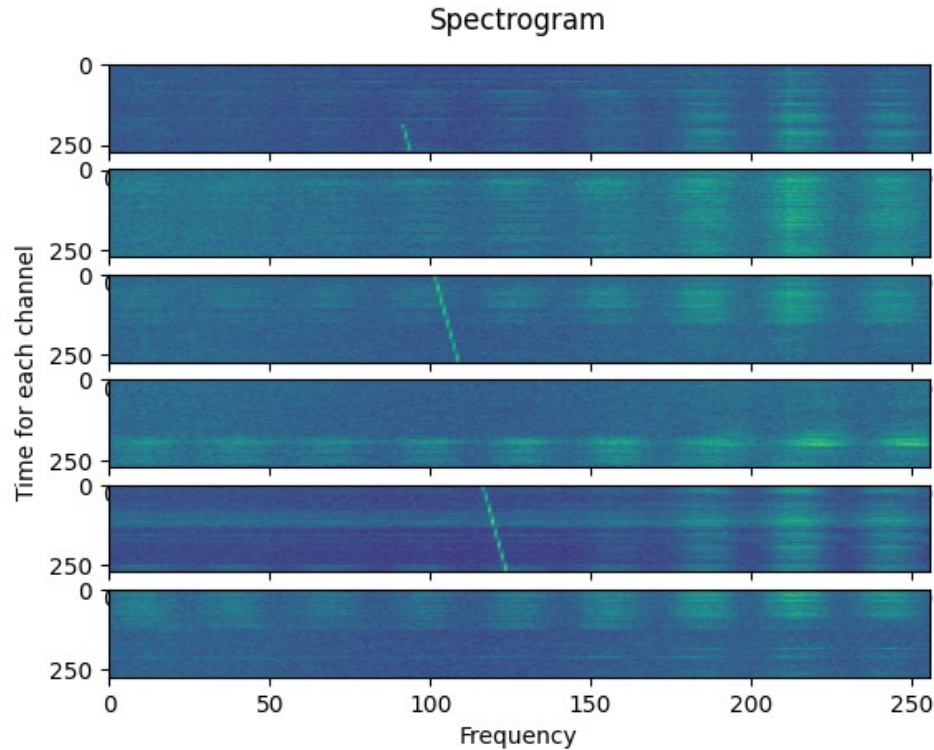
- Déséquilibre du jeu de données
- Calcul de **poids de classes** à appliquer à la perte durant l'entraînement
- Utilisation de la **perte focale**
- Métrique utilisée → **AUROC** (imposée)

Projet 8 : Soutenance

Pré-traitement des données

Pré-traitement des données

Conservation des *on-target* (étoile observée):



Pré-traitement des données

- **Cross validation** :
 - Séparation du jeu de données en 5 plis
 - **Stratification** selon la distribution des classes
- **Redimensionnement** des images en 250x250 pixels
- **Data augmentation** :
 - Variation aléatoire des **couleurs** (*random_hue*)
 - **Symétrie** aléatoire (*random_flip_up_down*)
 - **Mixup**

Pré-traitement des données

- Création de **dataloaders** :
 - **Mélange** (*shuffle*)
 - **Transformations** (*on-target, resize, data augmentation**)
 - **Mini-lots** (*batch*)
 - **Prélecture** (*prefetch*)
- Avec *mixup* :
 - **Deux** dataloaders avec mêmes données du jeu d'entraînement
 - **Mixage** des images
 - **Nouveau dataloader** pour l'entraînement des modèles (contenant images mixées)

* data augmentation seulement pour jeu d'entraînement

Projet 8 : Soutenance

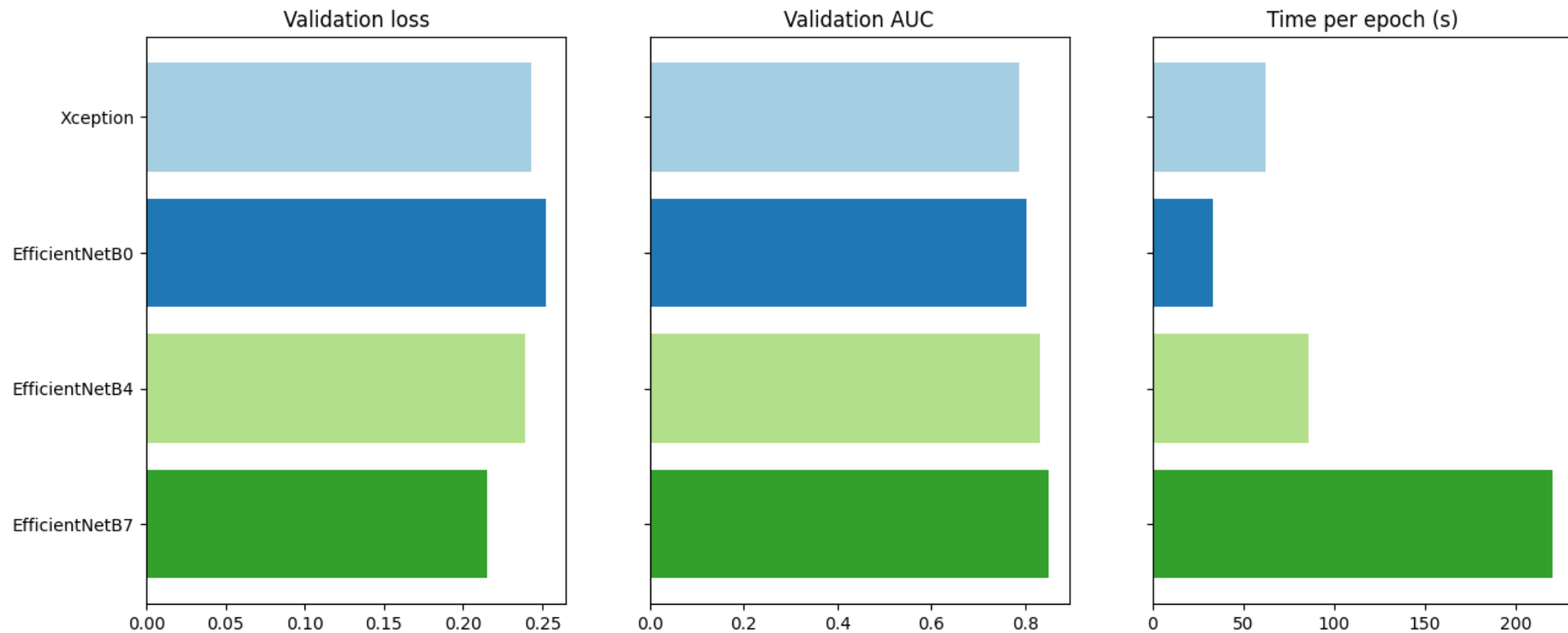
Choix d'une architecture de CNN

Choix d'une architecture de CNN

- Chargement de modèles par *Transfer Learning* :
 - Xception
 - EfficientNet B0
 - EfficientNet B4
 - EfficientNet B7
- Utilisation de **10 %** du jeu de données
- Rappel → *early stopping*
- Métrique → **AUC**

Choix d'une architecture de CNN

Comparaison des modèles :



Choix d'une architecture de CNN

- Modèle retenu → **EfficientNet B4**
- Hyperparamétrage :
 - Optimisation **bayésienne** (bibliothèque *keras-tuner*)
 - **Taux d'extinction** de la couche Dropout
 - **Taux d'apprentissage**
 - Utilisation de **10 %** des données

Projet 8 : Soutenance

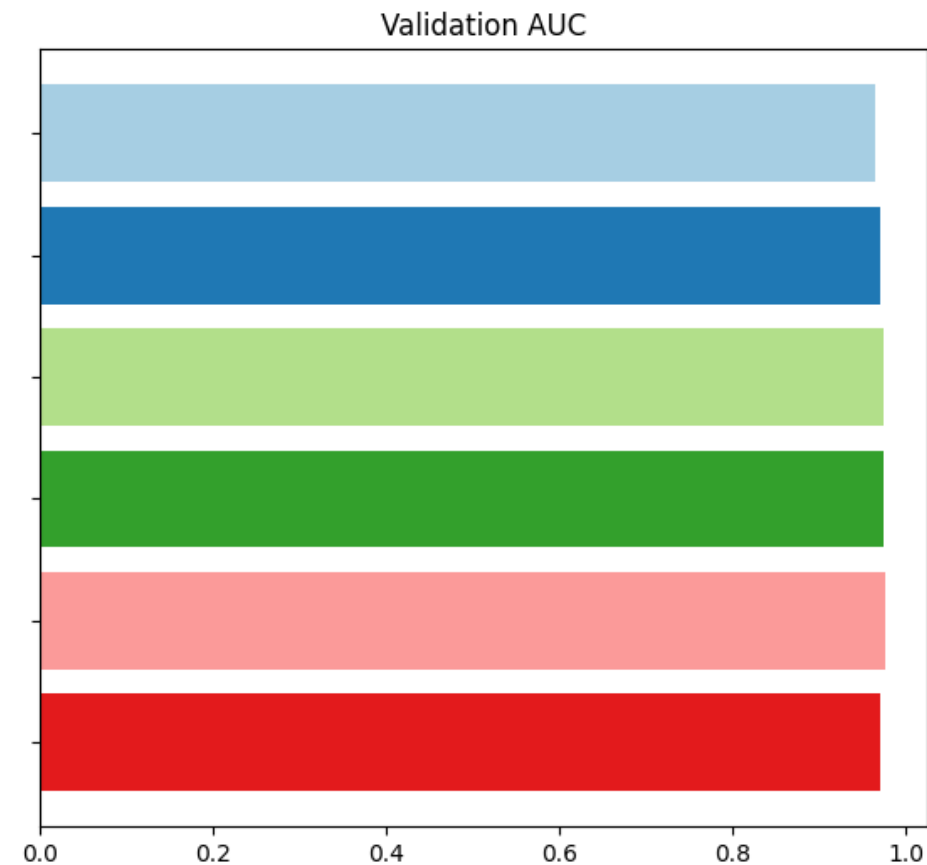
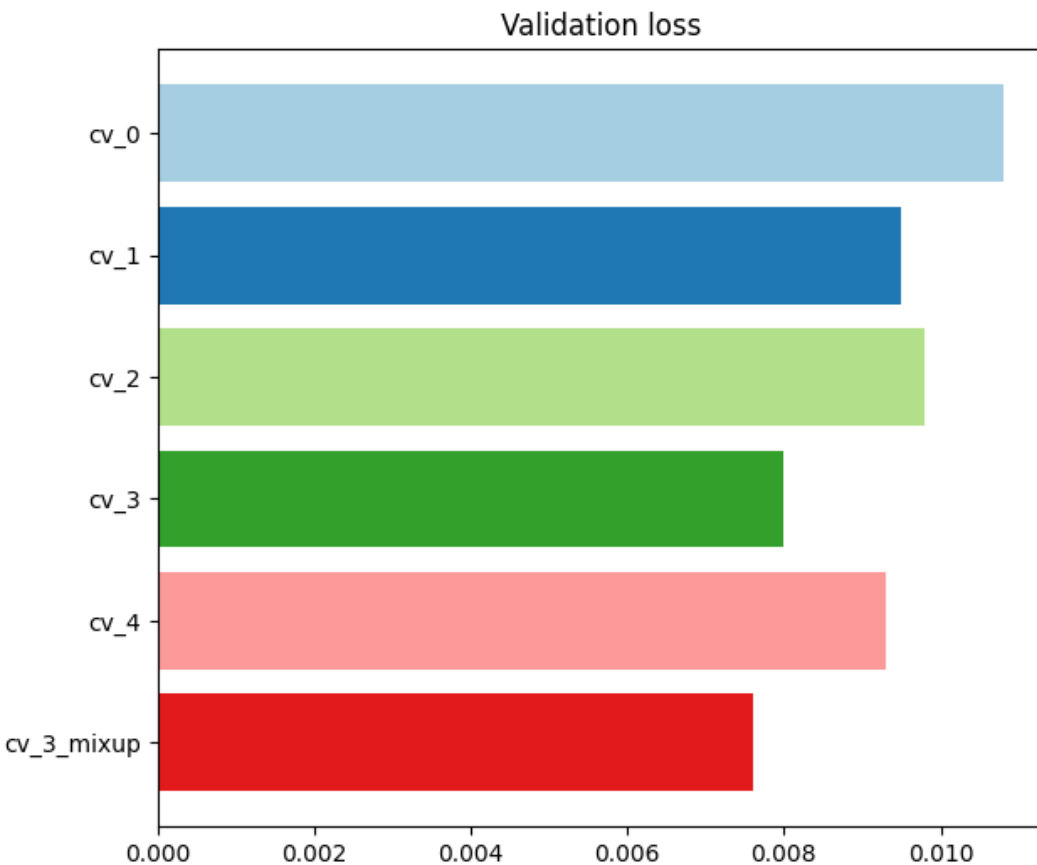
Modélisations proposées

Modélisations proposées

- Modèle optimisé → **EfficientNet B4**
- Utilisation jeu de données **complet**
- Entraînement pour chaque **pli**
- « Meilleur pli » → entraînement avec ***mixup***
- Rappels :
 - Early stopping
 - Échéancier d'apprentissage
 - Points de restauration du modèle (*ModelCheckpoint*)

Modélisations proposées

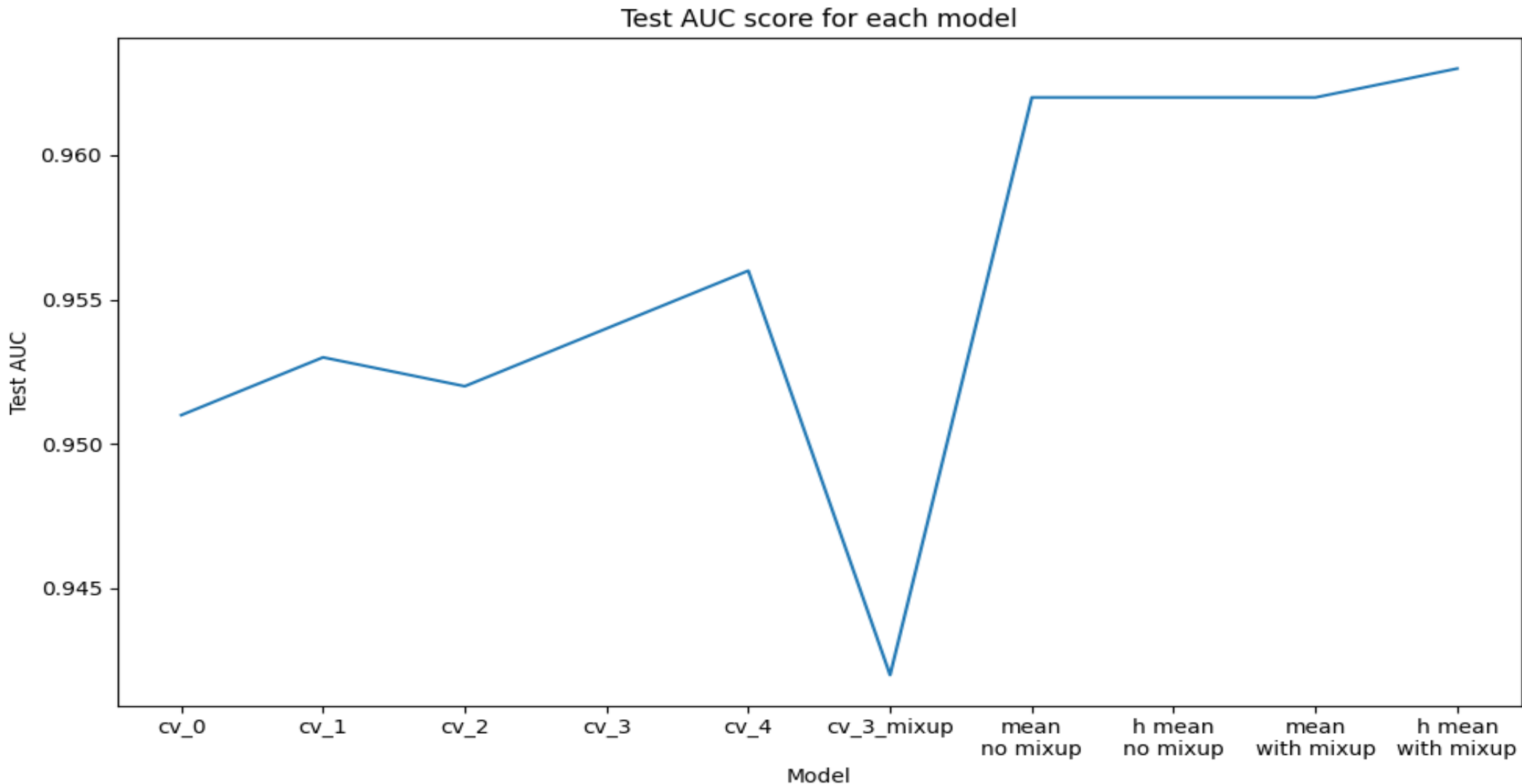
Performances EfficientNet B4 :



Modélisations proposées

- Deux stratégies :
 - Modèles **uniques** (avec ou sans *mixup*)
 - **Plusieurs** modèles :
 - **Moyenne** des probabilités d'appartenance à une classe
 - **Moyenne harmonique**
en fonction de l'**erreur** sur le jeu de validation
- Soumissions des prédictions sur Kaggle
- Obtention d'un score **AUC**

Modélisations proposées



Projet 8 : Soutenance

Kernel Kaggle :
AutoML

Kernel Kaggle : AutoML

- AutoML utilisé → **AutoKeras**
- Principal avantage → même **structure** que Keras
- Entraînement de **10** modèles (**10 %** des données)
- Meilleur modèle **moins bon** que précédents CNN :

Modèles	Xception	EFN B0	EFN B4	EFN B7	AutoKeras
Val loss	0,2432	0,2523	0,2395	0,2151	0,3722
Val AUC	0,7889	0,8038	0,8323	0,8511	0,4758

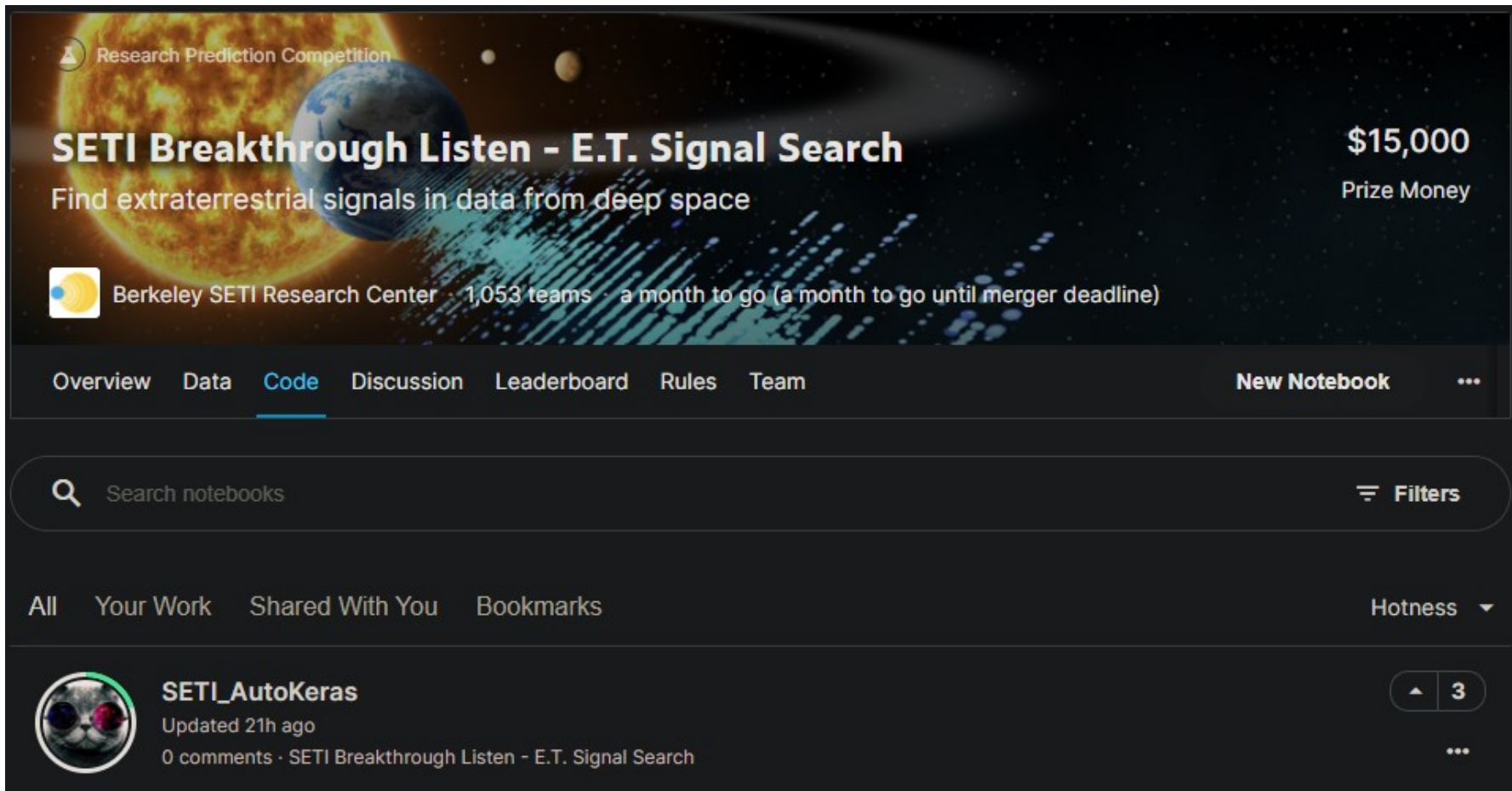
Kernel Kaggle : AutoML

Améliorations possibles :

- Jeu de données **complet**
- Data augmentation
- Entraîner **plus** de modèles
- **Imposer** architectures à utiliser
 - éviter d'entraîner modèles trop simples
 - gain de temps

Kernel Kaggle : AutoML

Partage du kernel avec la communauté Kaggle



The screenshot displays the Kaggle interface for the "SETI Breakthrough Listen - E.T. Signal Search" competition. The header features a cosmic background with a large orange sun, a blue Earth, and a comet. The competition title "SETI Breakthrough Listen - E.T. Signal Search" is prominently displayed, along with the subtitle "Find extraterrestrial signals in data from deep space". A prize money of "\$15,000" is shown in the top right. Below the title, the organizing institution "Berkeley SETI Research Center" is listed, along with "1,053 teams" and a note about the "a month to go (a month to go until merger deadline)".

The navigation bar includes tabs for "Overview", "Data", "Code" (which is selected and underlined), "Discussion", "Leaderboard", "Rules", and "Team". A "New Notebook" button is located on the right side of the navigation bar. Below the navigation bar, there is a search bar labeled "Search notebooks" and a "Filters" button. The main content area shows a list of notebooks, with the first one being "SETI_AutoKeras" by a user with a cat profile picture. This notebook was "Updated 21h ago" and has "0 comments". The notebook is associated with the "SETI Breakthrough Listen - E.T. Signal Search" competition. A "Hotness" dropdown menu is visible on the right side of the notebook list.

Projet 8 : Soutenance

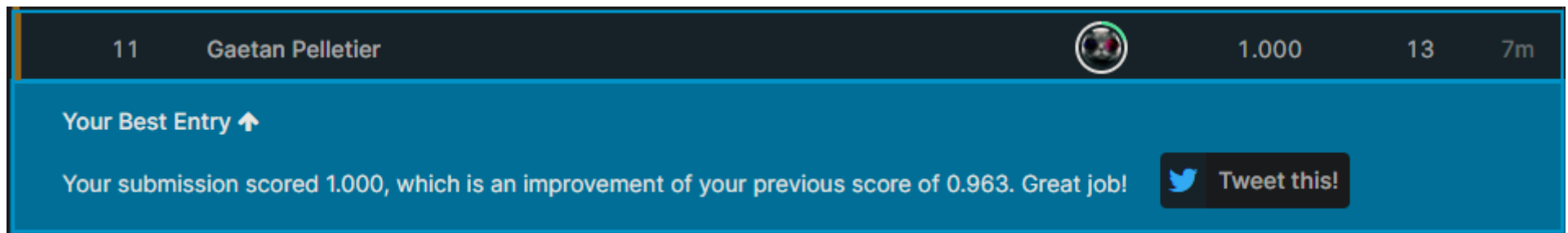
Limitations du projet

Limitations du projet

- Volume du jeu de données (plus de **60 Go**)
- **Quota** utilisation GPU : **30h/semaine**
- Nombre d'heures du projet fixé par OC : **50 h**
- Avec plus de temps, il serait possible :
 - d'utiliser **plus de données** (4 plis au lieu de 5)
 - de tester **différents dimensionnements** d'images
 - de tester plusieurs configurations de data augmentation
 - d'entraîner des modèles **plus complexes** (e.g. EFN B7)

Limitations du projet


- Time Stamp Leakage :
 - Téléchargement des données en local
 - Obtention date de création des spectrogrammes
 - Date non remise à zéro par Kaggle
 - Possibilité de **lier** cette **date** à la **classe** de l'image
- Obtention score AUC parfait avec AutoGluon :



Limitations du projet

Probable **fuite** d'informations au sein des **images** :

- **Pré-traitement** des **images** permettrait de trouver celles **créées par** le centre **SETI** (classe 1)
- En cours d'investigation par la communauté Kaggle




nyanp
Topic Author
195th place

Yet another leakage? (LB 0.995 without timestamp)

Posted in [seti-breakthrough-listen](#) 4 days ago

This leak probably indicates a problem with the data generation process in the simulation.

Some of the models of the top participants may already implicitly take advantage of this distributional difference, but I hope that this leak will be fixed along with the timestamp leak.



57

Projet 8 : Soutenance

Synthèse

Synthèse

- Étude du **déséquilibre** du dataset
- **Pré-traitement** des spectrogrammes
- Recherche **meilleure architecture** de modèles
- Optimisation du modèle **EfficientNet B4**
- **Prédictions** par :
 - Modèles uniques
 - Moyennes (harmoniques) → **AUC** test set **0,963**
- **Kernel Kaggle** pour la communauté : **AutoKeras**
- **Limitations** du projet :
 - **Temps** (alloué au projet, quota GPU)
 - **Data leakage** → remise à zéro de la compétition

Projet 8 : Soutenance

Merci de votre attention

Projet 8 : Soutenance

Annexe

Annexe 1

Mixup :

- Mixer deux images, ainsi que leur classe
- Modèle perd en confiance absolue
- Améliore la **généralisation** d'un modèle
- Technique efficace quand **peu de connaissances sur le problème étudiée** et les augmentations possibles

