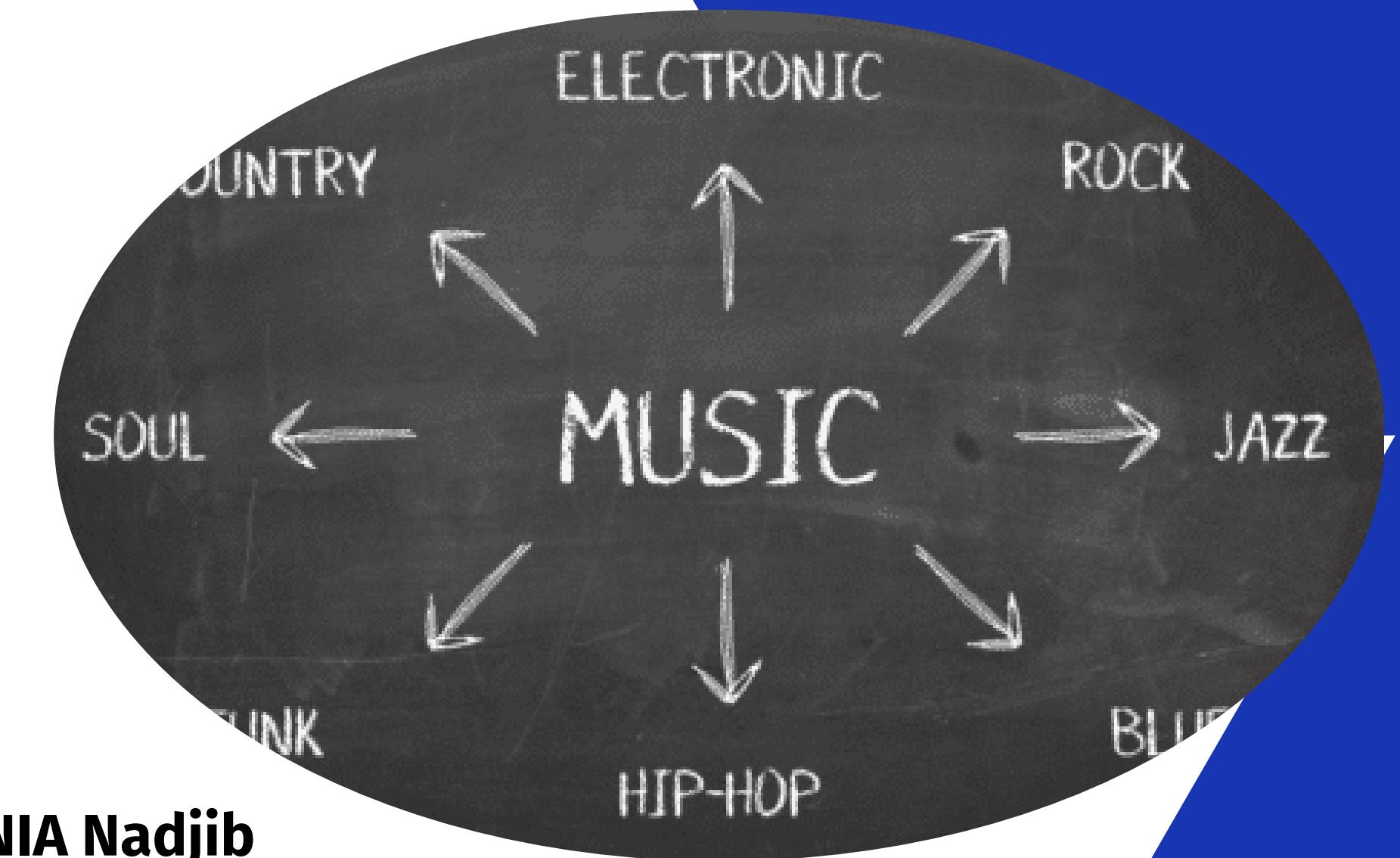


Thème:
**CLASSIFICATION DE
GENRES MUSICAUX
AVEC WAV2VEC 2.0**

Présenté par : **BOUTALBI Mohammed Iliass & ATAMNIA Nadjib**



Plan

Contexte et Problématique

Objectifs du Projet

Caractéristiques principales du dataset GTZAN

Architecture CNN Baseline

Résultats CNN Baseline

Architecture Wav2Vec 2.0

Fine-tuning Complet Wav2Vec 2.0

Fine-tuning Tête Seule

Comparaison des Trois Approches

Conclusion

Contexte et Problématique



Contexte général

L'explosion de la musique numérique via les plateformes de streaming nécessite des outils d'analyse automatique pour organiser et explorer les contenus audio.

La classification par genre facilite la recherche, améliore les recommandations et structure les bibliothèques musicales.

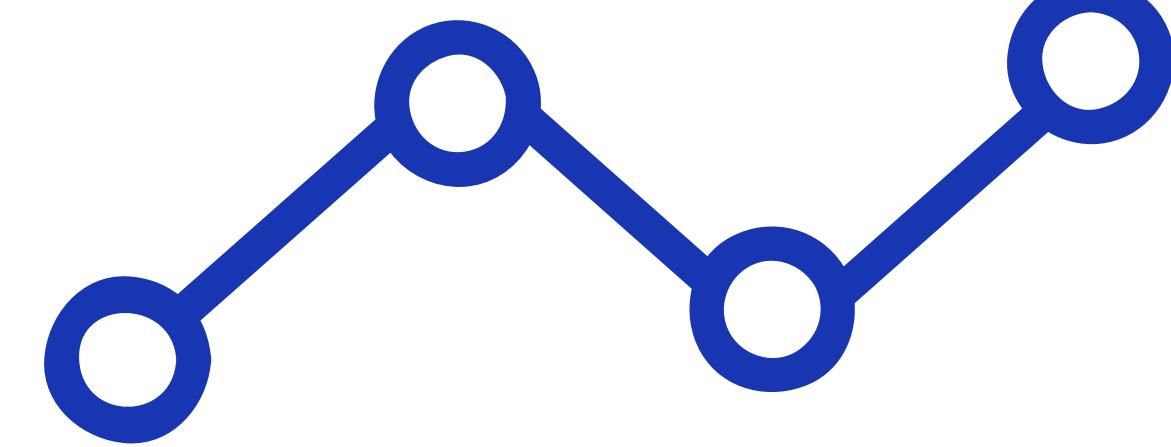
Problématique

La classification automatique reste complexe : frontières floues entre genres, variabilité intra-genre importante, et facteurs techniques (qualité, mixage, bruit).

Les approches traditionnelles (MFCC, SVM) atteignent leurs limites face à cette diversité.



Objectifs du Projet



01

02

03

04

Approche de référence

Identifier et implémenter une baseline CNN sur spectrogrammes depuis Kaggle pour GTZAN.

Wav2Vec 2.0

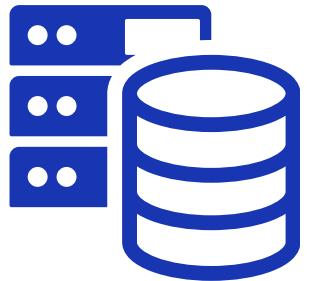
Mesurer les performances de l'approche CNN classique sur le dataset GTZAN

Évaluation baseline

Entraîner un modèle auto-supervisé Wav2Vec 2.0 avec Fairseq sur audio brut

Comparaison

Analyser quantitativement et qualitativement les avantages de l'apprentissage auto-supervisé.



Caractéristiques principales du dataset GTZAN

1000

Extraits audio

Fichiers de 30 secondes au format .wav, échantillonnés à 22050 Hz

10

Genres musicaux

Blues, Classical, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock

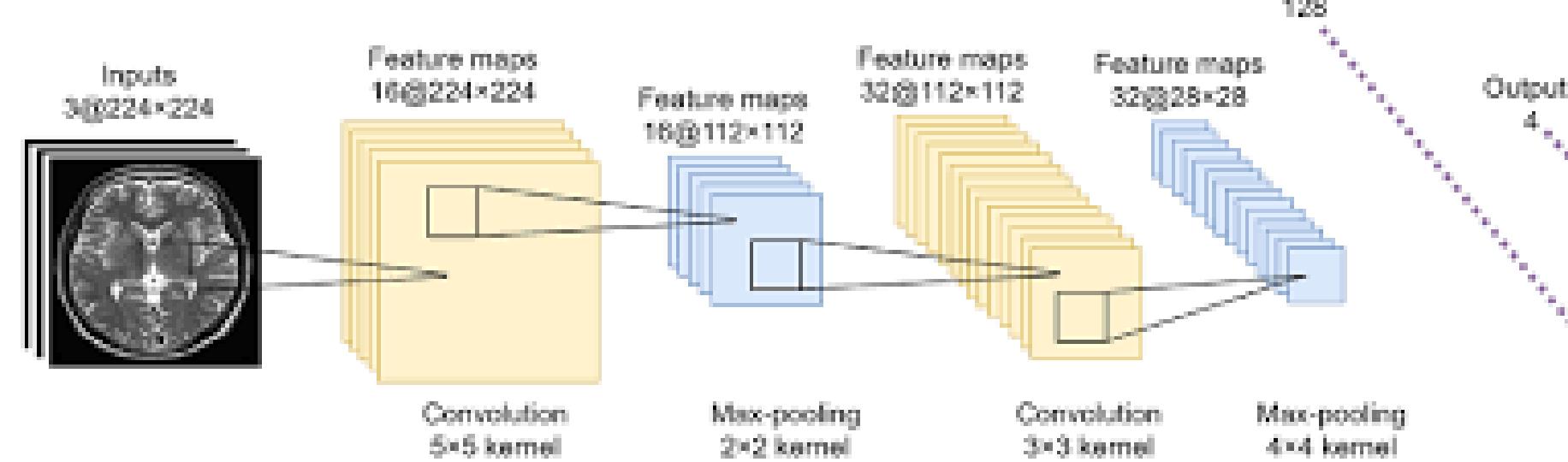
100

Fichiers par genre

Distribution équilibrée garantissant l'absence de biais de classe



Architecture CNN Baseline



01 Entrée : Mél-spectrogrammes

Signal audio -> spectrogramme 128 bandes ->
normalisation [0,1] -> tenseur $[1 \times 128 \times T]$

02 4 Blocs Convolutionnels

Conv2D (1 -> 32 -> 64 -> 128 -> 256) + BatchNorm + ReLU +
MaxPool + Dropout (0.25)

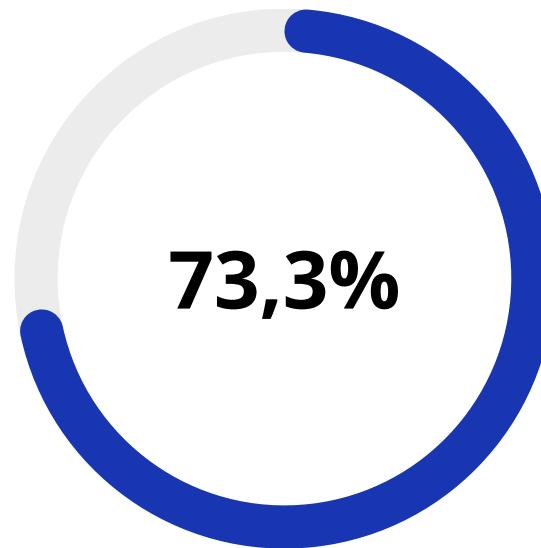
03 Global Average Pooling

Réduction à un vecteur de dimension 256, indépendant de la
longueur

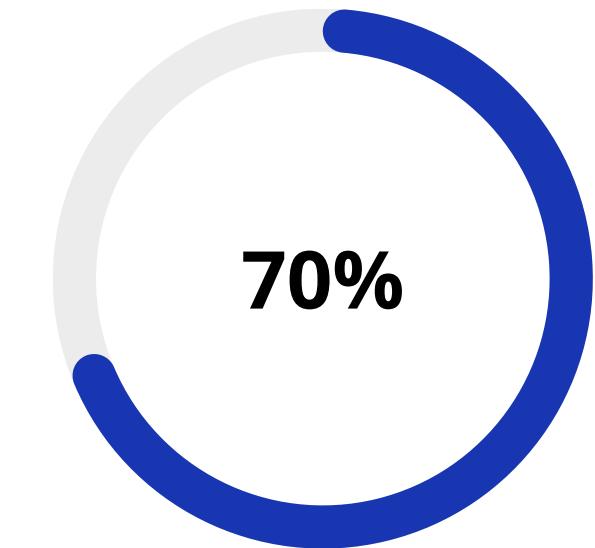
04 Tête de classification

FC 256 -> 512 -> 256 -> 10 avec ReLU et Dropout (0.5) pour
prédir les 10 genres

Résultats CNN Baseline



Accuracy test



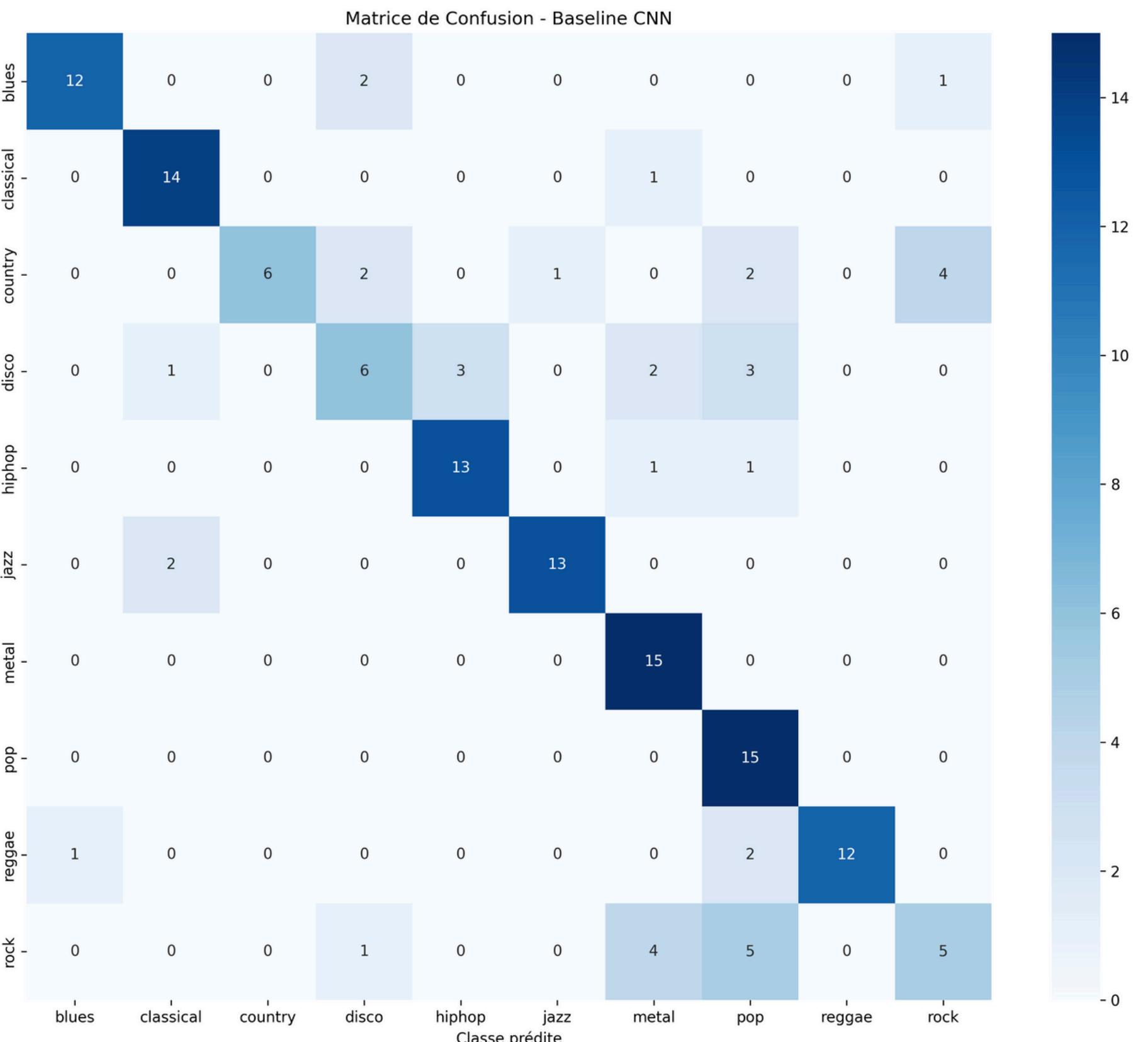
F1-score macro

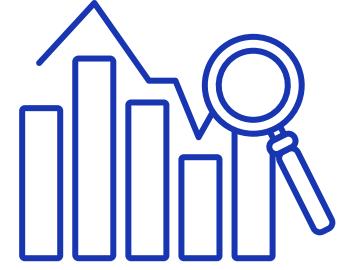
Genres bien reconnus :

- Classical, Hiphop, Metal : $F1 > 0.8$
- Reggae : excellente séparation

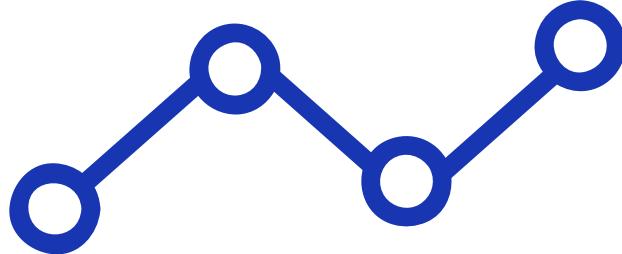
Genres difficiles :

- Disco : $F1 = 0.33$
- Rock : $F1 = 0.30$
- Confusions fréquentes entre genres proches





Architecture Wav2Vec 2.0



Encodeur CNN 1D

Transforme le signal brut en séquence de vecteurs compacts (représentation temps-fréquence apprise)

Masquage latent

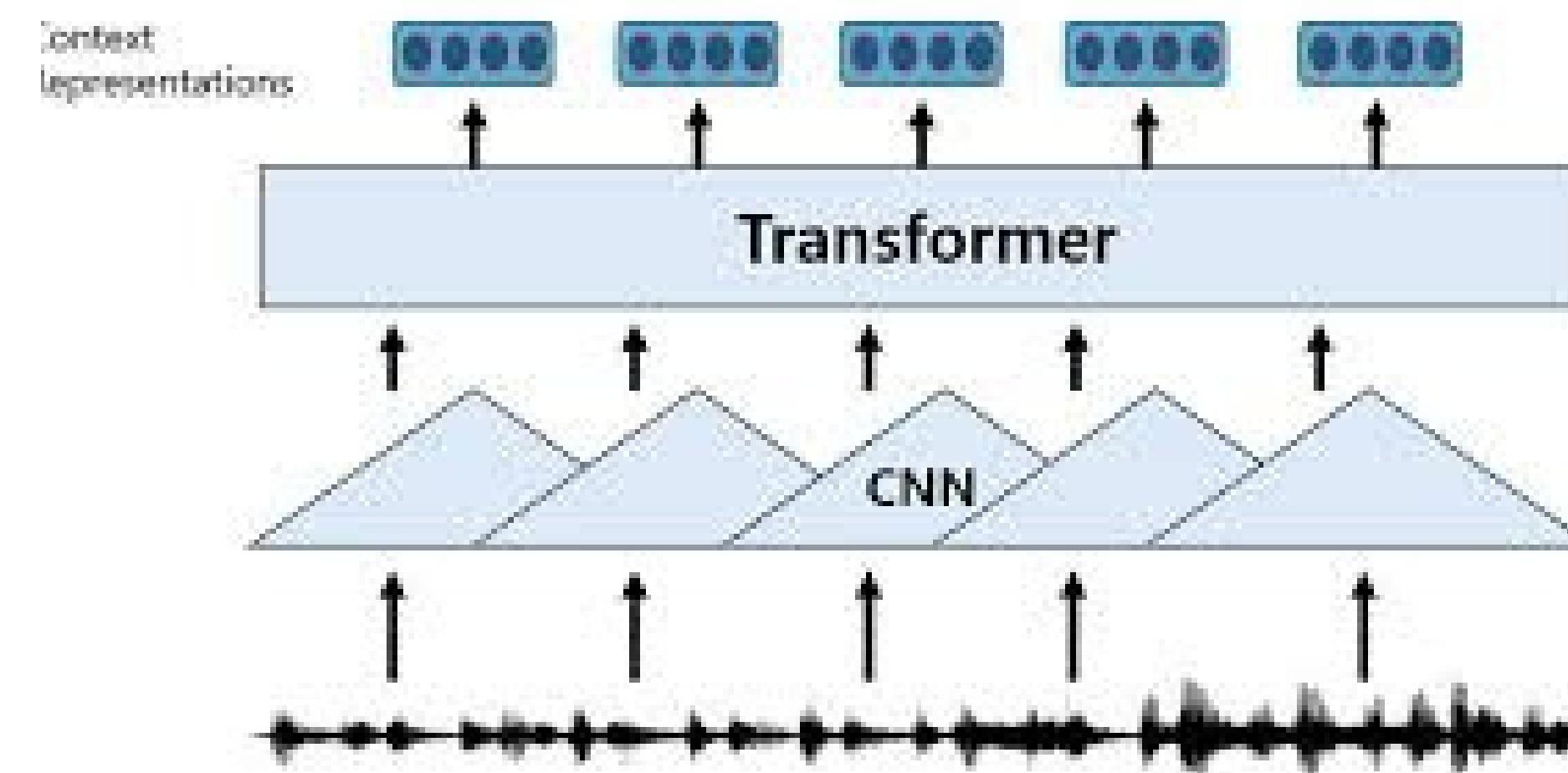
Segments masqués aléatoirement pour apprentissage auto supervisé par prédiction contextuelle

Transformers

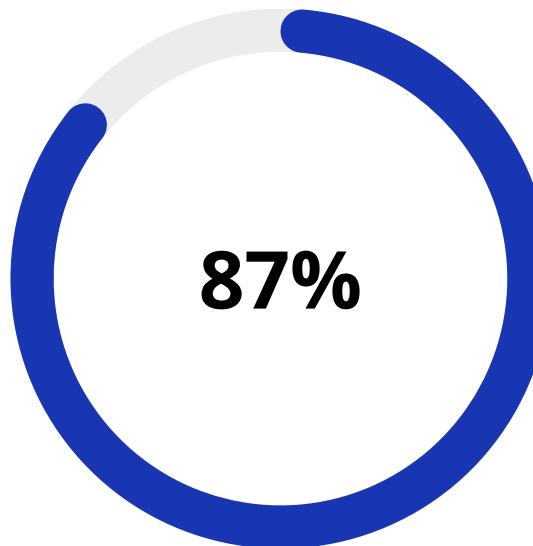
Blocs d'attention multi-tête modélisant les dépendances temporelles de long terme

Objectif contrastif

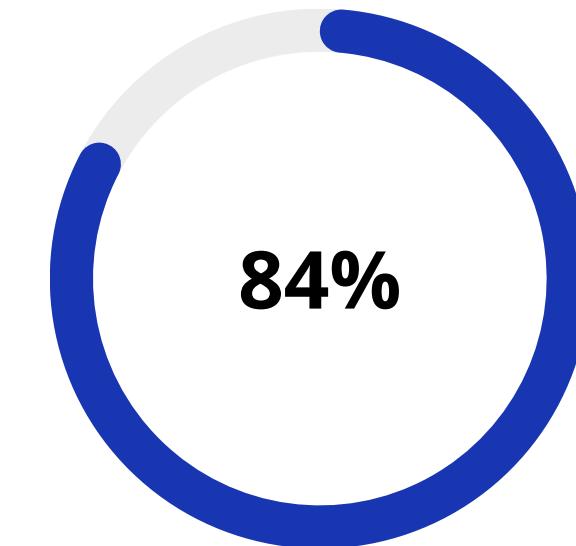
Quantification + loss contrastive rapprochant vraies cibles et éloignant négatifs



Fine-tuning Complet Wav2Vec 2.0



Accuracy test



F1-score macro

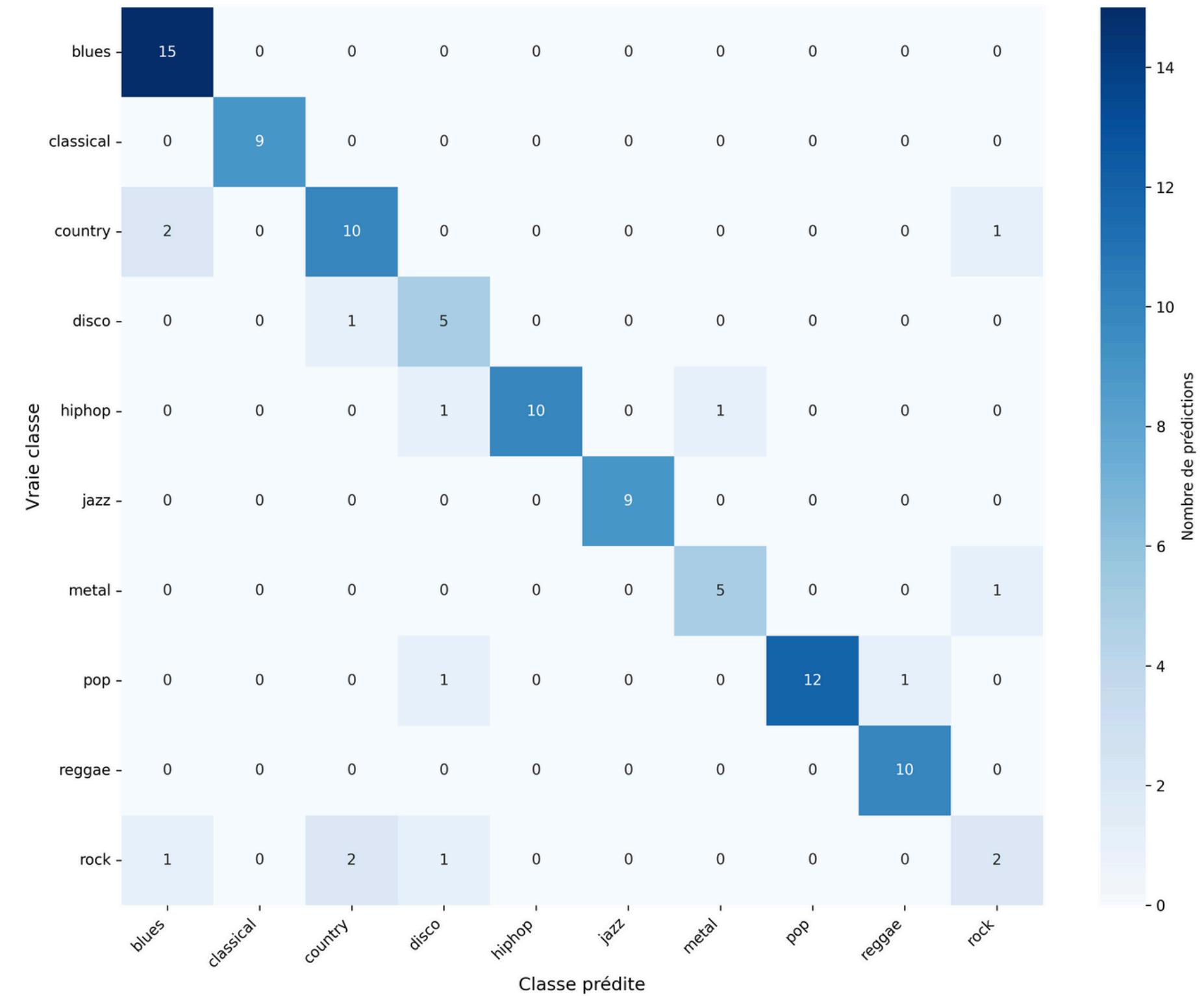
**Excellent
résultats**

- Classical, Jazz : F1 = 1.00
- Reggae : F1 = 0.95
- Pop, Blues, Hiphop : F1 > 0.90

**Performances
intermédiaires**

- Metal : F1 = 0.83
- Country : F1 = 0.77
- Disco : F1 = 0.71

Matrice de Confusion - test Set



Fine-tuning Tête Seule

Accuracy test

75%

F1-score macro

72%

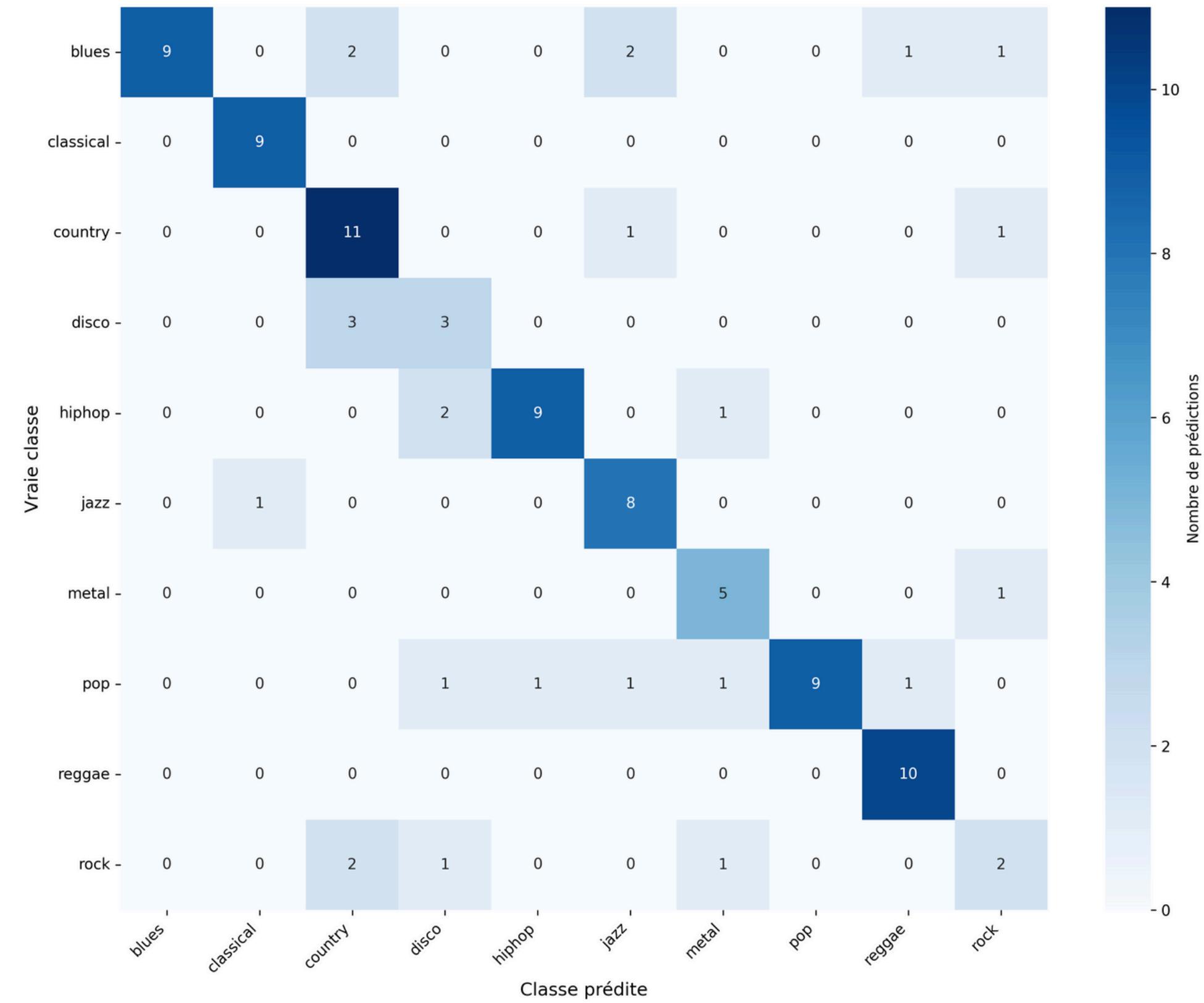
Points forts

- Classical (F1 = 0.95)
- Reggae (F1 = 0.91)
- Hiphop (F1 = 0.82)

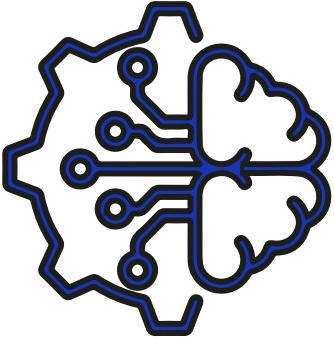
Limites

- Rock (F1 = 0.36)
- Disco (F1 = 0.46)

Matrice de Confusion - Test Set



10



Comparaison des Trois Approches

Hyperparamètre	Valeur choisie	Accuracy	F1 macro
CNN baseline	Spectrogrammes	73.33%	0.7
Wav2Vec 3 tête seule	Audio brut	75%	0.72
Wav2Vec 3 complet	Audio brut	87%	0.84

Apport décisif

L'apprentissage auto-supervisé sur audio brut surpassé nettement les approches traditionnelles sur spectrogrammes

Fine-tuning optimal

L'adaptation complète du backbone maximise les performances, particulièrement sur genres ambigus

Compromis efficace

La stratégie tête seule offre un gain rapide avec coût réduit, idéale pour prototypage

Conclusion



L'apprentissage auto-supervisé avec **Wav2Vec 2.0** surpassé nettement les approches classiques. **Le fine-tuning complet** exploite pleinement les représentations audio brutes pour une classification optimale des genres musicaux.

Merci de votre
attention !