Détection de genre musical à l'aide de technique de Machine Learning

Gabriel Lucchini, Nicolas Pluven, Malek Bennabi

Professeur : Nizar Ouarti
git : https://github.com/Malekbennabi3/Song-Classification

1. Introduction

La classification automatique des genres musicaux est un défi important dans le domaine de la reconnaissance de contenu audio, avec des applications dans la recommandation musicale, les systèmes de diffusion en continu, et l'analyse de grandes bases de données musicales. Avec l'explosion des plateformes de streaming et la diversité des styles musicaux disponibles, il devient crucial de développer des outils capables d'organiser et de catégoriser efficacement les morceaux de musique.

Dans ce contexte, notre projet s'inscrit dans une démarche visant à exploiter les avancées des techniques de Machine Learning pour classifier automatiquement les morceaux de musique selon leur genre. L'objectif principal est de concevoir un système capable d'identifier le genre musical parmi 20 catégories définies à partir d'enregistrements audio. Cette tâche présente plusieurs défis, notamment la gestion de grands volumes de données, le déséquilibre dans la distribution des genres, et la complexité inhérente des représentations musicales.

Pour atteindre cet objectif, nous avons adopté une méthodologie reposant sur différentes étapes, incluant le prétraitement des données audio, la conception et l'évaluation de plusieurs modèles de classification, allant des approches traditionnelles (SVM) aux architectures modernes de Deep Learning (CNN, EfficientNet, Transformers). Ce rapport présente les différentes étapes de notre travail, les résultats obtenus, les limitations rencontrées, ainsi que les pistes d'amélioration envisagées pour optimiser la classification des genres musicaux.

2. Données

Le projet s'appuie sur le dataset public disponible sur Hugging Face, intitulé "lewtun/music genres". Ce dataset contient un total de 25 000 enregistrements audio représentant environ 9,8 Go de données, avec une durée moyenne de 30 secondes par fichier audio. Chaque enregistrement est associé à un genre musical parmi 20 catégories, telles que le rock, le jazz, le hip-hop ou encore le blues. Les données sont fournies au format .parquet avec les colonnes suivantes :

- **Bytes** : représentation des données audio.
- **Song_id** : identifiant unique pour chaque enregistrement.
- **Genre_id** : identifiant numérique du genre musical.
- **Genre** : nom du genre musical.

Une analyse préliminaire de la distribution 1 des genres révèle un déséquilibre important dans le dataset. Certains genres, comme le rock et le hip-hop, sont largement représentés, tandis que d'autres, comme le jazz ou le blues, sont sous-représentés. Ce déséquilibre introduit un défi significatif pour les modèles de classification, susceptibles de biaiser les prédictions vers les genres majoritaires.

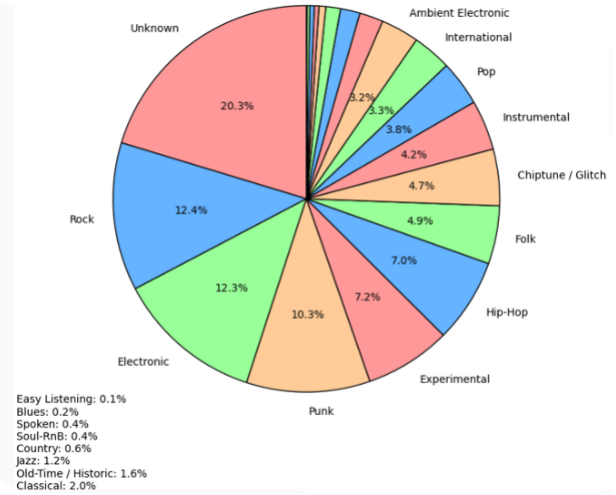

Figure 1. répartition dataset

Pour exploiter pleinement les enregistrements audio, une conversion des données brutes en spectrogrammes MEL a été effectuée. Le spectrogramme MEL est une représentation visuelle de l'énergie spectrale d'un signal audio dans le domaine des fréquences, ajustée selon l'échelle de Mel. Cette transformation est choisie car elle reflète mieux la manière dont l'oreille humaine perçoit les sons, facilitant ainsi l'apprentissage par les modèles de Machine Learning. Le pipeline de prétraitement suit les étapes suivantes :

1. Conversion des fichiers audio au format *.wav*.
2. Génération des spectrogrammes MEL pour capturer les caractéristiques audio pertinentes.
3. Division des données en ensembles d'entraînement, de validation et de test pour garantir une évaluation robuste des modèles.

Lors de l'analyse et du traitement des données, plusieurs défis ont été identifiés, certains fichiers .wav étaient endommagés ou inexploitables, nécessitant leur exclusion du pipeline. Aussi, les genres sous-représentés ont complexifié l'entraînement des modèles, nécessitant des stratégies comme la pondération des classes ou l'augmentation des données. Enfin, certaines chansons pourraient appartenir à plusieurs genres, rendant la classification plus difficile. Une évaluation basée sur le Top-3 ou Top-5 pourrait être envisagée pour pallier cette limite.

En dépit de ces défis, le dataset offre une base solide pour tester et comparer différentes approches de classification, en mettant en lumière les performances et les limites des modèles étudiés.

3. Méthodologie

3.1. Prétraitement des données

Le prétraitement des données est une étape cruciale dans la préparation des enregistrements audio pour la classification des genres musicaux. Cette étape vise à transformer les données brutes en représentations exploitables par les modèles de Machine Learning tout en réduisant le bruit et les redondances.

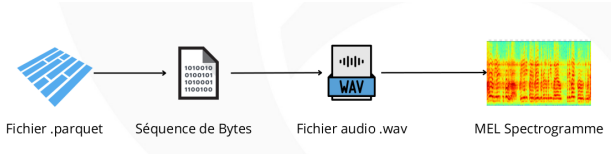

Figure 2. Pipeline du prétraitement

Les enregistrements audio fournis au format *.parquet* ont été convertis en fichiers de séquence de bytes puis en fichiers audio *.wav*. Cette conversion permet de manipuler plus facilement les données audio à l'aide de bibliothèques spécialisées dans le traitement des signaux.

Les fichiers audio *.wav* ont ensuite été transformés en spectrogrammes MEL, une représentation visuelle de l'énergie spectrale d'un signal audio dans le domaine des fréquences.

Un contrôle de qualité a été effectué pour identifier et exclure les fichiers *.wav* endommagés ou inexploités afin de ne pas biaiser l'entraînement des modèles.

3.2. Modèles étudiés

Pour répondre à l'objectif de classification des genres musicaux, plusieurs modèles de Machine Learning et de Deep Learning ont été implémentés et évalués. Chaque modèle a été testé sur les données prétraitées pour mesurer son efficacité en termes de précision et de performance globale.

3.2.1. Support Vector Machines (SVM)

Le SVM, une méthode classique de Machine Learning, a été utilisé comme point de départ pour évaluer les performances sur un problème multi-classe.

1. Prétraitement des données en spectrogrammes.
2. Division des données en ensembles d'entraînement (70%) et de test (30%) avec validation croisée.
3. Ajustement des hyperparamètres (niveau de régularisation, noyau RBF).
4. Évaluation des performances sur l'ensemble de test.

On obtient une précision de 41%

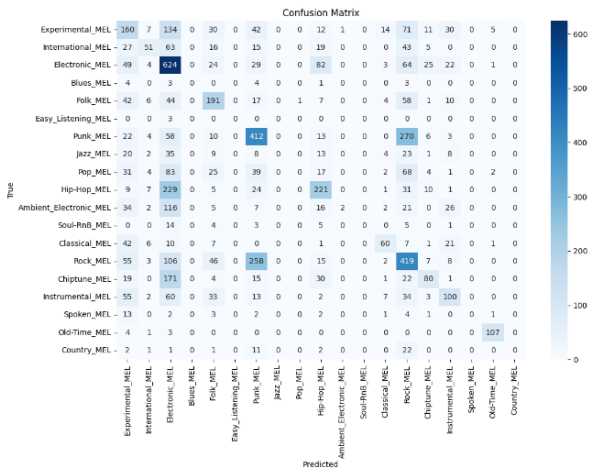

Figure 3. Matrice de confusion du SVM

On observe que certaines classes, comme Electronic_MEL (valeur diagonale de 624) et Folk_MEL (191), montrent des performances élevées en termes de classification correcte (valeurs élevées sur la diagonale).

Certaines classes semblent fortement attirées par d'autres. Par exemple, Rock_MEL est confondu avec Punk_MEL (270 fois), probablement à cause de similitudes stylistiques.

Au final, certaines classes comme Electronic_MEL et Classical_MEL montrent une forte précision. Des genres comme Folk_MEL, Punk_MEL, et Experimental_MEL souffrent de confusions importantes avec d'autres genres.

3.2.2. Convolutional Neural Network (CNN) From Scratch

Un modèle CNN simple a été conçu pour classer les spectrogrammes directement en images. Chaque couche convolutive est suivie de couches de MaxPooling et Dropout pour réduire le sur-apprentissage.

Table 1. Architechture pour le CNN from scratch

Couches	Valeurs
Couche d'entrée : Images RGB	128x128
Couche convolutive 1	32 filtres
Couche convolutive 2	64 filtres
Couche convolutive 3	128 filtres
Couche dense	128 neurones + ReLU
Couche dense lunaire	20 neurones + Activation softmax

On obtient une précision de 36.8%. Nous avons aussi calculé la précision top-5 à 85.91%.

3.2.3. EfficientNet (CNN)

EfficientNet, un modèle préentraîné, a été utilisé pour exploiter sa capacité à extraire des caractéristiques complexes de manière efficace.

On augmente les données et on normalise les Mel. On fait une division ensemble de train/validation/test à 70/15/15. Puis on fait le chargement d'un modèle préentraîné EfficientNet, ajusté sur les données de classification.

Table 2. Parametres pour EfficientNet

Parametre	Valeur
Taille des images	128x128
Taille des batchs	16
Optimiseur	Adam (lr=1e-4)
Epoch	10

On obtient une précision de 51%.

On observe que les genres tels que Electronic_MEL, Hip-Hop_MEL, Punk_MEL, et Rock_MEL montrent de bonnes performances de classification avec des valeurs diagonales élevées.

Des confusions significatives existent pour des genres comme Experimental_MEL, Folk_MEL, et Instrumental_MEL, qui semblent se chevaucher avec d'autres genres en termes de caractéristiques.

Les genres Blues_MEL, Spoken_MEL, et Jazz_MEL sont difficilement identifiables, ce qui suggère des problèmes dans la quantité ou la qualité des données.

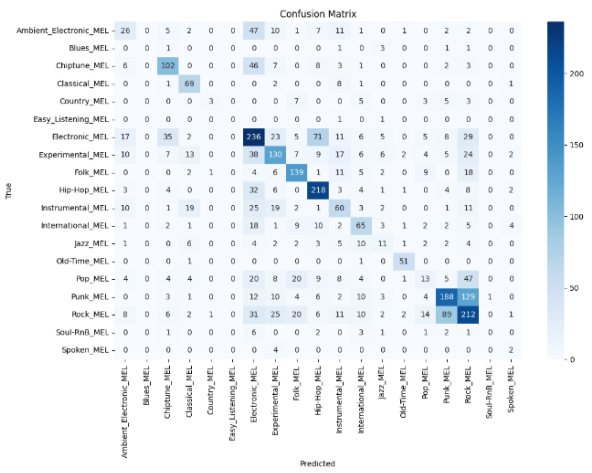

Figure 4. Matrice de confusion du EfficientNet

3.2.4. ViTb16 (Vision Transformer)

Les Transformers, conçus initialement pour le traitement du langage naturel, ont été adaptés pour la classification des images grâce à l'architecture ViTb16.

On augmente les données et on normalise les Mel. On fait une division ensemble de train/validation/test à 70/15/15. Puis on fait le chargement d'un modèle préentraîné ViTb16 et on ajoute les hyperparamètres et entraîne sur 250 epochs. Pour le ViT, on prend des images 256*256 et des batchs de 64.

On obtient une précision de 62%.

4. Résultats et Analyse

L'évaluation des modèles étudiés a permis de comparer leurs performances respectives pour la classification des genres musicaux. Les résultats obtenus mettent en évidence les forces et les faiblesses de chaque approche en termes de précision, de temps d'entraînement, et de capacité à gérer les limites des données.

Table 3. Comparaison des résultats des modèles

Modèle	Précision top-1	Epoch optimale
SVM	41%	-
CNN (From Scratch)	36.8%	20
EfficientNet	51%	10
ViTb16	62%	150

Le SVM est rapide (entraîné en 3min) et peu gourmand en ressources, ce qui en fait une solution adaptée pour des expérimentations rapides. Sa précision de 41% est relativement faible en raison de sa difficulté à gérer des données volumineuses et un problème multi-classe complexe.

Le CNN from scratch (entraîné en 33min) a une précision Top-5 qui atteind 85.91% mais la précision top-1 reste limitée à 36.8%. Cela suggère que le modèle capture certaines caractéristiques pertinentes mais n'est pas assez robuste pour des prédictions précises.

Avec une précision de 51%, EfficientNet (entraîné en 1h12) montre une nette amélioration par rapport aux approches précédentes. Sa conception préentraînée permet de tirer parti des caractéristiques complexes sans entraîner un modèle entièrement à partir de zéro. Les genres peu représentés restent mal classifiés, et le modèle nécessite des ressources matérielles importantes.

Le modèle Vision Transformer (entraîné en 4h23) affiche la meilleure précision à 62%. Sa capacité à capturer des relations complexes entre les caractéristiques des spectrogrammes en fait l'approche la plus performante du projet. Les temps d'entraînement sont significativement plus longs, ce qui le rend moins accessible pour des configurations matérielles modestes.

5. Défis rencontrés

Lors de la réalisation de ce projet, plusieurs défis ont été identifiés, principalement liés à la nature des données utilisées et à la complexité de la tâche de classification.

La distribution des genres musicaux dans le dataset est fortement déséquilibrée. Certains genres, tels que le rock et le hip-hop, sont sur-représentés, tandis que d'autres, comme le jazz et le blues, sont sous-représentés. Cette inégalité a entraîné une tendance des modèles à favoriser les genres majoritaires, réduisant ainsi leur capacité à bien classifier les genres moins fréquents. Pour y remédier, des techniques comme la pondération des classes ou l'augmentation des données pourraient être envisagées.

Une même musique peut parfois correspondre à plusieurs genres, ce qui complique la classification stricte dans une seule catégorie. Par exemple, une chanson pourrait être à la fois classée comme rock et punk en raison de leurs similarités stylistiques. Cette ambiguïté limite la précision des modèles lorsqu'une évaluation stricte est utilisée.

Ces défis soulignent l'importance de travailler sur des datasets plus homogènes et d'adopter des méthodes d'évaluation adaptées pour mieux refléter la nature complexe de la musique et des genres qui la composent.

6. Conclusion et Perspectives

Ce projet a exploré différentes approches de Machine Learning et de Deep Learning pour la classification automatique des genres musicaux. Parmi les modèles testés, le Vision Transformer (ViTb16) a obtenu les meilleures performances avec une précision de 62%, surpassant les approches traditionnelles comme les SVM et les réseaux convolutifs (CNN). Ce résultat met en évidence l'efficacité des architectures modernes pour capturer les relations complexes dans des données audio représentées par des spectrogrammes MEL.

Cependant, les résultats ont également révélé certaines limitations, notamment liées à la qualité et à la distribution des données. Le déséquilibre du dataset et l'ambiguïté des genres ont impacté les performances des modèles, en particulier pour les genres sous-représentés. Ces défis soulignent l'importance d'un traitement adapté des données pour améliorer la généralisation des modèles.

Pour aller plus loin, plusieurs pistes d'amélioration peuvent être envisagées :

- Amélioration du dataset : Homogénéiser la répartition des genres pour réduire l'effet du déséquilibre. Combiner plusieurs datasets pour accroître la diversité et la robustesse des données.
- Modèles et architectures : Tester des architectures avancées telles que ResNet ou des Transformers spécialisés pour les données audio. Expérimenter avec des modèles d'apprentissage non supervisé pour extraire des représentations générales à partir des données.
- Évaluation : Intégrer des métriques adaptées comme la précision Top-3 ou Top-5 pour mieux refléter les performances des modèles sur des tâches multi-genre.

En combinant ces améliorations, il serait possible de développer un système plus performant et mieux adapté à la complexité de la classification des genres musicaux. Ce projet ouvre également des perspectives prometteuses pour l'application des techniques modernes de Machine Learning à des problèmes similaires dans d'autres domaines audio.