Machine learning équitable et interprétable

Reproductibilité du modèle “XAIguiFormer”

Réalisé par :
Ghalia Chaoui - Habibata Samake - Nour Nounah - Safae Hariri - Zeina Gebran

1. Introduction

La détection précoce et précise des troubles psychiatriques à partir de signaux EEG présente un enjeu crucial pour la recherche clinique et la médecine personnalisée. Dans l’article *“XAIGuiFormer: Explainable Artificial Intelligence Guided Transformer for Brain Disorder Identification”* (ICLR 2025), ils proposent un modèle innovant combinant des transformeurs et des mécanismes d’explicabilité (XAI) afin d’extraire des biomarqueurs de connectivité multi-bande électrophysiologique.

Notre objectif est de reproduire fidèlement toutes les étapes de ce travail : du prétraitement des données brutes jusqu’à la visualisation des explications, en passant par l’implémentation des modules du modèle, l’entraînement, et l’évaluation des performances.

2. Implémentation

2.1. Jeu de données et prétraitement

Le pipeline de prétraitement transforme les enregistrements EEG bruts du jeu TDBRAIN en représentations de graphes exploitables par nos modèles. Les étapes principales sont :

- **Chargement des fichiers** BrainVision (*.vhdr, .eeg, .vmrk*) pour chaque sujet.
- **Nettoyage des canaux** : suppression des canaux non-EEG, application du montage 10-20, référence moyenne.
- **Filtrage et rééchantillonnage** : passe-bande 0,5–45 Hz, puis resampling à 250 Hz.
- **Suppression d’artéfacts** : ICA (15 composantes) + classification ICLabel pour retirer les artéfacts (œil, muscle, cœur).
- **Epochage** : découpage en segments de 30 s; sujets/époques sans données sont exclus.

- **Connectivité fonctionnelle** : calcul de la cohérence et du wPLI pour huit bandes de fréquence (Delta, Theta, Low/High Alpha, Low/Mid/High Beta, Low Gamma), puis ajout du ratio Theta/Beta comme neuvième bande (wPLI de cette bande mis à zéro).
- **Agrégation et export** : moyenne des métriques sur les époques, sauvegarde des matrices et métadonnées (*coherence.npy, wpli.npy, demographics.npy, label.npy*).
- **Répartition des jeux** : 80 % train, 10 % validation, 10 % test.

Pour plus de détails, se référer au **README** du dossier *data_pipeline/* sur le dépôt Git.

2.2. Architecture du modèle

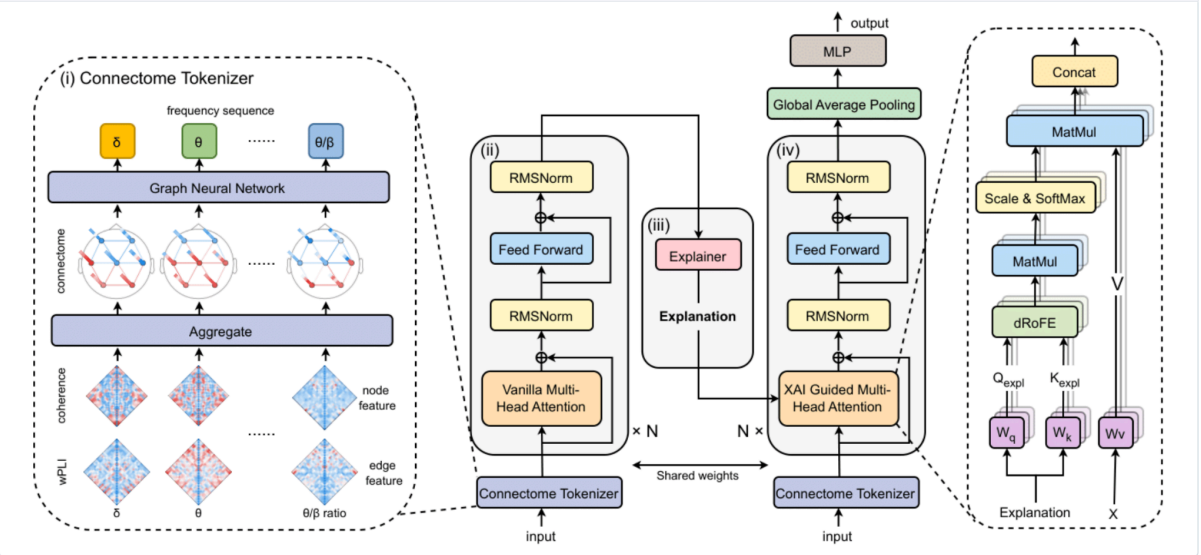

Figure 1 : Architecture du modèle XAIguiFormer

Le pipeline de classification se compose de six blocs principaux, chacun implémenté dans un module distinct :

1. **Connectome Tokenizer** : un réseau de neurones à graphes (GINEConv + mean pooling) encode chacun des neuf connectomes en un vecteur de dimension *out_dim=128.*
2. **dRoFE** : un schéma d'encodage rotatif qui injecte, dans chaque embedding de bande, des informations démographiques (âge, genre), modulées par des fréquences bornées (paramètres *FL, FU).*
3. **Transformeur vanilla** : un encodeur standard de deux couches (Multi-Head Attention suivi d'un MLP), traitant la séquence des neuf tokens.

4. **Explainer** : intégration de Captum DeepLIFT pour calculer les attributions d'importance de chaque token. Une méthode *explain_dataset* permet d'agréger ces scores sur tout un jeu de données.
5. **Transformeur XAI-guidé** : raffinement de l'attention en utilisant les attributions *Qexpl* et *Kexpl* issues de la phase d'explicabilité, suivant l'équation du papier pour renforcer les connexions importantes.
6. **Tête de classification** et **perte duale** : un simple MLP convertit les représentations en logits multiclasses, et la fonction de perte combine deux termes (vanilla et XAI-guidé) pondérés par $\alpha=0.7$.

2.3. Entraînement et validation

Le script **train.py** orchestre l'entraînement sur 30 époques, avec :

- Optimiseur : AdamW (lr=1e-4) avec scheduler cosinus;
- Split : 80 % entraînement, 10 % validation, 10 % test;
- Métriques : Balanced Accuracy, Sensitivity (macro recall), AUC-PR et AUC-ROC.

Pour chaque époque, nous accumulons les logits produits sur tous les lots afin de calculer les métriques globales. Les courbes de convergence sont tracées pour la perte au dessous:

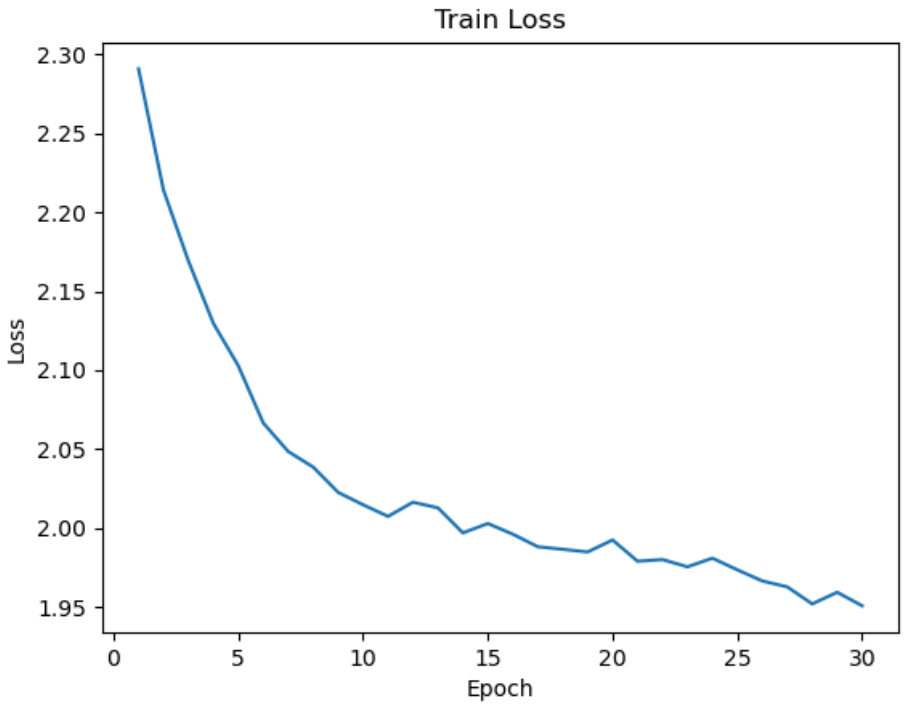

Figure 2 : Évolution de la perte d'entraînement (30 époques).

2.4. Résultat et interprétabilité

Les attributions DeepLIFT moyennées en valeur absolue sur l'ensemble du jeu de validation mettent en avant :

- **Theta/Beta ratio** (score ≈1.35e-3),
- **Low Beta** et **Low Alpha** (score ≈4e-4),
- puis **Delta**, **Mid Beta**, **Theta**, etc.,
- **High Alpha** restant la bande la moins contributive.

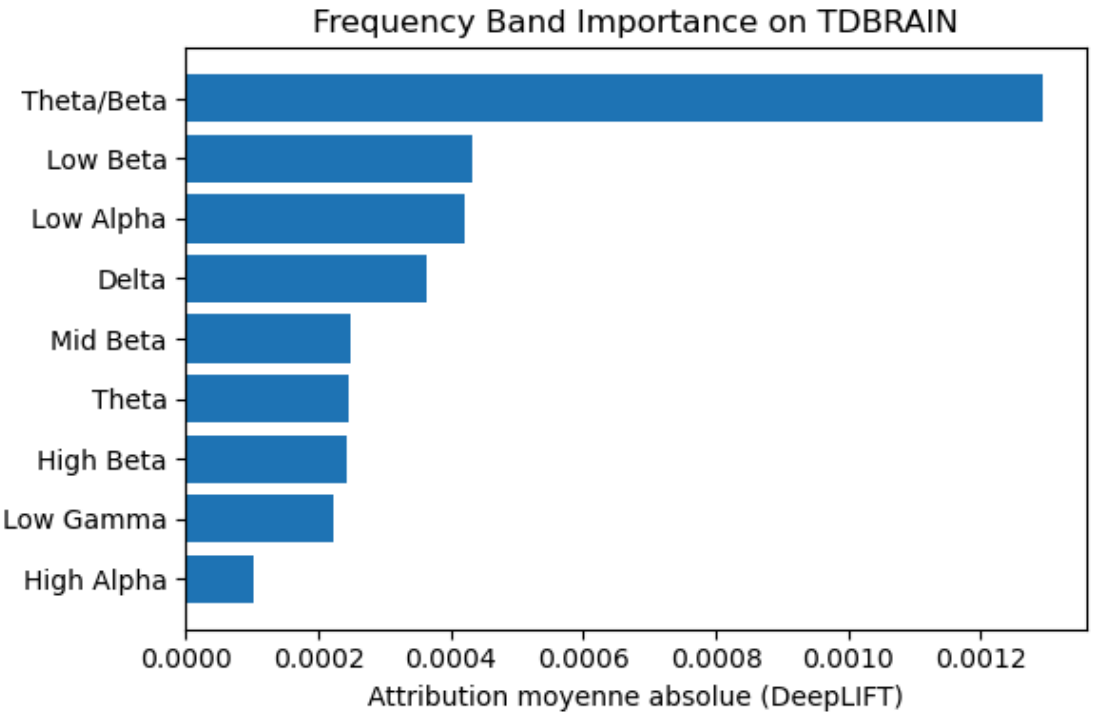

Figure 3 : Importance moyenne des bandes de fréquence sur TDBRAIN.

Ces observations sont en accord qualitatif avec la *Figure 5* du papier, validant que le ratio Theta/Beta agit bien comme biomarqueur clé.

3. Conclusion

Dans ce travail, nous avons fidèlement reproduit l'intégralité du flux expérimental présenté dans l'article *XAIGuiFormer*, de la transformation des signaux EEG bruts en connectomes multi-bandes jusqu'à l'entraînement et à l'évaluation du modèle guidé par XAI. Nos résultats confirment :

- Une **convergence stable** du modèle sur 30 époques grâce à la dual loss combinant sorties vanilla et XAI-guidée.
- La **cohérence** des biomarqueurs identifiés par DeepLIFT, notamment le ratio Theta/Beta.

Certaines différences quantitatives s'expliquent par notre usage d'un échantillon plus restreint du jeu TDBRAIN essentiellement pour des raisons de temps de calcul et de ressources. Néanmoins, le comportement qualitatif du modèle reste identique, démontrant la robustesse et la généralisabilité de XAIGuiFormer.