ALEKSANYAN Volodya
COURNIL-RABEUX Clément
LDD3 Informatique et Mathématiques

Projet de fairness en IA
Données de X-Ray chest images

Objectif du projet

On utilise les donnée de Clement, sous le dossier nommé “clem”

L'objectif de ce projet est d’entraîner un modèle d’intelligence artificielle capable de distinguer les personnes malades et les personnes saines à partir d’images de radiographies thoraciques. Le dataset utilisé contient environ **5000 images**, accompagnées de métadonnées telles que l’âge, le genre et la position de prise de vue (dos ou épaule).

Note :
Les priorités sont notées comme suit :

> **I** : Action prioritaire, réalisée dans le cadre du projet.
> **II** : Action pertinente mais non prioritaire ou non réalisée.
> **III** : Idée ou piste future, pas encore explorée.

Pre-processing

Avant de passer les images au modèle, un travail de prétraitement a été essentiel pour l'équité des données et réduire les biais potentiels.

1. Gérer le déséquilibre des âges

- **II** → *Sur-échantillonnage* des tranches d’âge sous-représentées afin d’éviter que le modèle soit sur-entraîné sur certaines catégories d’âge.

2. Gérer le déséquilibre des positions d’imagerie (dos vs. épaule)

Le dataset contient une majorité d’images prises de dos, avec seulement ~5% d’images issues de la vue par l’épaule. Ce déséquilibre peut générer un biais.

- **II** → *Augmentation de données* (rotations, contrastes, etc.) envisagée mais non réalisée afin de préserver l’indépendance des données.

- **I** → *Pondération dans la fonction de perte* : un poids plus important a été attribué aux images issues de la vue par l'épaule.

3. Normalisation et pré-traitement général

- **I** → *Normalisation des images* : standardisation des pixels pour assurer l'uniformité du jeu de données.

- **I** → *Vérification des fuites de données* : nous avons analysé les métadonnées pour éviter qu'elles n'induisent le modèle en erreur ou le rendent dépendant d'informations indirectes (ex : âge ou genre utilisé involontairement comme indice de prédiction).

Post-processing & Analyse des biais

Une fois le modèle entraîné, nous avons cherché à évaluer s'il présentait des biais sur certaines sous-populations, puis à appliquer des corrections éventuelles.

1. Détection des biais

- **I** → *Analyse de performance selon l'âge*.

- **I** → *Analyse selon le sexe* : nous avons évalué l'AUC, la précision, le taux de faux positifs et de faux négatifs séparément pour les hommes et les femmes.

- **II** → *Analyse selon la position d'imagerie* : envisagée mais non prioritaire dans notre première passe.

2. Réduction des biais

- **I** → CalibratedEqualizedOdds
- **II** → *Re-pondération de la fonction de perte* pour accorder plus d'importance aux classes sous-représentées.

- **III** → *Débiasing adversarial* : envisagé mais non implémenté, cette approche consisterait à entraîner un modèle secondaire pour détecter les biais.

- **III** → *Calibration post-processing* (Platt Scaling, Isotonic Regression) : piste à explorer pour homogénéiser les scores de prédiction entre groupes.

ALEKSANYAN Volodya
COURNIL-RABEUX Clément
LDD3 Informatique et Mathématiques

Plan d’actions

1. **Analyser les biais initiaux** à partir des métadonnées (âge, genre, position d’image).

2. **Choisir une stratégie de correction adaptée :**

    - **Tranches d’âge :** Nous avons opté pour une *pondération des classes* plutôt qu’un ré-échantillonnage explicite.

    - **Genre :** Aucun déséquilibre majeur dans les métadonnées pour le genre, mais des différences physiologiques peuvent induire des biais. Nous avons donc appliqué une *re-pondération selon le genre*.

    - **Position d’imagerie :** L’analyse a été amorcée pour étudier l’impact des vues “épaule” vs “dos”.

3. **Réentraîner le modèle et évaluer l’impact des corrections** :

Nous avons dans un premier temps évalué le modèle sur les **données brutes**, sans traitement particulier. Celui-ci a atteint une **accuracy moyenne de 0.74**, ce qui semble satisfaisant à première vue. Cependant, l’analyse plus poussée a révélé une **forte disparité dans les prédictions**, avec un score de **0.16**, indiquant la présence de biais notables dans les résultats.

Le deuxième entraînement a été effectué en intégrant une **pondération par tranche d'âge** et en appliquant des **transformations sur les images**. Cette approche, bien que légèrement moins performante en termes d’accuracy (**0.69** en moyenne), a permis une **réduction significative des biais**, avec un score de **Statistical Parity passant à 0.10**.

Nous avons ensuite affiné cette stratégie en **intégrant également la variable de genre dans la pondération**. Cette modification n’a pas changé l’accuracy moyenne, qui reste stable (**0.67**), mais elle a permis d’atteindre une **meilleure équité entre groupes**, avec un score de **0.07 en Statistical Parity**.

Enfin, bien que la **re-pondération en fonction de la position d’imagerie (dos vs épaule)** n’ait pas été mise en œuvre, notre analyse suggère qu’elle représente une très piste d’optimisation très prometteuse. En effet, les écarts de perspectives peuvent influencer la détection, et nous pensons que l’ajout de cette étape dans une future itération du pipeline aurait un **impact significatif sur la robustesse du modèle**.

Conclusion

Nos expérimentations montrent que **l'amélioration de l'équité** du modèle est non seulement possible, mais peut être **obtenue sans compromettre fortement la performance globale**.