Figure 1: Sample images of disease severity

1 Description

Dans ce projet, vous participerez à une compétition Kaggle de classification d'images rétiniennes. L'objectif est de développer un modèle d'apprentissage automatique capable d'évaluer la gravité de la rétinopathie diabétique à partir de photographies du fond d'œil. Le fond d'œil, situé à l'arrière de l'œil, contient la rétine où apparaissent les signes de la maladie. Votre but est de classer chaque image dans l'un des niveaux de sévérité ordonnés fournis. Bien que ces distinctions soient souvent subtiles et difficiles à repérer à l'œil nu, les modèles d'apprentissage automatique peuvent saisir des motifs complexes et de grande dimension, ce qui permet d'obtenir des prédictions précises.

Afin d'accomplir cette tâche, les données suivantes vous sont fournies:

- `train_data.pkl` - Il s'agit d'un fichier au format pickle qui contient, dans un dictionnaire, à la fois les images et les étiquettes du "training set". Vous pouvez l'ouvrir à l'aide du script montré au Listing 1
- `test_data.pkl` - Un autre fichier pickle contenant les images de l'ensemble de test. Cette fois, vous pouvez ouvrir ce fichier exactement comme auparavant, mais le dictionnaire ne contiendra pas la clé "labels" que vous devrez prédire.

2 Inscription des équipes (22 novembre)

`Note importante:` Tous les étudiants doivent inscrire leurs équipes sur Kaggle AVANT le 22 novembre ; sinon, AUCUNE inscription ne sera autorisée sur le site. Pour la section des cycles supérieurs (IFT6390), la tâche doit être réalisée **individuellement.** Les étudiants d'IFT3395 participent à la compétition en **équipes de 2 ou 3.** Une participation individuelle est également autorisée. Pour participer à la compétition, vous devez:

- Créer un compte Kaggle si vous n'en avez pas déjà un.
- Rejoindre la compétition via le lien d'invitation suivant: `https://www.kaggle.com/t/8e1f67cbcad64991b0b707ec9478901e`.
- Dès lors, vous pouvez accéder à la compétition.

Listing 1: Loading data from a pickle file

```python
import pickle

# Load the pickle file
with open('path_to_file.pkl', 'rb') as f:
    data = pickle.load(f)

# Access images and labels
images = data['images']
labels = data['labels']
```

- Dans la section "Invite Others", entrez les noms de vos coéquipiers ou le nom de votre équipe.
- Votre coéquipier peut accepter la fusion de l'équipe.
- Remplissez le formulaire google `https://docs.google.com/forms/d/e/1FAIpQLSeP60RhGxei-EdIM viewform?usp=dialog` avec les informations de votre équipe.
- IMPORTANT: Veuillez utiliser votre email institutionnel (procuré par l'université)

Note importante: Le nombre maximal de soumissions est de 3 par jour, par ÉQUIPE. Toute équipe dont les membres individuels dépassent le nombre de soumissions autorisé jusqu'à la date sera INCAPABLE de former une équipe. Exemple : Aujourd'hui est le premier jour de la compétition. A, B et C sont trois coéquipiers qui n'ont pas encore formé d'équipe.

- A a soumis 0 fois.
- B a soumis 3 fois.
- C a soumis 1 fois.

Comme le nombre maximum de soumissions est de 3 par équipe et par jour, le total des soumissions possibles pour une équipe est de 3. Cependant, le nombre cumulé de soumissions pour A, B et C est de 4. Par conséquent, ils ne pourront pas former une équipe (Ils devront attendre demain et ne soumettre aucune soumission le jour suivant).

3 Premier jalon : Battre le score de référence (30 novembre) [50pts]

Vous pouvez voir deux scores de référence sur le tableau de classement. Vous devrez battre les deux modèles de référence au classement public afin d'obtenir la note maximale. Vous pouvez utiliser la méthode de votre choix, mais gardez à l'esprit que vous travaillez avec un petit jeu de données ($<$ 1k examples).

Note importante : Pour battre le score de référence, vous n'êtes PAS autorisé à utiliser de bibliothèque d'apprentissage automatique, comme `scikit-learn`. Vous devez implémenter votre solution à partir de zéro en utilisant uniquement NumPy et les fonctionnalités de base de Python.

4 Deuxième jalon : Compétition (8 décembre)

Dans cette phase, vous êtes libre d'implémenter toute méthode et d'utiliser toute bibliothèque, comme scikit-learn, Pytorch ou Tensorflow.

Note importante : Le tableau de classement de Kaggle comporte une composante publique et une composante privée pour éviter le "surapprentissage" sur le tableau de classement. Le tableau de classement public montre votre score calculé sur 50% de l'ensemble de test, tandis que le tableau de classement privé est basé sur votre score sur l'autre moitié. Vous ne pourrez voir que le tableau public pendant la compétition. Les points de cette phase seront attribués en fonction de votre classement sur le tableau privé.

Note importante : Vous devez soumettre deux solutions distinctes, une pour la première phase (battre le score de référence) et une pour la deuxième phase (votre modèle le plus performant). Vous devez nommer vos fichiers de soumission pour distinguer les deux. Pour votre soumission de code sur Gradescope, vous devez également séparer les deux solutions.

5 Troisième jalon : Soumettre le code et le rapport (12 décembre) [40pts]

Vous devez rédiger un rapport détaillant votre pipeline d'apprentissage, y compris le prétraitement, les algorithmes, l'optimisation et l'apprentissage, le réglage des hyperparamètres et la validation. Vous devez également présenter et comparer les résultats des autres méthodes que vous avez mises en œuvre avant d'aboutir au modèle le plus performant. Le rapport doit inclure des détails sur les méthodes utilisées lors du premier et du deuxième jalon. Le rapport doit inclure :

- Titre du projet
- Nom de votre équipe sur Kaggle et liste des membres, noms complets et numéros d'étudiant.
- Introduction : description du problème et résumé de votre approche.
- Conception des caractéristiques : description et justification de vos méthodes de prétraitement.
- Algorithmes : aperçu des algorithmes d'apprentissage utilisés.
- Méthodologie : répartition entraînement/validation, régularisation, optimisation, etc.
- Résultats : analyse détaillée avec comparaisons de différentes valeurs d'hyperparamètres.

- Discussion : avantages/inconvénients de votre approche, idées d'amélioration.
- Références (obligatoire pour les idées empruntées).
- Annexe (optionnelle). Ici, vous pouvez inclure des résultats supplémentaires, des détails sur les méthodes, etc.

Le texte principal du rapport ne doit pas dépasser 6 pages. Les références et l'annexe peuvent être ajoutées. Vous devez soumettre votre code (premier et deuxième jalons) ainsi que votre rapport (troisième jalon) sur Gradescope au plus tard le **12 décembre à 23 h 59**.

Instructions de soumission

- Vous devez avoir des fichiers `.py` / notebooks distincts pour les premier et deuxième jalons. Le code doit être bien documenté. Si vous n'utilisez pas de notebooks Jupyter, vous devez inclure un fichier `README` contenant les instructions pour exécuter le code. Vous devrez soumettre un fichier `ZIP` contenant votre code et les fichiers associés sur **Gradescope**.
- Le rapport au format PDF (rédigé selon la structure générale décrite plus haut) doit être déposé sur **Gradescope**.
- Le fichier de prédictions contenant vos prédictions sur l'ensemble de test doit être soumis uniquement sur **Kaggle**. Votre fichier de soumission doit être un fichier `CSV` comportant deux colonnes et une ligne d'en-tête. **IMPORTANT :** Assurez-vous que vos identifiants sont indexés à partir de 1 et non à partir de 0, comme suit :

```
// sample_submission.csv
ID,Label
1,7
2,5
3,8
...
```

6 Critères d'évaluation

- **Premier jalon de la compétition de données (50 points)**
  - **20 points** : Battre le modèle de référence aléatoire.
  - **30 points** : Battre le modèle de référence solide.
- **Deuxième jalon de la compétition de données (10 points)**
  - **5 points** : Se classer au-dessus de la performance médiane.
  - **5 points** : Obtenir un classement dans le top-3.

- **Rapport (40 points)**
  - Votre note reposera sur la clarté, la profondeur et la rigueur technique de votre rapport final.

7 Résumé de la date limite

Les dates limites de ce projet sont strictes, et chaque remise doit inclure tous les éléments exigés pour chaque jalon. Le non-respect de ces exigences aux dates indiquées peut entraîner une perte de points.

- **22 novembre, 23:59** – Date limite pour former les équipes, soumettre le formulaire Google et s'inscrire sur Kaggle.
- **30 novembre, 23:59** – Date limite pour battre les modèles de base sur Kaggle.
- **8 décembre, 23:59** – Fin de la compétition Kaggle.
- **12 décembre, 23:59** – Soumettez votre rapport et votre code sur Gradescope.