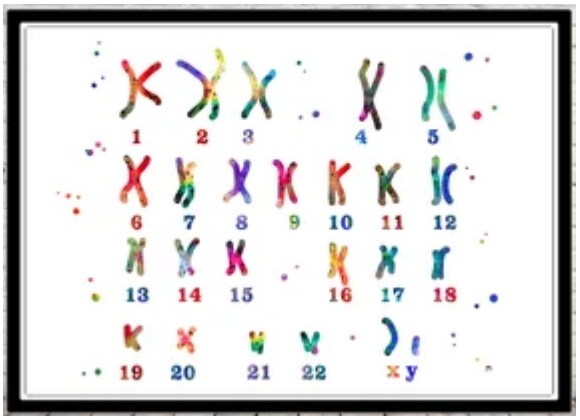
Reconnaissance automatique des anomalies chromosomiques afin d’accélérer le diagnostic pronostic dans la prise en charge des cancers du sang

---

Baptiste Le Goff
Guillaume Verpoest
Thomas Chaigneau



2020/2022 - Ecole Microsoft IA by Simplon

I. Contexte du projet

A l'heure actuelle, les cancers du sang, hémopathies malignes, représentent 30% des cancers de l'enfant et 5% des cancers chez l'adulte. Afin de pouvoir classifier ces cancers, il est nécessaire de prendre en compte le tissu d'origine, la morphologie, les marqueurs protéiques, ainsi que les données chromosomiques de ces cancers.

Les cancers du sang sont diagnostiqués grâce à la réalisation du caryotype de l'individu, car ils sont dus à une translocation de gènes, dans laquelle les chromosomes 9 et 22 ont échangé réciproquement des fragments de chromosomes.

Pour pouvoir mettre en évidence ce type d'anomalie chromosomique, il est donc nécessaire de classer les chromosomes afin d'établir le caryotype, puis de l'analyser. Ce projet est composé de deux parties, la première se charge de la segmentation et de la classification, et la deuxième se consacre à la détection d'anomalies.

Un caryotype est réalisé à partir d'une photographie d'une cellule en vue microscopique lors de la métaphase de la mitose. A cette étape, la chromatine est condensée ce qui rend les chromosomes visibles. Dans un caryotype, les chromosomes sont classés par paire, par taille et en fonction de la position du centromère. Il y a 23 paires de chromosomes autosomes (de 1 à 22) et une paire de gonosomes (XX pour les femmes et XY pour les hommes).

Aujourd'hui, il n'existe pas encore de logiciel capable de réaliser entièrement un caryotype, de manière automatique. Les logiciels actuels sont semi-automatiques et nécessitent encore l'intervention des techniciens de laboratoire qui savent reconnaître les chromosomes.

A partir d'une photographie d'une cellule issue d'un microscope (sous format TIF, JPEG ou BMP), le logiciel Ikaros Meta System, utilisé par le laboratoire de cytogénétique de Brest, traite l'image. Les techniciens séparent ensuite les chromosomes de manière semi-automatique, car ils doivent détourer les chromosomes qui se chevauchent. Les autres chromosomes sont séparés par le logiciel lui-même. A partir de ces séparations, le logiciel classe par taille les chromosomes et les range ensuite dans le caryotype.

II. Première approche

Après avoir récupéré les images des caryotypes (figure 1), nous avons commencé par préparer les données.

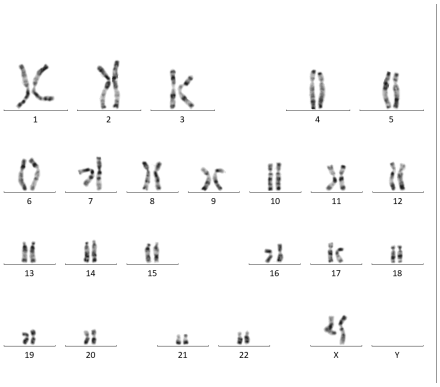


Figure 1: Exemple d'un caryotype

Dans un premier temps, afin d'avoir une base d'images pour entraîner un modèle, nous avons séparé chaque paire de chromosomes du caryotype (figure 2). L'un des premiers obstacles est le nombre d'images. Il fallait récupérer 23 images (une image par paire de chromosomes) pour les caryotypes de femme, 24 pour les caryotypes d'homme et ce sur 203 images.

Puisque chaque image de caryotype avait la même organisation, nous avons créé un script permettant de retirer la barre et le nombre puis de crop chaque paire. Nous avons ainsi récupéré un jeu d'images (203 par classe) pour chaque classe afin de lancer un entraînement de classifieur.

Figure 2: Paire 1 de chromosomes

Pour l'avoir utilisé plusieurs fois, nous avons choisi d'utiliser un CNN *MobileNet* comme classifier. Les résultats obtenus n'ont pas du tout été probants (figure 3) nous avons alors réfléchi à une autre solution.

```
Epoch 192/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0015 - accuracy: 0.9999 - val_loss: 285.0042 - val_accuracy: 0.0765
Epoch 193/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0014 - accuracy: 0.9997 - val_loss: 0.8649 - val_accuracy: 0.8580
Epoch 194/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0016 - accuracy: 0.9997 - val_loss: 79.1810 - val_accuracy: 0.0843
Epoch 195/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0016 - accuracy: 0.9997 - val_loss: 194.1971 - val_accuracy: 0.0440
Epoch 196/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0014 - accuracy: 0.9999 - val_loss: 126.8047 - val_accuracy: 0.1587
Epoch 197/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0010 - accuracy: 0.9999 - val_loss: 0.4992 - val_accuracy: 0.9277
Epoch 198/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0031 - accuracy: 0.9991 - val_loss: 1608.1890 - val_accuracy: 0.0435
Epoch 199/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0072 - accuracy: 0.9978 - val_loss: 5.6614 - val_accuracy: 0.3436
Epoch 200/200
239/239 [==============================] - 5s 20ms/step - loss: 0.0023 - accuracy: 0.9995 - val_loss: 28153.6660 - val_accuracy: 0.0435
60/60 [==============================] - 0s 5ms/step - loss: 28153.6660 - accuracy: 0.0435
60/60 - 0s - loss: 28153.6660 - accuracy: 0.0435
```

figure 3: résultats de l'entraînement avec MobileNet

Nous voyons que l'entraînement du modèle de classification donne des résultats très médiocres, puisqu'en fonction de l'epoch les scores sont complétement opposés. L'un des problèmes possible est que certaines paires sont très proches voire identiques (figure 4), ce qui mène à un entraînement très aléatoire.

Figure 4: Exemple des paires 13, 14 et 15 d'un caryotype.

III. Brainstorming

Au vu des difficultés rencontrées pour traiter les données que ce soit au niveau de la classification mais également de la labellisation (23 classes représentées sur 203 images), nous avons pris le temps de l'alternance pour réfléchir à des solutions qui pourraient nous aider.

Pour la partie classification, nous n'avons pas encore trouvé de solutions, mais pour ce qui est de la détection et de la labellisation, nous avons trouvé une solution dans les projets effectués avec notre entreprise.

Nous sommes arrivés à la conclusion que nous n'étions pas partis sur des outils adaptés au projet, malgré que nous les maîtrisions. Nous avons donc dû trouver d'autres outils, potentiellement récents, qui permettent de solutionner les problèmes posés par le projet.

+

IV. Présentation de la solution

Voici notre proposition pour résoudre le problème posé par le projet de segmentation des chromosomes.

A. Label Studio

Label Studio est un outil de labellisation de données. Il permet de labelliser tous les types de données, audio, texte, vidéo, image, etc. Cet outil est développé par la société Heartex, basée à San Francisco en Californie.

Voici l'éventail des templates disponibles pour cet outil :

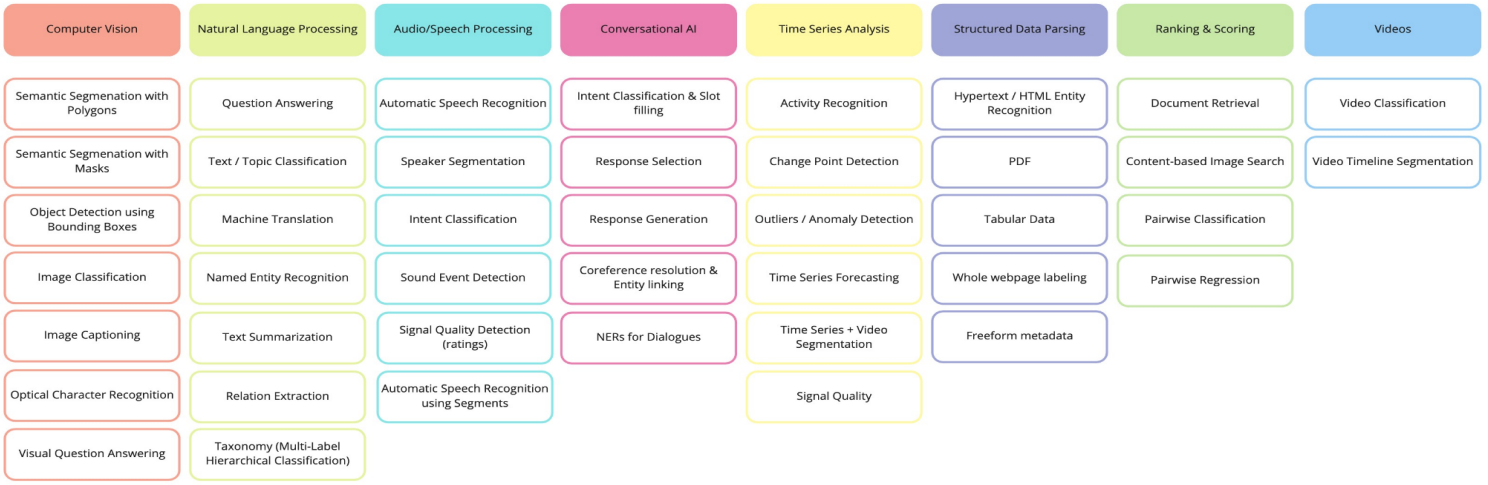


Figure 5: Templates disponibles pour la labellisation dans Label Studio

L'avantage de Label Studio est qu'il n'intègre pas uniquement une partie frontend comme la plupart des autres outils de labellisation, mais aussi une partie backend qui permet de bénéficier de la puissance de l'*active learning*. Le principe consiste à entraîner en parallèle du travail de labellisation un modèle qui va permettre une assistance en temps réel lors de la fastidieuse étape de labellisation.

L'autre avantage de ce logiciel est qu'il embarque un système d'identification et d'utilisateurs. Ce qui permettra l'utilisation de notre solution par plusieurs experts du CHRU pour la tâche de réalisation des caryotypes.

Une fois le projet initialisé et les données importées, le travail de labellisation peut commencer, voici un exemple d'une interface pour la détection de véhicules sur la piste d'un aéroport :

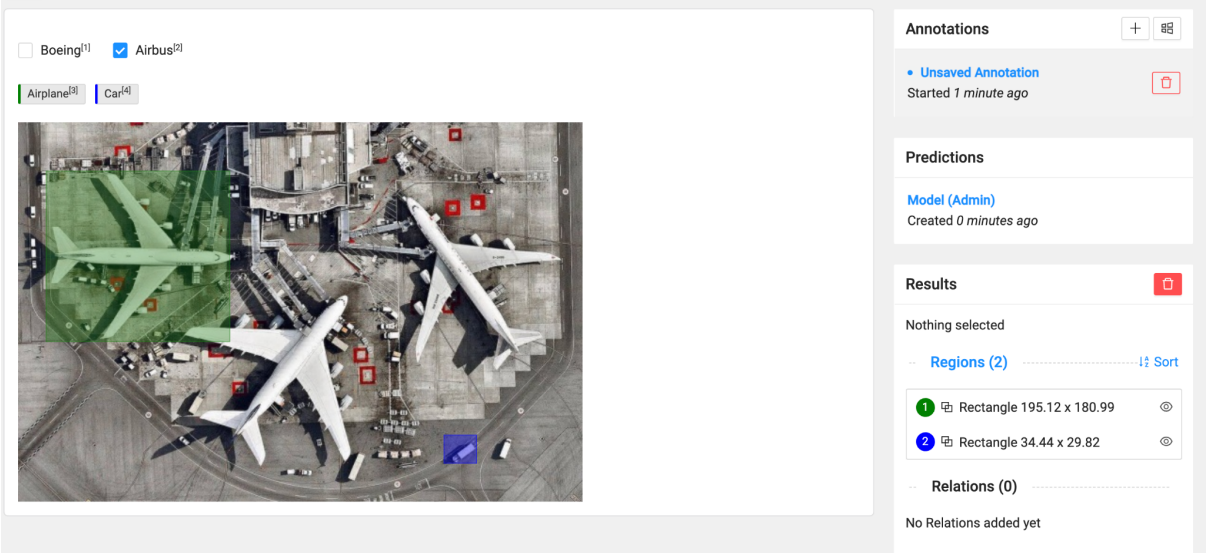


Figure 6: Exemple de l'interface de labellisation sur Label Studio

Il faut désormais connecter un modèle via le système de backend de Label Studio pour permettre un entraînement et une assistance dans la labellisation. Une base d'environ 300 images labellisées devrait permettre d'avoir un premier modèle assez performant pour commencer l'assistance pour poursuivre le travail de labellisation.

B. Yolov5

YOLOv5 🚀 est une famille d'architectures et de modèles de détection d'objets pré-entraînés sur l'ensemble de données COCO (un dataset constitué de 328K images pour le développement de model), et représente la recherche open source d'Ultralytics sur les futures méthodes d'IA de vision, incorporant les leçons apprises et les meilleures pratiques développées au cours de milliers d'heures de recherche et développement.

Voici un graphique présentant les performances du modèle :

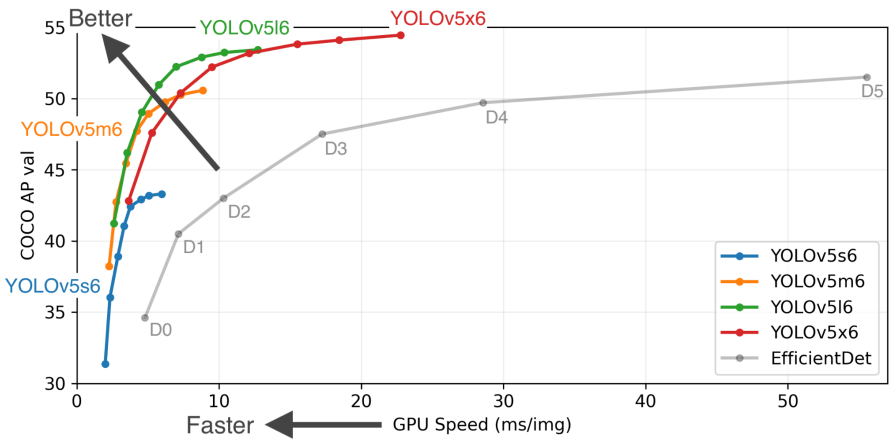


Figure 7: Graphique comparatif des performances des différents modèles YoloV5 avec EfficientDet

On peut constater que cette architecture récente permet d'obtenir des performances nettement supérieures à des modèles bien établis comme par exemple *EfficientDet*, même pour des formats *small* de l'architecture.

En effet, il existe plusieurs tailles disponibles pour l'architecture, voici les détails pour chacune d'entre elles :

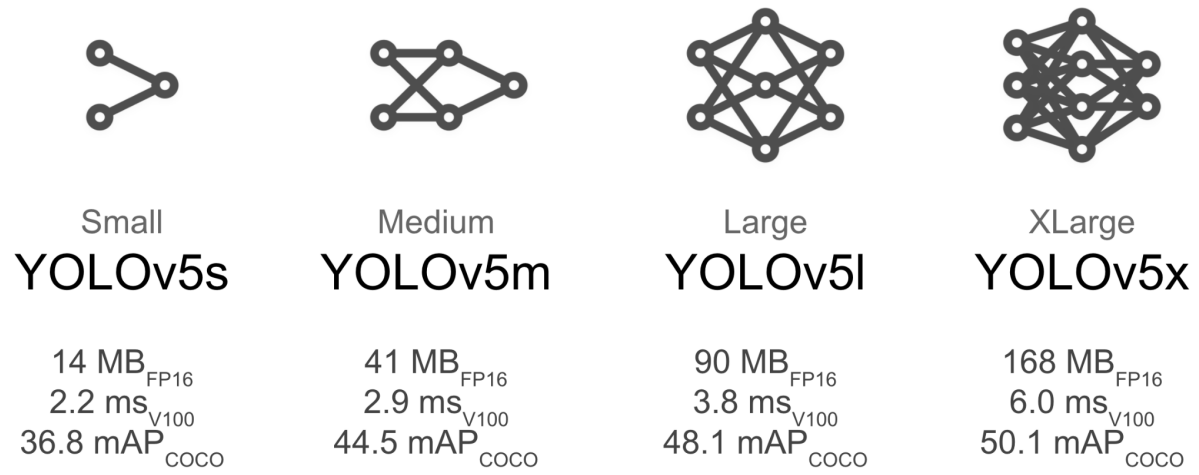


Figure 8: Différentes tailles disponibles pour les modèles YoloV5

C. Installation du système au CHRU

Une interface sera présentée aux experts de caryotypes qui sera connectée à leur système de base données. Cette dernière stocke les images de métaphase utilisées pour que les experts métier puissent séparer les chromosomes et établir des caryotypes.

Pour un modèle pertinent, un minimum de 300 images doivent être labellisées manuellement, pour ensuite permettre la mise en place de l'active learning et de générer un modèle semi-automatique qui labellise les nouvelles images au fur et à mesure.

Le système proposé permet un gain de temps mais également de conserver l'étape de vérification par l'expert métier qui assure une qualité dans le processus de création de caryotypes.

Un script de création de caryotypes suit l'étape de labellisation. Chaque image labellisée passe par ce script pour ensuite automatiquement créer un caryotype à partir de la labellisation.

Une fois le caryotype créé il est ajouté sur une base de données qui stocke les résultats et il ne reste plus qu'à le faire entrer dans une étape de vérification des chromosomes, soit par un expert métier, soit par un autre modèle que nous n'avons pas encore défini.

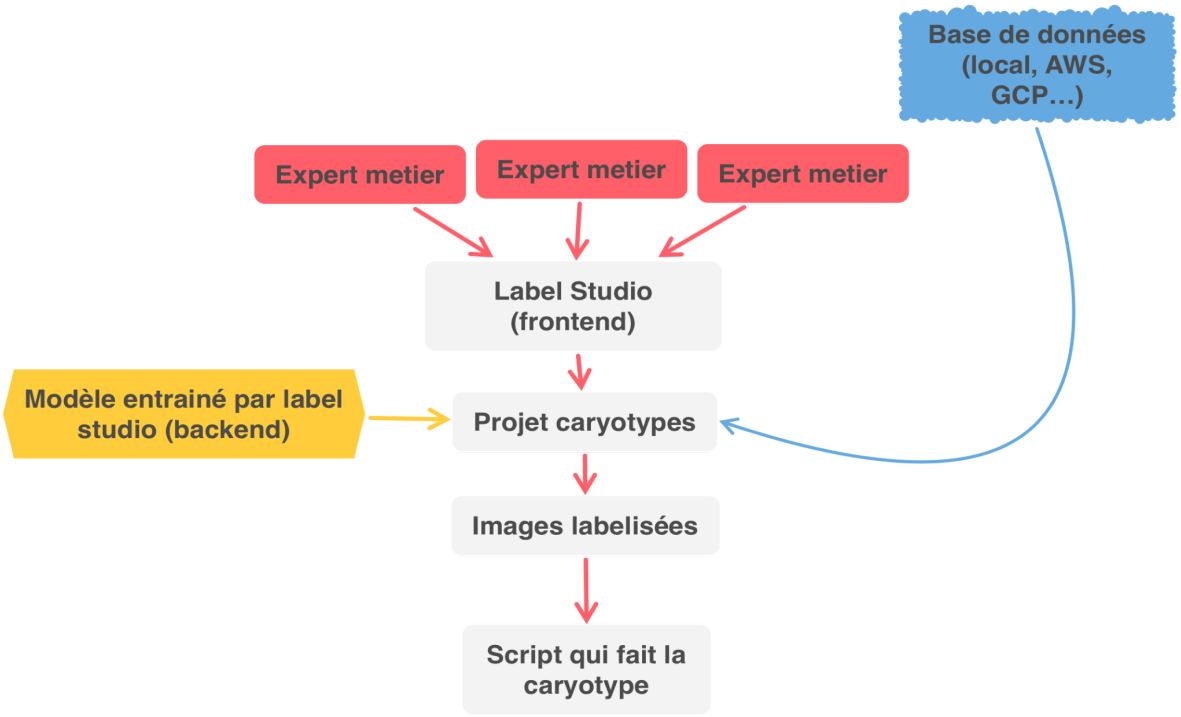


Figure 9: Schéma explicatif du système proposé

V. Conclusion

Avec ce projet, nous avons pu nous rendre compte qu'il est important de rester informé en permanence des nouvelles solutions qui apparaissent sur le marché puisque chaque projet demande une solution adaptée et qu'il est impossible de résoudre tous les problèmes avec une solution générique qui fonctionnait dans d'autres projets.

Grâce au système proposé par notre solution, le CHRU connaîtrait un gain effectif de temps sur la réalisation des caryotypes, grâce à l'assistance d'un modèle d'IA très performant pour les travaux de segmentation/détection.

De plus, le contrôle qualité serait intégré dans l'effort de labellisation et permettrait également un gain non négligeable sur la qualité de la réalisation des caryotypes par les experts métier qui auraient ainsi plus de temps à allouer à des tâches à plus haute valeur ajoutée.