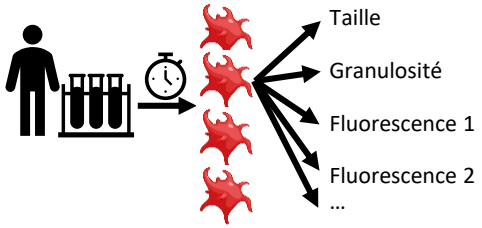
L'apprentissage machine pour le diagnostic par cytométrie en flux de la TIH

Auteurs : Matthieu Stoll, Frédéric Allemand (Emosis SAS, Illkirch-Graffenstaden)

Contexte :

Le diagnostic de la thrombopénie induite par l'héparine (TIH) est d'importance vitale, mais les tests fonctionnels de référence (SRA, HIPA) nécessitent plusieurs jours pour avoir les résultats.

La cytométrie en flux offre un test fonctionnel alternatif **rapide**
- **multiparamétrique** (taille, granulosité, fluorescences,…)
- **cellule par cellule** (*single cell analysis*)
- sur des **milliers de cellules**

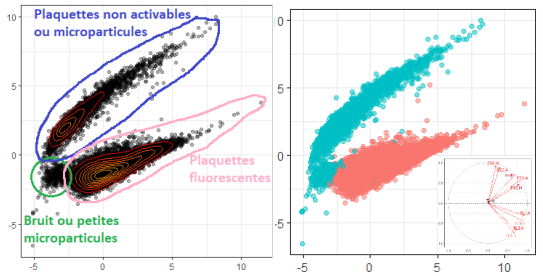


→ **Richesse informationnelle**

Objectif : Augmenter la précision du test fonctionnel par cytométrie en flux « FCA » à l'aide de méthodes d'intelligence artificielle (apprentissage statistique, *machine learning*,…)

Méthode :

1. Automatisation de la sélection des plaquettes parmi les cellules mesurées



- Méthodes algorithmiques de *clustering* (kmeans, classification hiérarchique,…)
- Validation visuelle de la meilleure méthode : une classification ascendante hiérarchique en deux classes

→ Remplace un *gating* réalisé visuellement par l'opérateur

2. Modélisation

- Requiert une expertise conjointe en biologie et en statistique, notamment pour le choix des variables liées à la présence de TIH
- Données : 154 individus pour lesquels la présence ou l'absence de TIH est connue
- Plusieurs types de modèles (régressions logistiques, arbres de décision, réseaux de neurones,…)
- Optimisation des modèles dans un but prédictif (*machine learning, « tuning »*)

Résultats :

Modèle prédictif le plus précis : un arbre de décision par la méthode « XGBoost »

Méthode	Sensibilité	Spécificité	Précision
FCA	90,0 %	93,7 %	92,0 %
FCA + *machine learning*	90,0 %	**95,8 %**	**93,3 %**

L'apprentissage machine (*machine learning*) a permis dans cette étude :
- Une **procédure automatisée, rapide, et indépendante de l'opérateur** en phase de réalisation du test
- L'**extraction de l'information**, très riche en cytométrie
- Une estimation **fiable** des performances
- Un **gain de spécificité et de précision**, peu significatif ici, en particulier car la méthode « FCA » est déjà très précise

Conclusion :

Cette étude montre le potentiel de l'IA pour l'amélioration des performances de diagnostic par cytométrie en flux : vers un diagnostic « augmenté » ?
L'appréciation du gain de performance nécessite des études complémentaires (autres algorithmes de classification, autres variables caractérisant l'activation plaquettaire, standardisation,…)

JIB 2020 JOURNÉES DE L'INNOVATION EN BIOLOGIE

Processus de mise en œuvre :

REALISATION DU TEST FCA « HISTORIQUE »

REALISATION DU TEST AVEC LES MODELES STATISTIQUES

L'EXPERTISE HUMAINE DANS LE DEVELOPPEMENT DU TEST

L'APPRENTISSAGE STATISTIQUE DANS LE DEVELOPPEMENT DU TEST

Manipulation et acquisition des données par cytométrie

Gating manuel

Classification

Variables relatives à l'activation plaquettaire

Modélisation « à la main »

Modélisation par *machine learning*

Valeur %HEPLA

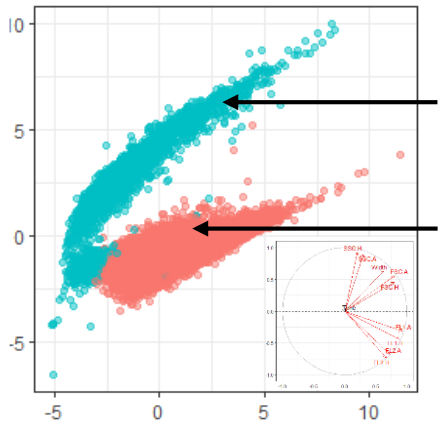
Probabilité(TIH+)

Classement du patient comme TIH+ ou TIH-

Performance (précision, sensibilité, spécificité) du test diagnostic

Classification et identification des plaquettes :

- Les cellules mesurées sont agrégées en *clusters* par la méthode de Ward. Cet algorithme regroupe les cellules et les classes de cellules de sorte à maximiser l'inertie inter-classe
- Un *clustering* en deux classes est validé visuellement, à l'aide des représentations graphiques des cellules :



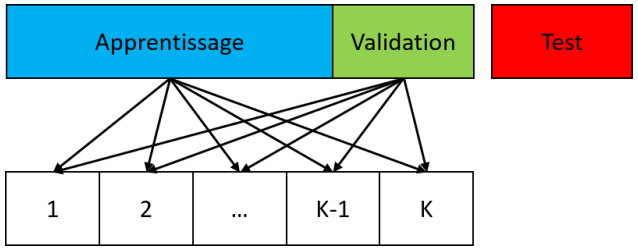
Résidus et microparticules divers, bruit

Plaquettes (forte luminescence PE)

→ Méthode de sélection des plaquettes automatisable

Modélisation de la présence de TIH :

- 25% des données (jeu de « test ») pour estimer les performances du modèle final
- 75% des données pour estimer et comparer des centaines de modèles par validation croisée :

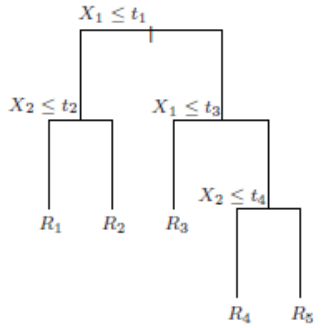


Partition en K blocs. Chaque bloc est tour à tour ensemble de validation ; il reste K-1 blocs pour l'apprentissage, c'est-à-dire l'estimation du modèle. On estime l'erreur par la moyenne des K erreurs de classement.

→ Estimation de la précision de chaque modèle
→ Optimisation des paramètres du modèle pour la meilleure capacité prédictive : « *tuning* »
→ Comparaison des modèles et sélection du meilleur modèle : un arbre de décision par la méthode « XGBoost »

Arbre de décision : un raisonnement naturel pour l'humain.

L'arbre est un partitionnement récursif des données. Pour prédire l'affection, il suffit de parcourir l'arbre depuis le sommet pour déterminer à quel nœud terminal appartient le patient, à qui on attribue la classe majoritaire dans ce nœud.



→ La méthode « XGBoost » inclut de nombreuses fonctionnalités permettant la construction d'un arbre décisionnel optimisé

→ Un tel modèle fournit, en sortie, une probabilité d'appartenance à la classe « TIH+ »