

Contexte et idée clé

L'électrophorèse des protéines sériques (EPS) sépare et quantifie les principales fractions protéiques. En routine, l'interprétation est souvent **visuelle** : gel et électrophérogramme.

Beaucoup d'approches automatisées imitent cette lecture via des CNN [?]. Pourtant, l'électrophérogramme est intrinsèquement une **courbe numérique**, donc exploitable comme **données tabulaires**.

Objectif. Comparer une approche tabulaire, CatBoost, à une baseline CNN [?], sur le même dataset [?], avec le même split test de 10%, **sans tuning**, afin d'évaluer un gain directement applicable.

Données

Cohorte : 1289 cas, comprenant gels et électrophérogrammes.

Classes étudiées :

- Acute phase protein : n = 74
- Monoclonal gammopathy : n = 264
- Polyclonal gammopathy : n = 244
- Hypoproteinemia : n = 249
- Nephrotic syndrome : n = 165
- Normal : n = 293

Point clé : forte variabilité des images, notamment en dimensions et en qualité, surtout pour les gels.

Méthodes : extraction tabulaire

Pipeline images → variables :

1. Passage en niveaux de gris et recadrage de la zone analytique.
2. Extraction colonne par colonne pour obtenir un profil 1D.
3. Interpolation à 150 points et normalisation min-max.
4. Extraction des fractions via minima locaux : albumine, α_1 , α_2 , β , γ .
5. Ajout des variables cliniques : sexe, âge, protéines totales, albumine.

Modèles testés : XGBoost, TabPFN et CatBoost.

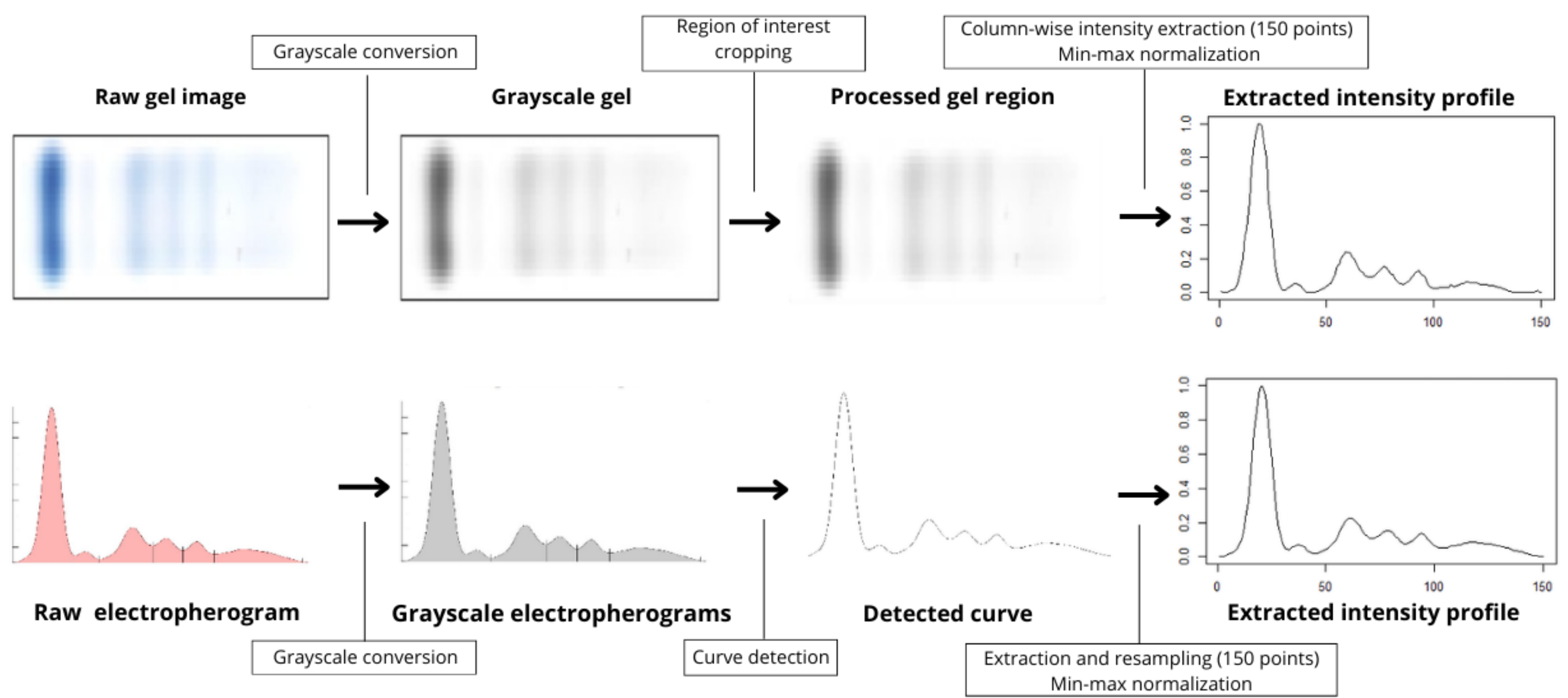


Figure 1. Pipeline d'extraction tabulaire à partir d'images EPS.

Résultats : CatBoost tabulaire vs CNN

Protocole : même split que [?], avec 10% de test, répété sur **100 seeds** avec IC95%.

Scores évalués : sensibilité, spécificité, F1-score par classe et score pondéré.

Résumé quantitatif :

- **Gels** : gain fort en F1 pondéré : **0.599** → **0.786**, soit $\Delta = +0.187$.
- **Électrophérogrammes** : gain modéré en F1 pondéré : **0.784** → **0.849**, soit $\Delta = +0.065$.
- Gains marqués sur **acute phase protein**, **polyclonal gammopathy**, **hypoproteinemia** et **normal**.

Cas monoclonal : sur électrophérogrammes, la baseline CNN est meilleure. Cela peut être lié à des **annotations manuelles de délimitation de pics** présentes sur les monoclonales, susceptibles d'avantager un modèle image, alors qu'elles sont ignorées par l'approche tabulaire.

Classe	Électrophérogrammes			Gels		
	F1 Ref	F1 CatBoost	Δ	F1 Ref	F1 CatBoost	Δ
Acute phase protein	0.429	0.648	+0.219	0.160	0.576	+0.416
Monoclonal gammopathy*	0.926	0.792	-0.134	0.844	0.788	-0.056
Polyclonal gammopathy	0.857	0.950	+0.093	0.711	0.925	+0.214
Hypoproteinemia	0.698	0.887	+0.189	0.531	0.738	+0.207
Nephrotic syndrome	0.786	0.783	-0.003	0.345	0.704	+0.359
Normal	0.727	0.868	+0.141	0.698	0.854	+0.156
Moyenne pondérée	0.784	0.849	+0.065	0.599	0.786	+0.187

Figure 2. F1-score par classe. Δ = CatBoost – Référence. * Délimitations de pics sur monoclonales.

Discussion

Pourquoi le tabulaire aide : les profils EPS sont des signaux 1D. Les modèles tabulaires exploitent directement la structure du signal, les intensités, les fractions et les variables cliniques, sans devoir réapprendre une représentation depuis l'image.

Robustesse et déploiement :

- **Robustesse à la qualité d'image**, particulièrement importante pour les gels.
- **Coût computationnel faible** : entraînement rapide, sans GPU.
- **Intégration clinique facilitée** : les analyseurs stockent déjà les courbes.
- **Interprétabilité** : importance des variables et des zones discriminantes de la courbe.

Limite : le dataset monoclonal peut être biaisé par des marquages visuels sur les électrophérogrammes.

Conclusion

Une approche tabulaire simple, basée sur CatBoost, améliore nettement la classification des EPS, surtout sur gels, **sans tuning**.

Après validation de la méthode, des optimisations complémentaires, comme le tuning ou l'utilisation d'algorithmes plus poussés, pourront faciliter un déploiement en routine.

Références

[1] Lee N., Jeong S., Jeon K., Song W., Park M.-J. (2022). Development and validation of a deep learning-based protein electrophoresis classification algorithm. PLOS ONE, 17(8) :e0273284.

[2] Lee N. (2022). Dataset of deep learning based protein electrophoresis classification. Harvard Dataverse. doi :10.7910/DVN/FG9NXZ.