

**RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
SCIENTIFIQUE**

UNIVERSITÉ KASDI MERBAH – OUARGLA

Faculté des Hydrocarbures, des Énergies Renouvelables, des Sciences de la Terre et de l'Univers
Département des Sciences de la Terre et de l'Univers

Mémoire de Master LMD

Domaine : Sciences de la Terre et de l'Univers
Spécialité : Géologie des Hydrocarbures

Thème

**Évaluation des modèles de prédiction des paramètres pétro-physiques
des réservoirs pétroliers à partir des données de diagraphies : Cas du
Champ de Hassi Tarfa**

Présenté par :

Dahdi Abdellah & Ben Abdelouahad Abdelwafi

Soutenu le 02/06/2026 devant le jury composé de :

Belekseir Mohamed Salah	Pr	UKMO – Ouargla	Président
Benkhelifa Randa	MCB	UKMO – Ouargla	Examinatrice
Chemam Asma	MCB	UKMO – Ouargla	Encadrante
Nemer Zoubida	MCB	UKMO – Ouargla	Co-Encadrante

Année universitaire : 2025 / 2026

Remerciements

Avant de commencer cet humble travail, il nous est particulièrement important d'exprimer notre gratitude et nos remerciements à tous ceux qui ont contribué directement ou indirectement à son élaboration.

Nous remercions en premier lieu Dieu Tout-Puissant, qui nous a accordé la force et le courage de poursuivre nos études.

Nous tenons à exprimer notre sincère appréciation à nos encadrantes, le Dr NEMER ZOUBIDA et Dr ASMA CHEMAM, MaîtreS de Conférences à l'Université Kasdi Merbah Ouargla, pour leur soutien constant, leur précieux conseils et leur disponibilité tout au long de l'élaboration de ce mémoire.

Nous adressons nos vifs remerciements aux membres du jury, Pr BELKSEIR Mohamed Salah (Professeur, Président) et Dr BENKHELIFA Randa (MCB, Examinatrice), pour l'honneur qu'ils nous font d'évaluer ce travail.

Nous remercions également nos familles et nos proches, dont le soutien moral et les encouragements constants ont été une source de motivation inestimable.

Enfin, nos sincères remerciements s'adressent à tous nos enseignants, collègues et à toute personne ayant contribué, de près ou de loin, à l'aboutissement de ce mémoire.

Sommaire

Remerciements	2
Liste des figures.....	5
Liste des tableaux	5
Liste des Abréviations	6
Introduction Générale	11
Chapitre I : Contexte Géologique et Description des Données du Champ de Hassi Tarfa .	12
I.1. Situation Géographique.....	12
I.2. Contexte Structural et Tectonique.....	13
I.3. Stratigraphie	13
I.3.2. Paléozoïque (Cambro-Ordovicien) Réservoir Principal	15
I.3.3. Mésozoïque	15
I.3.4. Cénozoïque.....	15
I.4. Historique de l'Exploration	16
I.5. Caractéristiques du Réservoir.....	16
I.5.1. Lithologie et Paramètres Pétrophysiques	16
I.5.2. Aquifère.....	17
I.6. Présentation des Puits Disponibles.....	17
I.7. Description des Données de Diagraphies	17
I.8. Sources et Formats des Données	18
Chapitre II : Contexte Théorique.....	20
II.1. Fondamentaux des Paramètres Pétrophysiques	20
II.1.1. La Porosité (Phi)	20
II.1.2. La Perméabilité (K).....	20
II.1.3. La Saturation en Eau (Sw)	21
II.1.4. Le Volume d'Argile (Vcl / VSH)	21
II.2. Techniques d'Apprentissage Automatique Appliquées à la Pétrophysique	22
II.2.1. Principes Généraux du Machine Learning.....	22
II.2.2. Modèles de Régression Utilisés	22
II.2.3. Métriques d'Évaluation des Modèles	23
II.2.4. Prétraitement des Données.....	23
Chapitre III : Méthodologie.....	24
III.1. Introduction et Vue d'Ensemble	24
III.2. Description du Jeu de Données	26
III.3. Contrôle Qualité des Données	26
III.3.1. Remplacement des Valeurs Sentinelles Nulles	26
III.3.2. Filtrage par Plage Physique Valide	26

III.3.3. Disponibilité des Attributs et Imputation	26
III.4. Découpage Train/Test Basé sur les Puits	27
III.5. Modèles d'Apprentissage Automatique	27
III.5.1. Régression Ridge	27
III.5.2. Bagging	27
III.5.3. Arbres Extrêmement Aléatoires (Extra Trees)	27
III.5.4. XGBoost	28
III.6. Métriques d'Évaluation des Modèles	28
III.7. Déploiement du Modèle et Génération des Sorties	28
III.8. Limites et Sources d'Incertitude	28
Chapitre IV : Résultats et Discussion	30
IV.1. Introduction	30
IV.2. Analyse Exploratoire des Données	30
IV.2.1. Matrice de Corrélation de Spearman	30
IV.2.2. Importance des Attributs	31
IV.3. Prédiction de la Saturation en Eau (Sw)	32
IV.3.1. Comparaison des Modèles	32
IV.3.2. Distributions Réelles vs. Prédites	33
IV.3.3. Diagramme Croisé	33
IV.3.4. Performance par Puits de Test	34
IV.3.5. Profils en Profondeur et Analyse des Résidus	34
IV.4. Prédiction du Volume d'Argile (Vcl)	36
IV.4.1. Comparaison des Modèles	36
IV.4.2. Diagramme Croisé et Profils par Puits	37
IV.5. Porosité (ϕ) : Validation Croisée sur HTF-19	39
IV.6. Synthèse Spatiale	39
IV.7. Discussion Générale	40
IV.7.1. Performances Globales et Benchmark Littérature	40
IV.7.2. Cohérence Physique	40
IV.7.3. Recommandations	41
IV.8. Synthèse Finale	41
Conclusion Générale	42
Références Bibliographiques	44

Liste des figures

Figure 1: Situation géographique du champ de Hassi Tarfa (SONATRACH, 2007)	12
Figure 2: Colonne lithostratigraphique type de la région de Hassi Tarfa (SONATRACH, 2007)	14
Figure 3: Matrice de disponibilité des données de diagraphe au niveau de chaque puit....	18
Figure 4: Organigramme de la méthodologie de prédiction pétrophysique par apprentissage automatique dans le champ HTF	26
Figure 5: Matrice de corrélation de Spearman (puits d'apprentissage Sw)	31
Figure 6: Importance relative des attributs pour la prédiction de Sw (Random Forest)	32
Figure 7: Distributions de Sw (gauche) et Vcl (droite) réelles et prédites (puits de test aveugles)	33
Figure 8: Diagramme croisé Sw prédite vs. réelle (XGBoost)	34
Figure 9: Profils Sw vs. profondeur (six puits de test)	35
Figure 10: Profils résiduels ΔSw (six puits de test)	36
Figure 11: Diagramme croisé Vcl prédite vs. réelle (XGBoost)	37
Figure 12: Profils Vcl vs. profondeur (quatre puits de test)	38
Figure 13: Boîtes à moustaches- Sw et Vcl réelles vs. prédites par puits	39
Figure 14: R^2 XGBoost par puits de test pour Sw et Vcl avec code couleur	40

Liste des tableaux

Tableau 1: Répartition des puits pour chaque variable cible	27
Tableau 2: Récapitulatif des performances XGBoost par variable cible	28
Tableau 3: Métriques de performance des quatre modèles pour la prédiction de Sw (6 puits de test aveugles, $n = 4\,098$)	32
Tableau 4: Performance XGBoost par puits de test pour la prédiction de Sw	34
Tableau 5: Métriques de performance pour la prédiction de Vcl (4 puits de test aveugles, $n = 2\,990$)	36
Tableau 6: Performance XGBoost par puits de test pour la prédiction de Vcl	38
Tableau 7: Résultats de la validation croisée 5 plis pour la porosité (HTF-19, $n = 667$)	39
Tableau 8: Synthèse des performances et niveaux de fiabilité opérationnelle	41

Liste des Abréviations

Abréviation	Signification
ϕ	Porosité
Sw	Saturation en eau
K	Perméabilité
Vcl / VSH	Volume d'argile (Volume of Clay / Volume of Shale)
PHIE	Porosité effective
RHOB	Densité volumique apparente (Bulk Density)
NPHI	Porosité neutronique
GR	Rayon gamma (Gamma Ray) [API]
RT / RT10-RT90	Résistivité vraie (niveaux d'induction)
CALI	Diamètre du puits (Caliper)
DTC / DTS	Temps de transit sonique compression / cisaillement
PE	Facteur photoélectrique
POTA / URAN / THOR	Potassium / Uranium / Thorium (spectroscopie gamma)
ML	Apprentissage automatique (Machine Learning)
IA	Intelligence Artificielle
ACP / PCA	Analyse en Composantes Principales
EDA / AED	Analyse Exploratoire des Données
SCAL	Analyse spéciale sur carottes (Special Core Analysis Laboratory)
MAE / EAM	Erreur Absolue Moyenne (Mean Absolute Error)
RMSE / REMC	Racine de l'Erreur Quadratique Moyenne (Root Mean Square Error)
MAPE / EAMP	Erreur Absolue Moyenne en Pourcentage (Mean Absolute Percentage Error)
R ²	Coefficient de détermination
AIC	Critère d'information d'Akaike
KDE / EDN	Estimation de Densité par Noyau (Kernel Density Estimation)
XGBoost	Extreme Gradient Boosting
RF	Forêt aléatoire (Random Forest)
SVM	Machine à vecteurs de support (Support Vector Machine)
CNN	Réseau neuronal convolutif (Convolutional Neural Network)
DST	Test en tige de forage (Drill Stem Test)
API	American Petroleum Institute (unité du GR)
G.O.R.	Rapport gaz/huile (Gas-Oil Ratio)
HTF	Hassi Tarfa Field
LAS	Log ASCII Standard
CSV	Comma-Separated Values
VC	Validation Croisée (k-fold Cross-Validation)

Résumé

Cette étude académique vise à évaluer la performance des modèles d'apprentissage automatique pour prédire avec précision les paramètres pétrophysiques clés d'un réservoir hydrocarboné, tels que la porosité, le volume d'argile et la saturation en eau, en prenant comme cas d'étude le champ de Hassi Tarfa (HTF), situé dans la province triasique du Sahara algérien au sud de Hassi Messaoud. Ce réservoir est classé comme une formation compacte (tight) caractérisée par une faible porosité (5–10 %) et une très faible perméabilité (1–2 mD), rendant son interprétation traditionnelle complexe. L'analyse repose sur un large ensemble de données de diagraphies issues de 17 puits du champ, où les données ont subi un contrôle qualité rigoureux avant de tester quatre algorithmes de régression : Ridge, Bagging, Extra Trees et le modèle avancé XGBoost. Afin de garantir une validation robuste et d'éviter le surapprentissage, la stratégie de division des données a été strictement exécutée sur la base de puits aveugles (blind wells) complets. Les résultats démontrent la supériorité remarquable de l'algorithme XGBoost, qui a capturé efficacement les relations non linéaires complexes et prédit la saturation en eau avec une forte corrélation par rapport aux données de carottage réelles. Cette approche numérique innovante permet d'optimiser le temps et de réduire considérablement les coûts par rapport aux méthodes d'interprétation manuelles conventionnelles. En conclusion, ce travail ouvre des perspectives prometteuses pour l'intégration et le déploiement opérationnel de l'intelligence artificielle dans l'industrie pétrolière en Algérie.

Summary

This academic study evaluates the effectiveness of advanced machine learning models in predicting essential petrophysical properties such as porosity, clay volume, and water saturation, focusing on the Hassi Tarfa (HTF) field in the Triassic province of the Algerian Sahara. The targeted reservoir is classified as a tight formation, characterized by low porosity (5–10%) and very low permeability (1–2 mD), which complicates traditional interpretation workflow. The database incorporates comprehensive well-logging curves extracted from 17 operational wells, where data underwent rigorous quality control before implementing four regression algorithms: Ridge, Bagging, Extra Trees, and XGBoost. To ensure robust validation and prevent data leakage, the data splitting strategy was strictly executed based on full blind test wells. The experimental results demonstrate that XGBoost significantly outperformed the other models by accurately handling highly non-linear subsurface data and predicting water saturation with excellent agreement compared to laboratory core analyses. This data-driven workflow offers a faster and more cost-effective alternative to traditional manual log interpretation techniques used in the industry. Ultimately, this study provides valuable operational recommendations and highlights the high potential of artificial intelligence deployment in Algerian oilfields.

الملخص

تهدف هذه الدراسة الأكاديمية إلى تقييم كفاءة وقدرة نماذج التعلم الآلي المتطورة في التنبؤ بدقة بالخصائص البتروفيزيائية الجوهرية للمكامن النفطية، مثل المسامية، وحجم الطين، والتشبع المائي، وذلك بالتطبيق على حقل "حاسي طرفة" الواقع في المديرية الثلاثية بالصحراء الجزائرية جنوب حقل حاسي مسعود. يُصنف هذا المكن ضمن التكوينات الضيقة والمدمجة (Tight Reservoirs) التي تتميز بمسامية منخفضة تتراوح بين 5% و10% ونفاذية ضئيلة جداً لا تتعدى 1 إلى 2 مللي دارسي، مما يجعل تفسيرها التقليدي أمراً معقداً. وقد اعتمد البحث على قاعدة بيانات رقمية ضخمة شملت تسجيلات الآبار (Diagraphies) لـ 17 بئراً مختلفة، حيث خضعت البيانات لعمليات تنظيف ومعالجة مكثفة قبل اختبار أربعة خوارزميات رئيسية هي: Ridge Regression و Bagging و Extra Trees و XGBoost. ولضمان دقة التقييم ومنع التداخل، تم اعتماد استراتيجية صارمة لتقسيم البيانات بناءً على استبعاد آبار كاملة (Blind Wells) لاختبار النماذج بشكل حيادي تماماً. وقد أظهرت النتائج التجريبية تفوقاً ملحوظاً وخاصاً لخوارزمية "XGBoost" التي أثبتت قدرة استثنائية على محاكاة العلاقات غير الخطية المعقدة وتوقع قيم التشبع المائي بدقة توافقت بشكل ممتاز مع القياسات المختبرية المباشرة للعينات الأسطوانية. وتكمن الأهمية العملية لهذه المنهجية الرقمية في قدرتها العالية على تقليل التكاليف الباهظة واختصار الوقت الطويل مقارنة بالطرق التقليدية واليدوية المستعملة في تفسير تسجيلات الآبار. وفي الختام، تفتح هذه المذكرة آفاقاً واعدة لتطوير عمليات محاكاة المكامن وتعميم تقنيات الذكاء الاصطناعي كأداة موثوقة في الحقول النفطية الجزائرية مستقبلاً.

Introduction Générale

La caractérisation des réservoirs pétroliers constitue une étape fondamentale dans l'industrie pétrolière. Elle vise à déterminer les paramètres pétrophysiques essentiels, porosité, perméabilité et saturation en eau, qui conditionnent l'estimation des réserves et l'optimisation de la production ([Peters, 2007](#)).

Les méthodes traditionnelles reposent sur l'interprétation manuelle des diagraphies et les analyses de carottes. Cependant, la complexité croissante des réservoirs et le volume important de données disponibles motivent le recours à des approches numériques avancées, notamment les algorithmes d'apprentissage automatique. Comme le soulignent [Anifowose et al. \(2017\)](#), l'apprentissage automatique offre une capacité inégalée à capturer des relations non linéaires et multidimensionnelles entre les réponses diagraphiques et les propriétés de réservoir.

Le présent mémoire porte sur l'évaluation de modèles de prédiction des paramètres pétrophysiques à partir des données de diagraphies, appliqués au champ de Hassi Tarfa (HTF), situé dans la province triasique au sud de l'Algérie. Ce champ, caractérisé par des réservoirs gréseux cambro-ordoviciens à faible porosité et perméabilité, présente une forte hétérogénéité lithologique qui en fait un cas d'étude particulièrement pertinent pour les approches d'apprentissage automatique.

Ce travail est organisé en quatre chapitres :

- Chapitre I - Contexte géologique et description des données du champ de Hassi Tarfa : situation géographique, structure tectonique, stratigraphie, historique de l'exploration, caractéristiques pétrophysiques du réservoir et présentation des données de diagraphie disponibles.
- Chapitre II - Contexte théorique : fondamentaux des paramètres pétrophysiques et techniques d'apprentissage automatique appliquées à la pétrophysique.
- Chapitre III - Méthodologie : pipeline de traitement des données, sélection et entraînement des modèles, métriques d'évaluation.
- Chapitre IV - Résultats et discussion : présentation et analyse des performances des modèles, synthèse spatiale et recommandations opérationnelles.

Une conclusion générale synthétise les apports du travail et propose des perspectives pour les recherches futures.

Chapitre I : Contexte Géologique et Description des Données du Champ de Hassi Tarfa

I.1. Situation Géographique

Le champ de Hassi Tarfa est situé dans la province triasique, au sud du gisement de Hassi Messaoud, dans la région du Sahara algérien. Il constitue une structure satellite s'étendant vers le sud, le long du trend El Gassi – El Agreb – Hassi Messaoud, dans le bloc n° 427 . Cette province triasique représente l'une des zones les plus prolifiques d'Algérie en matière de production d'hydrocarbures, et le champ de Hassi Tarfa y occupe une position stratégique en raison de sa proximité avec les grandes infrastructures de production existantes.

Le champ est délimité par les structures suivantes :

- Au nord et au nord-est : par le champ de Hassi Messaoud, premier producteur d'huile d'Algérie.
- A l'ouest : par la structure anticlinale de Hassi D'zabat.
- Au sud-ouest : par les gisements d'El Gassi, Zotti et El Agreb.
- A l'est : par le champ de Mesdar et le bassin de Berkine.

Cette position géographique privilégiée confère à Hassi Tarfa un accès aux infrastructures de production existantes tout en présentant des caractéristiques géologiques distinctes qui justifient une étude approfondie. La Figure 1 illustre la situation géographique du champ dans son contexte régional.

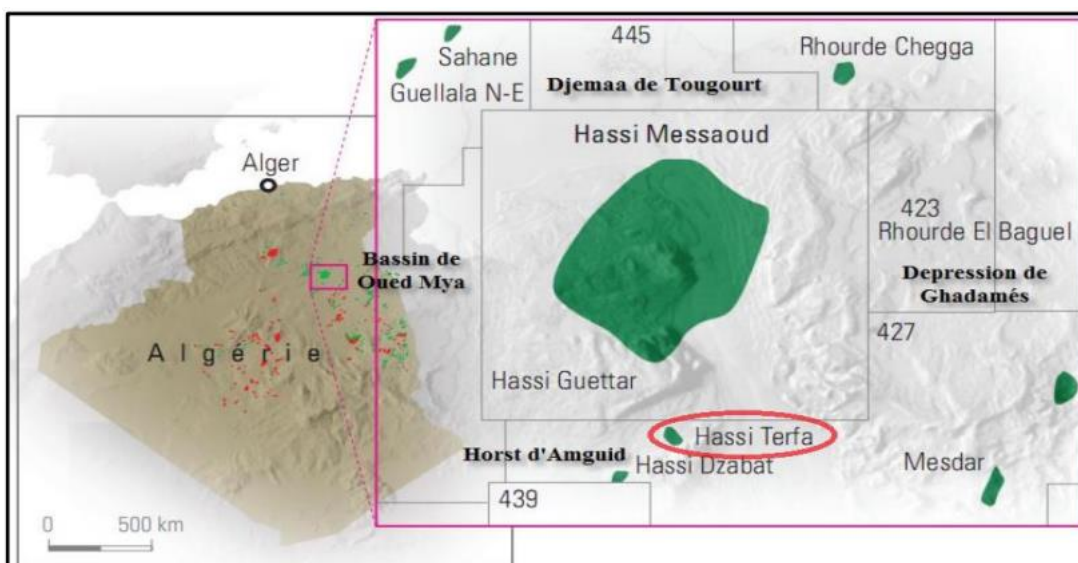


Figure 1: Situation géographique du champ de Hassi Tarfa (SONATRACH, 2007)

I.2. Contexte Structural et Tectonique

La structure de Hassi Tarfa se présente comme un pli anticlinal d'orientation NE-SO, découpé par un système de failles longitudinales (NNE-SSO) et de failles transversales secondaires (E-O), créant un réseau complexe de horsts et de grabens. Ces discontinuités structurales jouent un rôle déterminant dans la perméabilité et la fracturation du réservoir, conditionnant directement la distribution et la mobilité des hydrocarbures.

Le champ est subdivisé en 25 compartiments de production relativement indépendants, délimités par ces failles majeures. La communication entre compartiments est conditionnée par la fracturation intense des grès cambro-ordoviciens. Cette compartimentation complexe constitue l'un des défis majeurs pour la modélisation de réservoir et pour la prédiction des paramètres pétrophysiques à l'aide de méthodes d'apprentissage automatique. En effet, l'hétérogénéité structurale se traduit par une variabilité spatiale marquée des propriétés de réservoir d'un compartiment à l'autre ([Wood, 2025](#)).

L'histoire tectonique est marquée par trois phases de déformation principales :

- La phase hercynienne (Carbonifère–Permien) : responsable de l'érosion de la couverture paléozoïque et de la formation de la discordance majeure séparant le Paléozoïque du Mésozoïque.
- La phase triasique : dépôt des formations évaporitiques et argilo-salifères qui constituent le scellement du réservoir.
- Les réactivations crétacées et tertiaires : responsables de l'accentuation des failles et de la géométrie actuelle du champ.

I.3. Stratigraphie

La colonne stratigraphique du champ de Hassi Tarfa est caractérisée par une importante discordance hercynienne, qui marque la séparation entre le Paléozoïque et le Mésozoïque (Figure 2). Les formations rencontrées sont décrites ci-après, de la base au sommet.

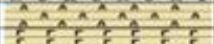
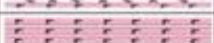
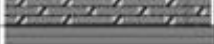
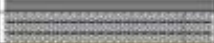
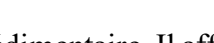
Ere	Système	Série	Étage	Ép.(m)	Stratigraphie	Lithologie
CENOZOÏQUE	NÉOGENÈ	Mio-Plio		178		Sable, Grès et Argile
	PALÉOGENÈ	Eocene		123		Calcaire crayeux
MESOZOÏQUE	CRÉTACÉ	SENONIEN	Carbonaté	180		Calcaire et Dolomie
			Anhydritique	217		Anhydrites, calcaire blanc et Dolomie
			Salifère	134		Sel massif incolore à blanc
			Turonien		116	
		Cénomanién		179		Anhydrite, Dolomie, parfois Argile Grise
		Albien		300		Grès Fin à Moyen et Intercalation d'Argile Brun Rouge et de Sable Grossier à la base
		Aptien		27		Dolomie et Marne
		Barremien		260		Grès, Argile silto-sableuse, et Dolomie
		Néocomien		208		Argile carbonate avec passées de Grès
	JURASSIQUE	MALM (Late Jurassic)		229		Argile Silteuse à intercalation de Dolomie de Calcaire et Marne
		DOGGER	Argileux	77		Argile indurée, Dolomie, Grès et Anhydrite
			Lagunaire	244		Anhydrite et Dolomie, passées d'Argile Silteuse
		LIAS	LD-1	38		Anhydrite et Argile
			LS-1	226		Sel et Argile.
			LD-2	55		Anhydrite et Argile
			LS-2	59		Sel et Argile
			H.B: horizon B	28		Argile et Dolomie
		TRIAS	TS1	12		Anhydrite Intercalé d'Argile Dolomitique
			TS2	159		Sel rose Massif, avec passées d'Argile indurée et Anhydrite
	TS3		195		Sel rose Massif à la base, avec passées d'Argile.	
	Argileux		96		Argile silteuse (Brun Rouge) parfois Salifère	
	Roches éruptives		68		Roches Éruptives	
DISCORDANCE HERCYNIENNE						
PALEOZOÏQUE	ORDOVICIEN	Grès de Ouargla	50		Argile silteuse avec des passées de Grès	
		Quartzites de Hamra	126		Grès Quartzites à Quartzite	
		Grès d'El Atchane	25		Grès (glauconieux) Gris Clair + Argile	
		Argile d'El Gassi	100		Argile Gris sombre	
		ZA: zone d'alternances	29		Argile et Grès	
	CAMBRIEN	Camb "R1"	49		Grès gris beige fin à moyen, Tigite	
		Camb "Ra"	120		Grès blanc beige (moyen à grossier)	
		Camb "R2"	100		Grès micro-conglomératique	
		Camb "R3"	370		Grès grossier, conglomératique	
INFRA-CAMBRIEN		45		Grès Argileux rouge		
SOCLE				Granite porphyroïde rose		

Figure 2: Colonne lithostratigraphique type de la région de Hassi Tarfa (SONATRACH, 2007)

I.3.1. Socle précambrien

Le socle granitique constitue la base de la série sédimentaire. Il affleure dans certaines zones du Hoggar et est recoupé à grande profondeur par les forages dans la région de Hassi Tarfa. Ses propriétés mécaniques influencent la géométrie des structures sus-jacentes par l'intermédiaire de failles héritées (Ahmed T, 2018).

I.3.2. Paléozoïque (Cambro-Ordovicien) Réservoir Principal

Le Paléozoïque constitue la principale tranche réservoir du champ. On distingue, du bas vers le haut, plusieurs unités lithologiques aux propriétés pétrophysiques distinctes :

- Ra (Cambrien inférieur) : grès à stratification horizontale, d'origine fluviatile. Zone la plus productive du champ, avec une porosité légèrement supérieure aux autres niveaux gréseux (épaisseur ~60 m au puits HTF-1).
- Ri (Cambrien moyen) : grès à stratification subhorizontale, milieu marin côtier peu profond. Épaisseur constante d'environ 49 m au puits HTF-1. Ce niveau contribue de manière significative à la production totale du champ.
- Zone des alternances : succession de grès et d'argiles à tigillites (terriers fossiles), d'épaisseur variable entre 13 et 29 m selon les puits. Cette hétérogénéité verticale est à l'origine d'une forte variabilité des réponses diagaphiques et constitue un défi pour l'interpolation des propriétés pétrophysiques.
- Grès d'El Atchane (Ordovicien) : stratifications obliques à subhorizontales, témoignant d'un environnement marin côtier. Leur porosité est généralement plus faible que les niveaux cambriens sous-jacents.
- Quartzites de Hamra (Ordovicien) : formation à morphologie massive, avec des fissurations intenses et de faibles variations de faciès. Principal réservoir de la structure. Épaisseur : HTF-1 (116 m), HTF-2 (104 m), HTF-3 (126 m).

I.3.3. Mésozoïque

Le Mésozoïque comprend plusieurs formations dont le rôle vis-à-vis du réservoir est avant tout celui de roche-couverture :

- Trias : couverture argilo-salifère d'environ 150 m, assurant l'étanchéité du réservoir paléozoïque. L'intégrité de ce scellement est un paramètre critique pour la rétention des hydrocarbures.
- Jurassique : épaisseur croissante du sud vers le nord (HTF-2 : 1 333 m, HTF-3 : 1 372 m), reflétant la subsidence différentielle du bassin.
- Crétacé : membre inférieur à faciès détritique continental, membre supérieur à dominante carbonatée. Ces formations ne constituent pas de réservoir économiquement significatif dans le champ HTF.

I.3.4. Cénozoïque

Les dépôts de l'Éocène et du Mio-Pliocène, constitués de sables et de calcaires de plateforme épicontinentale, forment la partie superficielle de la colonne sédimentaire. Ils n'ont pas de signification pétrolière dans ce contexte.

I.4. Historique de l'Exploration

Le champ de Hassi Tarfa a été découvert en septembre 1999, lors du forage du puits HTF-1, qui a confirmé la présence d'hydrocarbures dans les formations gréseuses cambriennes (niveaux Ra et Ri), à une profondeur d'environ 3 318 m ([SONATRACH, 2007](#)). Ce premier forage a immédiatement mis en évidence les caractéristiques pétrophysiques singulières du champ : faible porosité, très faible perméabilité et forte compartimentation structurale.

En 2001, les puits HTFN-1 et HTF-2 ont été forés pour délimiter la structure et évaluer l'extension du réservoir ordovicien. En mai 2005, un nouveau puits a permis d'affiner la compréhension des Quartzites de Hamra et des grès cambriens, en particulier leur hétérogénéité verticale et latérale. Actuellement, le champ compte 22 puits de production répartis en plusieurs collecteurs, avec une production principalement issue des niveaux cambriens Ra et Ri.

I.5. Caractéristiques du Réservoir

I.5.1. Lithologie et Paramètres Pétrophysiques

Le réservoir principal est constitué par les grès quartzeux anisométriques des niveaux cambriens (Ra et Ri). Ces formations présentent une forte hétérogénéité verticale et latérale, liée à la silicification et à la présence d'argiles. Les propriétés pétrophysiques de référence, établies à partir des analyses de carottes (SCAL) et de l'interprétation des diagraphies, sont les suivantes :

- Porosité (ϕ) : faible, comprise entre 5 % et 10 %. Cette valeur est significativement inférieure à celle des réservoirs gréseux conventionnels (15–25 %), ce qui confère au champ HTF un caractère de réservoir compact (tight reservoir).
- Saturation en eau (S_w) : variant de 20 % dans les zones à hydrocarbures à 100 % au contact huile-eau, estimé à environ 3 380–3 444 m de profondeur. Cette large gamme de variation est l'une des raisons qui motivent le recours aux modèles d'apprentissage automatique.
- Perméabilité (K) : très réduite, entre 1 et 2 mD en moyenne. Cette faible perméabilité nécessite des stratégies de stimulation particulières (fracturation hydraulique) pour maintenir des débits de production économiquement viables.

Ces paramètres placent Hassi Tarfa dans la catégorie des réservoirs à faible qualité pétrophysique, pour lesquels les approches d'apprentissage automatique ont démontré leur intérêt dans la littérature ([Elkatatny et al., 2018](#) ; [Stadtmüller, M., & Jarzyna, J. A. 2023](#)).

I.5.2. Aquifère

L'aquifère associé est considéré comme inactif, ne contribuant pas de manière significative au maintien de la pression de réservoir. L'eau de gisement est très saline, avec une minéralisation de 360 à 370 g/L et une densité de 1,21 g/cm³ (SONATRACH, 2007). Cette haute salinité est un paramètre important pour l'interprétation des diagraphies de résistivité et le calcul de la saturation en eau via la loi d'Archie (2003).

I.6. Présentation des Puits Disponibles

L'étude utilise les données de diagraphie de 17 puits forés dans le champ HTF : HTF-01, HTF-10, HTF-13, HTF-14, HTF-19, HTF-20, HTF-24, HTF-25, HTF-28, HTF-29, HTF-30, HTF-32, HTF-33, HTF-36, HTF-37, HTF-38 et HTF-42. Ces puits ont été sélectionnés en raison de la disponibilité et de la qualité relative de leurs données de diagraphie.

Les 22 puits du champ sont raccordés à deux collecteurs principaux :

- Collecteur HTFM 1 : 5 puits (HTF-12, HTF-20, HTFN-5, HTFN-1, HTFN-3).
- Collecteur HTFM 2 : 6 puits (HTF-1, HTF-2, HTF-11, HTF-13, HTF-14, HTF-16).

La répartition spatiale des puits au sein du champ est inégale, certains compartiments structuraux étant mieux représentés que d'autres. Cette asymétrie spatiale a des implications importantes pour la stratégie d'apprentissage et d'évaluation des modèles de machine learning, en particulier pour la capacité du modèle à généraliser vers des zones peu représentées dans les données d'entraînement (Vallabhaneni et al., 2019).

I.7. Description des Données de Diagraphies

Chaque fichier de puits contient une colonne de profondeur (DEPH) et jusqu'à 24 courbes de diagraphie brutes. Les diagraphies disponibles pour les puits étudiés comprennent les outils suivants :

- *Rayon Gamma (GR)* : mesure de la radioactivité naturelle pour l'identification de la lithologie et le calcul du volume d'argile (Vcl). C'est l'outil diagnostique le plus universellement disponible dans les suites diagraphiques.
- *Résistivité vraie (RT10 à RT90)* : cinq niveaux d'induction à différentes profondeurs d'investigation, permettant la distinction entre fluides hydrocarbures et eau de formation. Ces courbes sont les entrées primaires pour le calcul de la saturation en eau via la loi d'Archie (2003).

- Densité volumique apparente (RHOB) : utilisée pour le calcul de la porosité densité.
- Porosité neutronique (NPHI) : mesure de l'indice hydrogène, corrélée à la porosité. La combinaison RHOB-NPHI permet de corriger les effets lithologiques et de détecter la présence de gaz.
- Temps de transit sonique de compression (DTC) et de cisaillement (DTS) : utilisés pour l'évaluation de la porosité sonique et de la lithologie.
- Diamètre du puits (CALI) : indicateur de la qualité du trou de forage et des conditions de mesure.
- Spectroscopie gamma naturelle : Potassium (POTA), Uranium (URAN), Thorium (THOR), permettant une identification minéralogique plus fine des argiles.
- Facteur photoélectrique (PE) : utilisé pour l'identification de la lithologie matricielle.

Une analyse détaillée de la disponibilité colonne par colonne a révélé qu'aucune suite diagraphique n'est systématiquement présente dans les 17 puits. Plusieurs puits ne disposent pas de certains outils (par exemple, HTF-01 ne possède pas de diagraphie de résistivité ; HTF-24 ne comporte que GR et DTC). De manière critique, la courbe de porosité interprétée (PORO) n'est présente que dans un seul puits (HTF-19), ce qui a imposé une contrainte fondamentale sur la stratégie de modélisation de ce paramètre.

Well	DEPH	CALI	GR	NPHI	PE	RHOB	DTC	POTA	RT10	RT20	RT30	RT60	RT90	URAN	THOR	RHOP	DTS	BIT	ZDEN	SW	PIGN	PERM	PORO	VCL	GR
HTF-01																									
HTF-10																									
HTF-11																									
HTF-13																									
HTF-14																									
HTF-19																									
HTF-20																									
HTF-25																									
HTF-28																									
HTF-29																									
HTF-30																									
HTF-32																									
HTF-33																									
HTF-36																									
HTF-36																									
HTF-37																									
HTF-38																									
HTF-42																									

Figure 3: Matrice de disponibilité des données de diagraphie au niveau de chaque puit

I.8. Sources et Formats des Données

Les données brutes proviennent des rapports techniques et des bases de données internes de la division Production de Sonatrach. Initialement au format LAS (Log ASCII Standard), elles ont été converties au format CSV (Comma-Separated Values) pour permettre leur traitement par les algorithmes de Machine Learning en Python. Des analyses de carottes (SCAL) issues des puits HTF ont également été utilisées comme référence pour les paramètres cibles (S_w , V_{cl} et ϕ).

Le format LAS est le standard industriel pour l'échange de données de diagraphe de puits. La conversion en CSV a nécessité un traitement particulier des valeurs sentinelles LAS (-999,25) et des métadonnées d'en-tête, détaillé dans la section méthodologique (Pedregosa et al., 2011). Comme le soulignent [Luengo et al. \(2020\)](#), la qualité des données constitue un enjeu majeur pour la fiabilité des modèles de prédiction, le principe « garbage in, garbage out » étant particulièrement pertinent pour les modèles d'apprentissage automatique appliqués aux données de subsurface.

Chapitre II : Contexte Théorique

II.1. Fondamentaux des Paramètres Pétrophysiques

La pétrophysique est la discipline qui étudie les propriétés physiques et chimiques des roches réservoirs en relation avec les fluides qu'elles contiennent (Peters, 2007). Elle constitue le lien entre les données mesurables par les diagraphies et les grandeurs d'ingénierie de réservoir nécessaires à l'estimation des ressources et à la simulation du comportement du réservoir.

II.1.1. La Porosité (Phi)

La porosité est une mesure quantitative de la capacité de stockage des roches poreuses. Physiquement, elle est définie comme le rapport entre le volume des vides (volume poreux) et le volume total de la roche (volume apparent) (Peters, 2007) :

$$\phi = V_{pores} / V_{total}$$

Elle quantifie le pourcentage de vides dans la roche et indique sa capacité à contenir des fluides (eau, pétrole, gaz). Dans les grès cambriens du champ HTF, la porosité est faible (5 à 10 %), limitant le potentiel de stockage des hydrocarbures. On distingue plusieurs types de porosité :

- La porosité absolue (totale), qui inclut tous les vides, connectés ou non.
- La porosité efficace (effective), limitée aux vides interconnectés contribuant à l'écoulement des fluides. C'est cette grandeur qui est pertinente pour l'ingénierie de réservoir.
- La porosité secondaire, résultant de fractures, dissolution ou dolomitisation postérieures au dépôt.

En diagraphie, la porosité est estimée à partir des outils de densité (RHOB), de neutron (NPHI) et de sonique (DTC). Ces estimations sont souvent combinées pour corriger les effets lithologiques et obtenir une porosité neutron-densité plus robuste.

II.1.2. La Perméabilité (K)

La perméabilité est la propriété fondamentale qui détermine la facilité avec laquelle un fluide peut s'écouler à travers un milieu poreux sous l'effet d'un gradient de pression. Elle est quantifiée par la loi de Darcy (1856) :

$$Q = -(K * A / \mu) * (dP/dL)$$

où Q est le débit volumique (m^3/s), K la perméabilité (Darcy ou mD), A la section transversale (m^2), μ la viscosité dynamique du fluide ($mPa \cdot s$) et dP/dL le gradient de pression (Pa/m). Dans le champ HTF, la perméabilité est très faible (1 à 2 mD), ce qui impose des stratégies de stimulation particulières. La perméabilité n'est pas directement mesurable par diagraphie ; elle est estimée de manière indirecte à partir de relations empiriques ou de modèles de machine learning (Bhatt & Helle, 2002).

II.1.3. La Saturation en Eau (S_w)

La saturation en eau est la fraction du volume poreux occupée par l'eau. Elle est définie comme :

$$S_w = V_{eau} / V_{poreux}$$

Dans les zones à hydrocarbures, S_w est faible (20–40 %) tandis que dans les zones aquifères, S_w tend vers 1 (100 %). Son estimation repose classiquement sur la loi d'Archie (1942) :

$$(S_w)^n = (a * R_w) / (R_t * \phi^n)$$

où R_w est la résistivité de l'eau de formation, R_t la résistivité mesurée par diagraphie, a la constante de tortuosité, m l'exposant de cimentation et n l'exposant de saturation (Archie, 1942). Dans les formations argileuses, la loi d'Archie est modifiée (équations de Simandoux, Waxman-Smits, etc.) pour tenir compte de la conduction supplémentaire liée à l'argile.

II.1.4. Le Volume d'Argile (V_{cl} / V_{SH})

Le volume d'argile représente la fraction de la roche constituée de minéraux argileux. Il est estimé principalement à partir de l'indice gamma-ray (IGR) :

$$IGR = (GR_{log} - GR_{min}) / (GR_{max} - GR_{min})$$

Des équations linéaires ou non-linéaires (Steiber, Clavier, Larionov) permettent ensuite de convertir l'IGR en V_{cl} . Le volume d'argile est un paramètre clé car les argiles réduisent la porosité effective, la perméabilité et compliquent l'interprétation des résistivités (Kamayou et al., 2021). Sa quantification précise est donc un prérequis pour l'interprétation correcte des autres paramètres pétrophysiques.

II.2. Techniques d'Apprentissage Automatique Appliquées à la Pétrophysique

II.2.1. Principes Généraux du Machine Learning

L'apprentissage automatique (Machine Learning, ML) désigne un ensemble de méthodes qui permettent à un algorithme d'apprendre des relations entre des variables d'entrée (features) et une ou plusieurs variables cibles (labels) à partir d'un jeu de données d'entraînement, sans être explicitement programmé pour chaque règle de décision ([Ghojogh & Crowley, 2019](#)). En pétrophysique, le ML offre la capacité de capturer des relations non linéaires et multidimensionnelles entre les réponses diagraphiques et les propriétés de réservoir, là où les méthodes déterministes reposent sur des constantes empiriques fixes ([Anifowose et al., 2017](#) ; [Elkatatny et al., 2019](#)).

Les problèmes de prédiction de S_w , V_{cl} et ϕ à partir de diagraphies constituent des tâches de régression supervisée. Le flux de travail typique comprend : collecte et contrôle qualité, analyse exploratoire, prétraitement, division train/test, entraînement, évaluation sur puits aveugles, et interprétation des résultats ([Luengo et al., 2020](#)).

II.2.2. Modèles de Régression Utilisés

Quatre algorithmes représentant un spectre allant de la régression linéaire régularisée aux méthodes d'ensemble les plus avancées ont été sélectionnés pour cette étude.

Régression Ridge : Extension régularisée L2 des moindres carrés ordinaires, pénalisant les grands coefficients de régression pour réduire le surajustement en présence de prédicteurs corrélés ([Hoerl & Kennard, 1970](#)). Elle sert de référence linéaire.

Bagging (Bootstrap AGGREGatING) : Entraîne plusieurs apprenants de base sur des échantillons bootstrap et fait la moyenne de leurs prédictions, réduisant la variance sans augmenter substantiellement le biais ([Breiman, 1996](#)).

Arbres Extrêmement Aléatoires (Extra Trees) : Étendent le bagging en randomisant également le seuil utilisé à chaque division, réduisant davantage la variance et accélérant l'apprentissage ([Geurts et al., 2006](#)). Performances compétitives sur des tâches de régression géophysique ([Bhatt & Helle, 2002](#)).

XGBoost (Extreme Gradient Boosting) : Construit un ensemble additif d'arbres peu profonds par descente de gradient séquentielle ([Chen & Guestrin, 2016](#)). Incorpore des régularisations L1 et L2, un sous-échantillonnage des lignes et des colonnes, et une

expansion de Taylor au second ordre. Résultats de pointe dans de nombreuses études de prédiction pétrophysique (Vallabhaneni et al., 2019 ; Tariq et al., 2022).

II.2.3. Métriques d'Évaluation des Modèles

La performance des modèles est quantifiée à l'aide de quatre métriques complémentaires. Le coefficient de détermination (R^2) mesure la proportion de la variance de la variable cible expliquée par le modèle (valeur parfaite = 1,0 ; valeurs négatives = pire que la moyenne) (Hodson, 2022). La Racine de l'Erreur Quadratique Moyenne (RMSE) pénalise davantage les grandes erreurs que la MAE (Hodson, 2022). L'Erreur Absolue Moyenne (MAE) est la déviation absolue moyenne, robuste aux valeurs aberrantes (Al-Mudhafar, 2017). L'Erreur Absolue Moyenne en Pourcentage (MAPE) facilite la comparaison entre des jeux de données aux plages de valeurs très différentes (Al-Mudhafar, 2017).

II.2.4. Prétraitement des Données

Le prétraitement des données constitue une étape critique dont la qualité conditionne directement la fiabilité des modèles entraînés (Luengo et al., 2020). Les opérations principales comprennent :

- Analyse Exploratoire des Données (AED) et Estimation de Densité par Noyau (EDN) : première étape pour comprendre la structure statistique des données, produisant des courbes de densité lisses indépendantes du choix de l'amplitude des classes.
- Nettoyage des données : imputation des valeurs manquantes par suppression ou estimation statistique ; filtrage du bruit par des algorithmes KNN ou forêts aléatoires.
- Détection des valeurs aberrantes : critères statistiques rigoureux (distance de Mahalanobis, écart interquartile, analyse KDE) pour identifier les régions à très faible densité de probabilité.
- Normalisation (Min-Max ou standardisation Z-score) : restructure les valeurs dans un intervalle uniforme, garantissant l'égalité des chances pour toutes les variables lors de l'apprentissage.
- Sélection de caractéristiques : trois approches, méthodes de filtrage (gain d'information, corrélation), méthodes d'encapsulation (wrapper), méthodes intégrées (arbres de décision, Lasso/Ridge) (Kotsiantis et al., 2006).
- Division train/test basée sur les puits entiers (et non sur les lignes individuelles) pour éviter la contamination entre données d'entraînement et de test (Wood, 2021).

Chapitre III : Méthodologie

III.1. Introduction et Vue d'Ensemble

Ce chapitre décrit la méthodologie complète adoptée pour la prédiction de trois paramètres pétrophysiques clés, la saturation en eau (S_w), le volume d'argile V_{cl} et la porosité (ϕ), à partir des données de diagraphie de puits du champ HTF. L'approche suit un pipeline d'apprentissage automatique supervisé comprenant l'acquisition et le contrôle qualité des données, l'ingénierie des attributs, la sélection des modèles, l'apprentissage et la validation sur puits aveugles.

L'apprentissage automatique s'est imposé comme un complément puissant à l'interprétation pétrophysique conventionnelle, offrant la capacité de capturer des relations non linéaires et multidimensionnelles entre les réponses de diagraphie et les propriétés de réservoir (Anifowose et al., 2017 ; Al-Mudhafar, 2017). La chaîne de traitement a été implémentée en Python et exécutée sur la plateforme de calcul en nuage Kaggle, garantissant la reproductibilité et l'accessibilité des résultats. La Figure 3 présente l'organigramme méthodologique résumant l'ensemble du pipeline.

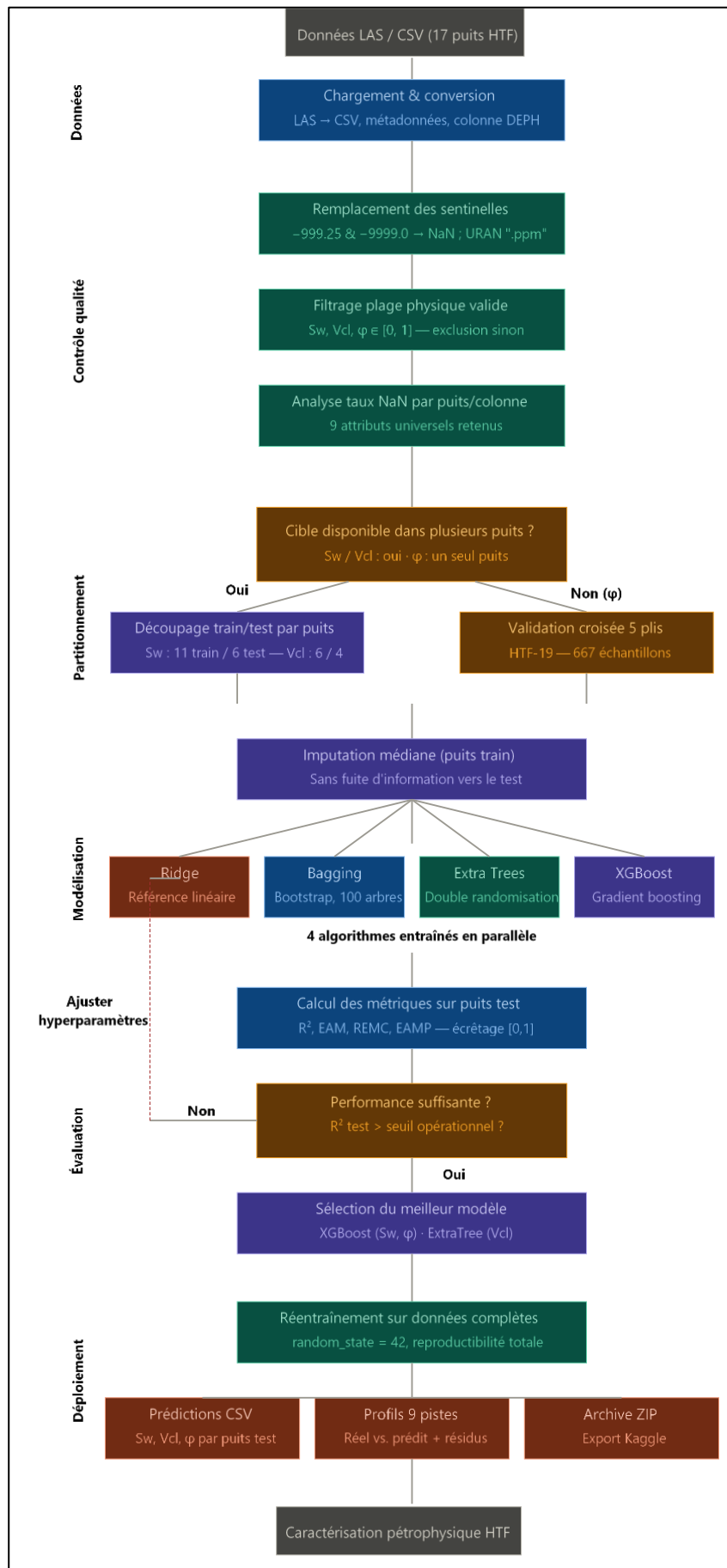


Figure 4: Organigramme de la méthodologie de prédiction pétrophysique par apprentissage automatique dans le champ HTF

III.2.

Description du Jeu de Données

L'étude utilise les données de diagraphie de 17 puits forés dans le champ HTF. Chaque fichier de puits contient une colonne de profondeur (DEPH) et jusqu'à 24 courbes de diagraphie brutes. Les courbes pétrophysiques interprétées (S_w , V_{cl} et ϕ) sont incluses lorsqu'elles sont disponibles. Une analyse détaillée de la disponibilité colonne par colonne a révélé qu'aucune suite diagraphique n'est systématiquement présente dans les 17 puits. De manière critique, la courbe de porosité interprétée (PORO) n'est présente que dans un seul puits (HTF-19), ce qui a imposé une contrainte fondamentale sur la stratégie de modélisation de ce paramètre.

III.3. Contrôle Qualité des Données

III.3.1. Remplacement des Valeurs Sentinelles Nulles

Deux valeurs sentinelles distinctes ont été identifiées dans les fichiers CSV bruts : la valeur nulle LAS standard de -999,25 et une sentinelle secondaire de -9999,0 utilisée dans les colonnes de courbes interprétées. Les deux ont été remplacées par NaN (Not a Number) de Python avant tout traitement ultérieur. L'omission de l'une de ces deux sentinelles était la principale cause d'échec des versions précédentes du pipeline (Pedregosa et al., 2011).

III.3.2. Filtrage par Plage Physique Valide

Les variables cibles (S_w , V_{cl} et ϕ) sont des fractions adimensionnelles physiquement contraintes à l'intervalle $[0, 1]$. Toute ligne dont la valeur cible se situait en dehors de cet intervalle a été exclue de l'apprentissage et de l'évaluation. Cette étape a réduit la taille effective du jeu de données mais a garanti que toutes les étiquettes utilisées correspondaient à des états de réservoir physiquement cohérents.

III.3.3. Disponibilité des Attributs et Imputation

Une analyse du taux de NaN par puits et par colonne a montré que de nombreuses courbes de diagraphie étaient absentes dans la majorité des puits (DTS : 82,5 % de NaN ; RHOB : 44,7 % ; NPHI : 50,5 %). Un ensemble universel de neuf attributs a donc été défini : GR, CALI, RT10, RT20, RT30, RT60, RT90, URAN et THOR. Pour les puits dans lesquels un ou plusieurs de ces attributs étaient manquants, les valeurs absentes ont été imputées par la

médiane colonne calculée exclusivement sur les puits d'apprentissage, garantissant l'absence de fuite d'information (Pedregosa et al., 2011).

III.4. Découpage Train/Test Basé sur les Puits

Une stratégie de découpage basé sur les puits a été adoptée plutôt que le découpage aléatoire classique au niveau des lignes. Avec un découpage aléatoire, des échantillons de profondeur provenant du même puits apparaissent à la fois dans l'ensemble d'apprentissage et dans l'ensemble de test, gonflant artificiellement les performances en test (Vallabhaneni et al., 2029 ; Wood, 2021). Le découpage basé sur les puits affecte des puits entiers soit à l'apprentissage, soit au test, obligeant le modèle à généraliser spatialement. C'est le protocole d'évaluation recommandé pour les applications d'apprentissage automatique en subsurface (Valentin et al., 2020).

Tableau 1: Répartition des puits pour chaque variable cible

Cible	Puits d'apprentissage	Puits de test aveugles	Justification
S_w	HTF-01, 10, 13, 14, 19, 20, 24, 25, 28, 29, 30 (11 puits, 11 774 éch.)	HTF-32, 33, 36, 37, 38, 42 (6 puits, 4 098 éch.)	Couverture maximale
V_{cl}	HTF-01, 10, 13, 14, 24, 30 (6 puits, 8 012 éch.)	HTF-32, 33, 36, 38 (4 puits, 2 990 éch.)	V_{cl} absent dans 7 puits
Porosité	HTF-19 uniquement (667 éch.)	N/A, validation croisée	Un seul puits disponible

Pour la porosité, un schéma de validation croisée k-plis ($k = 5$) a été appliqué (Kohavi, 1995).

III.5. Modèles d'Apprentissage Automatique

III.5.1. Régression Ridge

Régression Ridge avec hyperparamètre $\alpha = 1,0$ et standardisation des entrées via StandardScaler. Sert de référence linéaire (Hoerl & Kennard, 1970).

III.5.2. Bagging

100 estimateurs de base (arbres de décision) entraînés sur des échantillons bootstrap. Réduit la variance sans augmenter substantiellement le biais (Breiman, 1996).

III.5.3. Arbres Extrêmement Aléatoires (Extra Trees)

100 arbres avec double randomisation des attributs et des seuils de division (Geurts et al., 2006), réduisant davantage la variance et accélérant l'apprentissage.

III.5.4. XGBoost

Ensemble additif d'arbres peu profonds par descente de gradient (Chen & Guestrin, 2016).

Hyperparamètres retenus : $n_estimateurs = 300$, taux d'apprentissage = 0,05, profondeur maximale = 6, sous-échantillonnage = 0,8, $colsample_bytree = 0,8$.

III.6. Métriques d'Évaluation des Modèles

La performance des modèles a été quantifiée à l'aide de quatre métriques complémentaires : R^2 , EAM, REMC et EAMP. Pour Sw et Vcl, les métriques ont été calculées sur les puits de test aveugles. Toutes les valeurs prédites ont été écrêtées à [0, 1] avant le calcul des métriques.

Tableau 2: Récapitulatif des performances XGBoost par variable cible

Cible	Modèle	R2 Apprentissage	R2 Test	EAM	Remarques
Sw	XGBoost	0,87	0,40	0,189	Meilleure généralisation sur puits aveugles
Vcl	XGBoost	0,95	0,31	0,087	Bon ajustement ; écart test lié à l'hétérogénéité géologique
Porosité	ExtraTree	VC uniquement	VC uniquement	0,007	Validation croisée 5 plis (un seul puits)

III.7. Déploiement du Modèle et Génération des Sorties

Après évaluation, XGBoost a été réentraîné sur l'ensemble complet des données étiquetées pour maximiser la quantité d'information utilisée. Les prédictions ont ensuite été post-traitées : écrêtage des valeurs à [0, 1] pour garantir la validité physique ; export des fichiers CSV par puits ; génération de profils de diagraphie à neuf pistes pour chaque puits de test aveugle ; et compression des sorties dans une archive ZIP. Le pipeline complet est implémenté sous la forme d'un unique script Python autonome (~520 lignes), avec toutes les graines aléatoires fixées ($random_state = 42$) pour garantir la reproductibilité.

III.8. Limites et Sources d'Incertitude

Plusieurs limites de la méthodologie doivent être soulignées. Premièrement, les courbes diagraphiques interprétées utilisées comme étiquettes d'apprentissage ont elles-mêmes été dérivées d'une interprétation pétrophysique conventionnelle, comportant sa propre incertitude liée aux paramètres d'Archie (a , m , n et R_w). Les modèles sont donc entraînés à

reproduire le résultat d'un interpréteur particulier plutôt qu'une vérité terrain absolue. Deuxièmement, l'ensemble d'attributs restreint (9 diagraphies) a été imposé par la disponibilité des données ; des prédicteurs importants tels que NPHI, RHOB et DTC sont absents dans de nombreux puits. Troisièmement, le modèle de porosité est construit sur un seul puits (HTF-19) et ne peut pas être validé sur des contextes géologiques non vus ([Isah et al., 2025](#)).

Chapitre IV : Résultats et Discussion

IV.1. Introduction

Ce chapitre présente l'ensemble des résultats obtenus à l'issue de l'application du pipeline d'apprentissage automatique décrit au Chapitre III. Les résultats sont organisés en cinq parties : l'analyse exploratoire, les performances comparatives pour Sw et Vcl, la validation croisée pour la porosité, l'analyse spatiale par puits, et une discussion générale.

IV.2. Analyse Exploratoire des Données

IV.2.1. Matrice de Corrélation de Spearman

La Figure 4 présente la matrice de corrélation de Spearman calculée sur les puits d'apprentissage. Les résistivités RT10–RT60 présentent des corrélations négatives significatives avec Sw (de -0,42 à -0,47), conformément à la loi d'Archie (1942) : une formation riche en hydrocarbures (Sw faible) est plus résistante. GR affiche une corrélation positive ($\rho = +0,50$), les formations argileuses étant généralement saturées en eau. La forte colinéarité entre les résistivités (RT10–RT60 : $\rho > 0,92$) justifie l'utilisation de méthodes d'ensemble robustes plutôt que paramétriques (Chen & Guestrin, 2016).

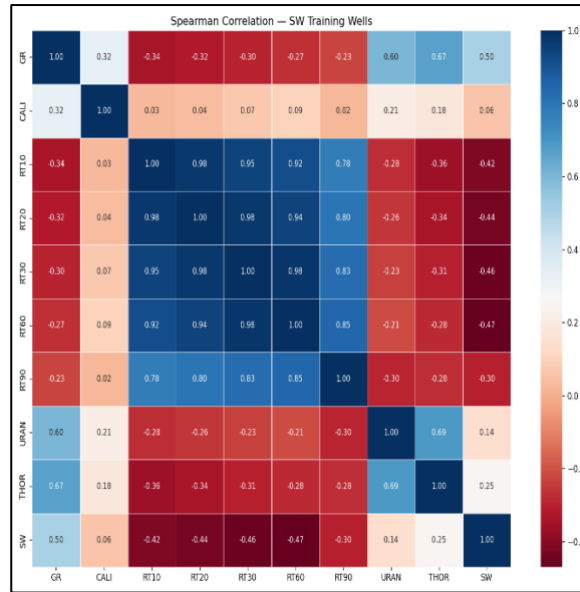


Figure 5: Matrice de corrélation de Spearman (puits d'apprentissage Sw)

IV.2.2. Importance des Attributs

La Figure 5 présente l'importance relative des neuf attributs calculée par Forêt Aléatoire (Breiman, 2001). GR arrive en tête (0,30) devant RT60 (0,26) et CALI (0,13). Cette hiérarchie est physiquement cohérente : GR contrôle la teneur en argile qui détermine le comportement électrique et donc indirectement Sw. Pour Vcl, GR domine encore plus fortement (0,47), ce qui valide la méthode classique de l'index gamma-ray pour la quantification de l'argilosité (Kamayou et al., 2021).

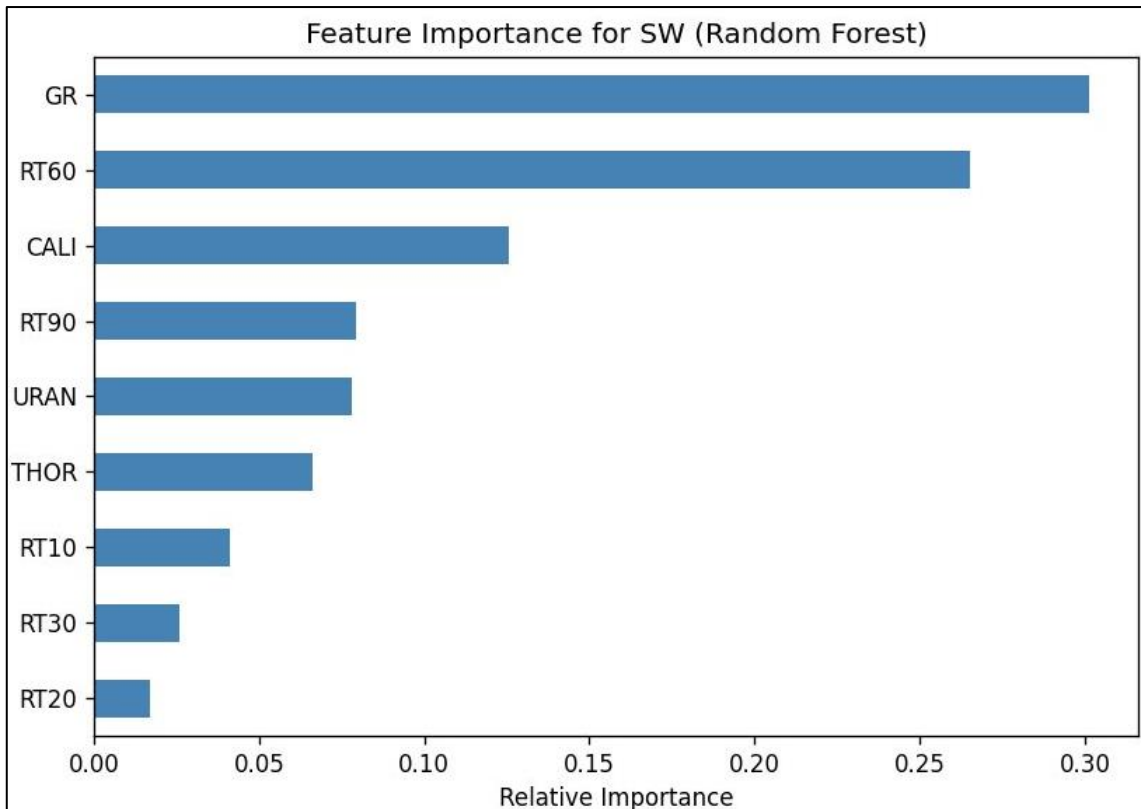


Figure 6: Importance relative des attributs pour la prédiction de Sw (Random Forest)

IV.3. Prédiction de la Saturation en Eau (Sw)

IV.3.1. Comparaison des Modèles

Le Tableau 3 compare les quatre algorithmes sur 4 098 échantillons des six puits de test aveugles. La régression Ridge ($R^2 = 0,29$) confirme l'inadéquation des modèles linéaires pour ce problème fondamentalement non linéaire (Anifowose et al., 2017). XGBoost et ExtraTree obtiennent des R^2 quasi identiques (0,4433 vs 0,4447), mais XGBoost est retenu pour sa meilleure EAM (0,184) et un R^2 d'apprentissage plus équilibré (0,83 vs 1,00), signe d'un surajustement moins prononcé.

Tableau 3: Métriques de performance des quatre modèles pour la prédiction de Sw (6 puits de test aveugles, $n = 4\ 098$)

Modèle	R2 App.	R2 Test	EAM	REMC	Commentaire
Ridge	0,2853	0,2938	0,2407	0,2864	Insuffisant — modèle linéaire inadéquat
Bagging	0,9649	0,3961	0,1925	0,2648	Performances modérées
ExtraTree	1,0000	0,4447	0,1898	0,2539	Meilleur R2 (surajustement fort)
XGBoost	0,8348	0,4433	0,1840	0,2543	Meilleure EAM — modèle retenu

IV.3.2. Distributions Réelles vs. Prédites

La Figure 6 montre que la distribution réelle de Sw est bimodale, avec un pic à $Sw \approx 0,35$ (zone réservoir) et un pic marqué à $Sw = 1,0$ (argiles et aquifères). Les trois modèles d'ensemble reproduisent le premier mode mais atténuent le second, illustrant le biais de compression systématique des méthodes d'ensemble (Breiman, 1996). Pour Vcl (panneau droit), tous les modèles capturent bien le pic principal ($\approx 0,05-0,10$) mais sous-estiment les valeurs élevées.

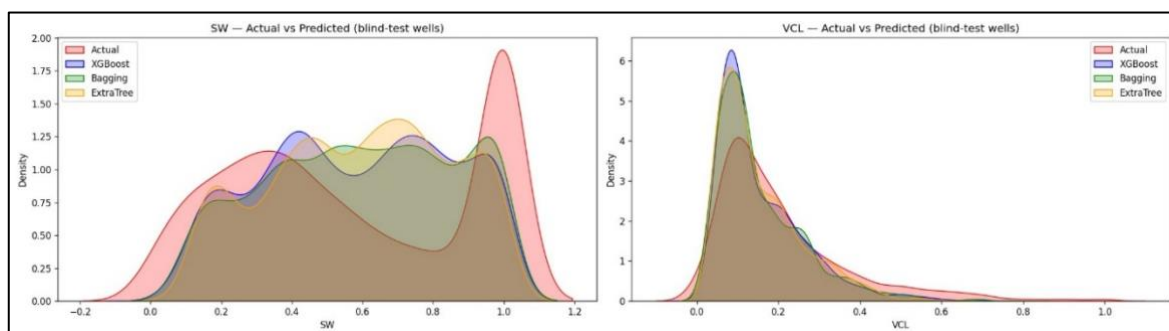


Figure 7: Distributions de Sw (gauche) et Vcl (droite) réelles et prédites (puits de test aveugles)

IV.3.3. Diagramme Croisé

La Figure 8 présente le diagramme jointplot Sw prédite vs. réelle. La droite de régression (pente légèrement < 1) traduit le biais de compression. La concentration de points sur l'axe $Sw = 1,0$ correspond aux zones argileuses où le modèle prédit $0,6-0,9$ au lieu de $1,0$.

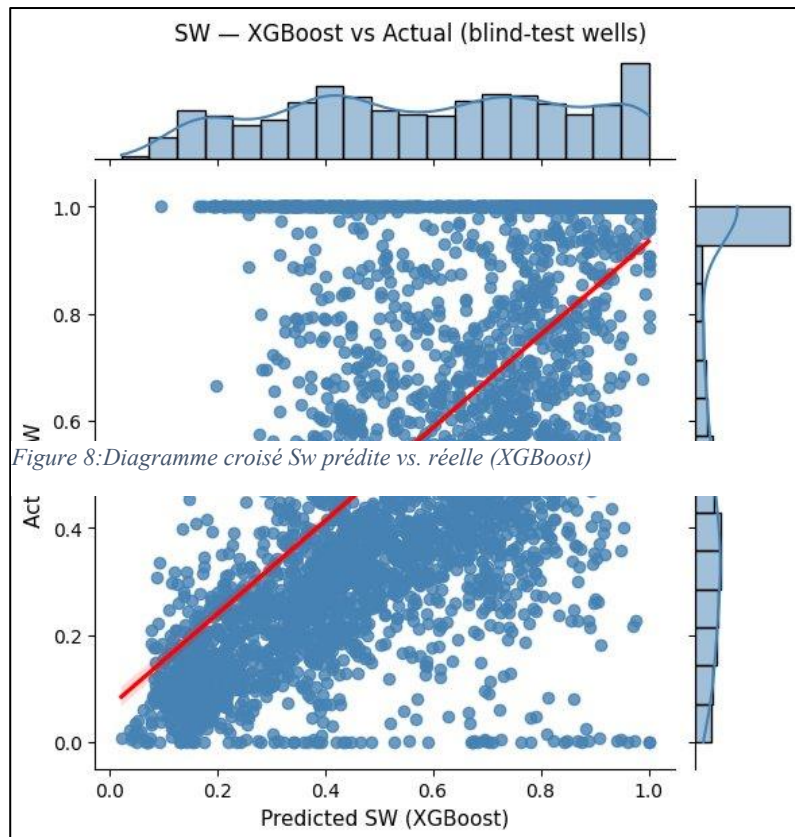


Figure 8: Diagramme croisé Sw prédite vs. réelle (XGBoost)

IV.3.4.

Performance par Puits de Test

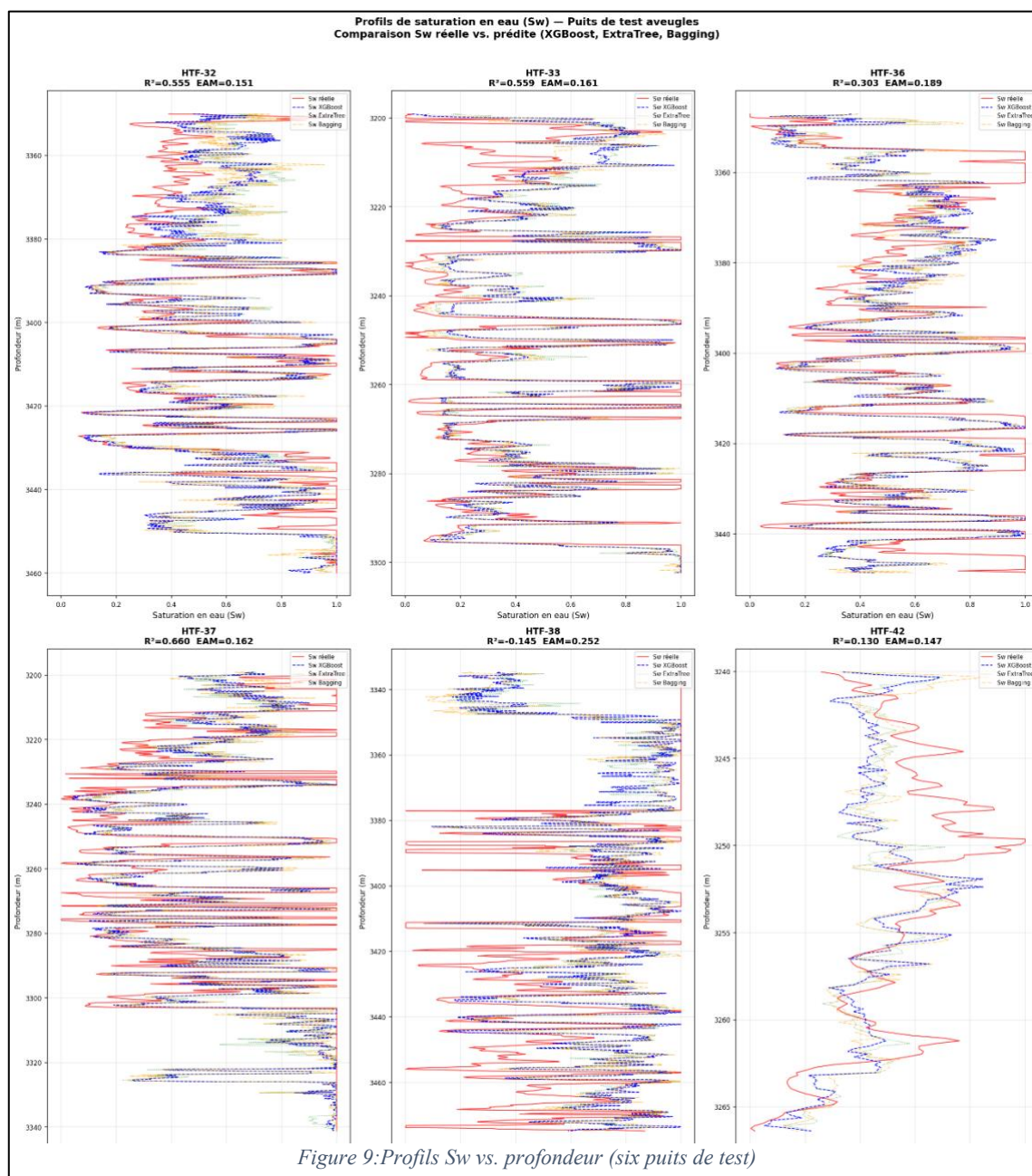
Le Tableau 4 détaille les performances puits par puits. HTF-37 ($R^2 = 0,66$) est le meilleur puits, avec une Sw moyenne prédite (0,603) quasi identique à la réelle (0,600). HTF-38 ($R^2 = -0,15$) présente une performance négative, indiquant des formations non représentées dans le jeu d'apprentissage. HTF-42 souffre d'un nombre d'échantillons insuffisant (174) pour une évaluation fiable.

Tableau 4: Performance XGBoost par puits de test pour la prédiction de Sw

Puits	n	Sw réelle	Sw prédite	R2	EAM	Commentaire
HTF-32	722	0,579	0,582	0,5547	0,1510	Bonne prédiction
HTF-33	679	0,382	0,458	0,5588	0,1614	Bonne prédiction
HTF-36	668	0,578	0,537	0,3029	0,1891	Performance modérée
HTF-37	934	0,600	0,603	0,6602	0,1617	Meilleure performance
HTF-38	921	0,688	0,720	-0,1455	0,2525	Performance négative, faciès atypique
HTF-42	174	0,512	0,440	0,1296	0,1472	Faible, peu d'échantillons

IV.3.5. Profils en Profondeur et Analyse des Résidus

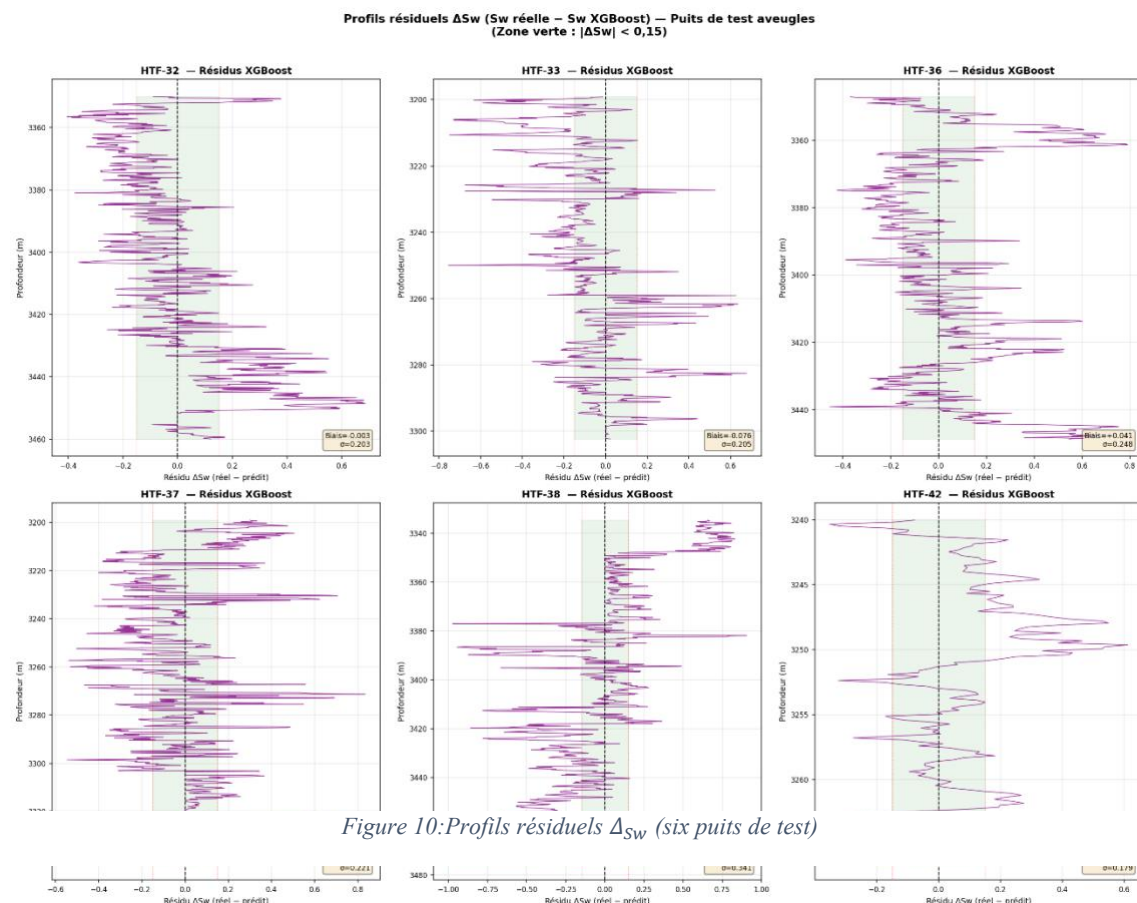
La Figure 9 présente les profils S_w vs. profondeur pour les six puits. Pour HTF-37, le suivi de la courbe réelle par XGBoost est excellent. Pour HTF-38, un décalage systématique est visible à toutes les profondeurs, cohérent avec la performance négative observée. La convergence des trois modèles d'ensemble suggère que les écarts sont liés aux données plutôt qu'aux algorithmes.



La Figure 10 présente les résidus $\Delta_{S_w} = S_{w_{réelle}} - S_{w_{XGBoost}}$. La zone verte indique la plage acceptable ($|\Delta_{S_w}| < 0,15$). Pour HTF-32, 33 et 37, la majorité des résidus s'inscrit

dans cette plage. Pour HTF-38, un biais négatif persistant (surestimation de S_w) est visible. L'absence de dérive avec la profondeur écarte tout artefact de compaction.

IV.4. Prédiction du Volume d'Argile (V_{cl})



IV.4.1. Comparaison des Modèles

Le Tableau 5 présente les résultats pour V_{cl} . Ridge obtient un R^2 négatif (-0,12), reflétant la distribution très asymétrique de V_{cl} combinée à la petite taille du jeu d'apprentissage (6 puits). ExtraTree est retenu comme modèle de référence ($R^2 = 0,41$, EAM = 0,081).

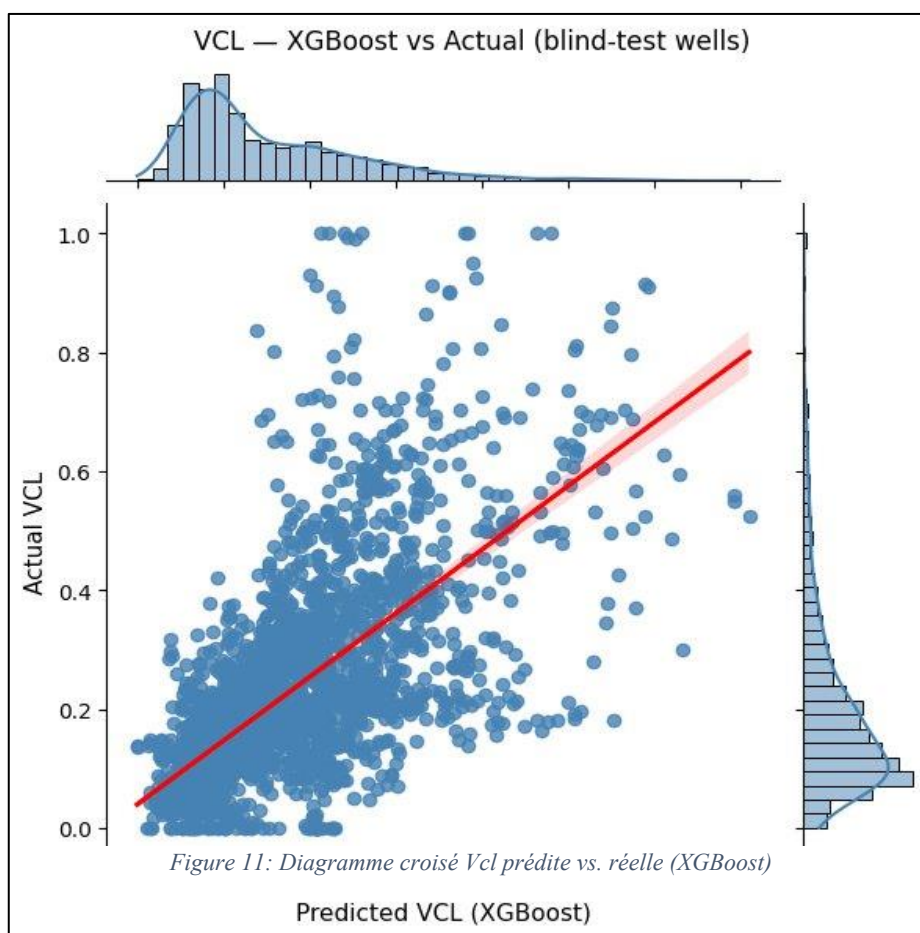
Tableau 5: Métriques de performance pour la prédiction de V_{cl} (4 puits de test aveugles, $n = 2\,990$)

Modèle	R2 App.	R2 Test	EAM	REMC	Commentaire
Ridge	0,5742	-0,1155	0,1122	0,1749	Inadéquat, R^2 négatif

Modèle	R2 App.	R2 Test	EAM	REMC	Commentaire
Bagging	0,9822	0,3698	0,0835	0,1314	Acceptable
ExtraTree	0,9990	0,4112	0,0808	0,1270	Meilleur R ² test , retenu
XGBoost	0,9407	0,3791	0,0839	0,1305	Bon équilibre biais-variance

IV.4.2. Diagramme Croisé et Profils par Puits

La Figure 11 montre une compression encore plus marquée que pour Sw : aucune valeur de Vcl prédite ne dépasse 0,71, alors que des valeurs réelles jusqu'à 1,0 existent. Cette limitation est directement liée à la sous-représentation des faciès fortement argileux dans le jeu d'apprentissage restreint à 6 puits. La sous-estimation systématique observée dans tous les puits (Tableau 6) suggère une composition minéralogique des argiles différente entre puits d'apprentissage et de test (Slatt, 2006).



Résultats et Discussion

Tableau 6: Performance XGBoost par puits de test pour la prédiction de Vcl

Puits	n	Vcl réelle	Vcl prédite	R2	EAM	Commentaire
HTF-32	722	0,210	0,146	0,3783	0,0758	Sous-estimation systématique
HTF-33	679	0,207	0,174	0,6894	0,0569	Meilleure performance
HTF-36	668	0,182	0,104	0,2445	0,0844	Forte sous-estimation
HTF-38	921	0,219	0,186	0,1049	0,1099	Faible — changement de faciès

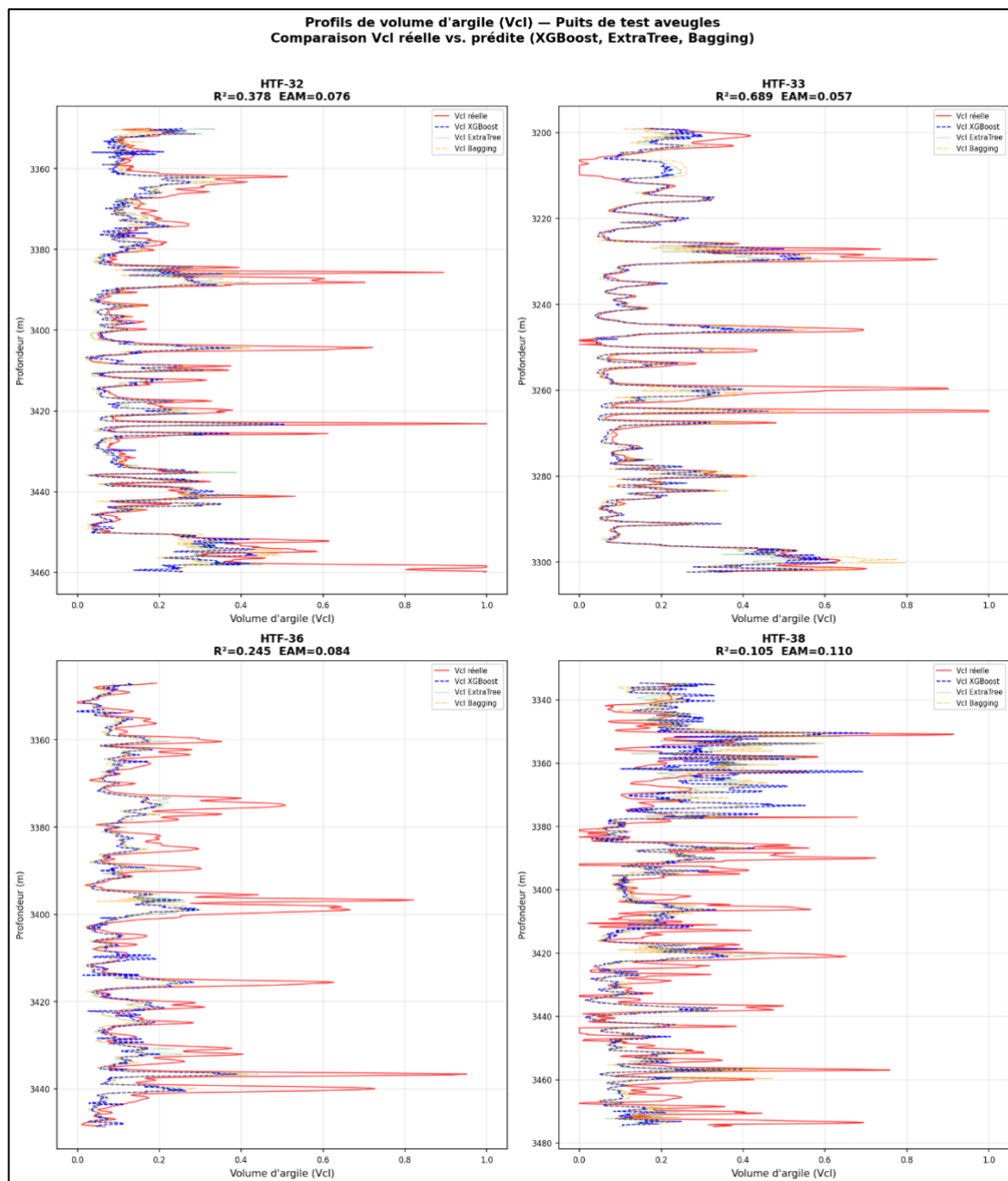


Figure 12: Profils Vcl vs. profondeur (quatre puits de test)

IV.5. Porosité (ϕ) : Validation Croisée sur HTF-19

Les résultats de la validation croisée 5 plis (Tableau 7) sont très élevés : $R^2 CV = 0,9558 \pm 0,0106$ pour ExtraTree et $0,9379 \pm 0,0189$ pour XGBoost, avec une EAM inférieure à 0,7 % de porosité absolue. Ces résultats doivent cependant être interprétés avec prudence : la validation croisée intra-puits mesure l'interpolation entre échantillons auto-corrélés et non la généralisation spatiale (Isah et al., 2025). La validation sur des puits indépendants reste indispensable avant toute application opérationnelle.

Tableau 7: Résultats de la validation croisée 5 plis pour la porosité (HTF-19, $n = 667$)

Modèle	R2 CV moy.	Écart-type	EAM CV moy.	Appréciation
Ridge	0,5079	0,0362	0,02625	Insuffisant
Bagging	0,9365	0,0199	0,00791	Très bon
ExtraTree	0,9558	0,0106	0,00694	Meilleur (faible écart-type)
XGBoost	0,9379	0,0189	0,00782	Très bon

IV.6. Synthèse Spatiale

La Figure 13 compare les distributions de Sw et Vcl réelles (rouge) et prédites (bleu) par puits. La compression des prédictions est visible sur les deux paramètres : les boîtes bleues sont systématiquement plus étroites. Pour Vcl, la sous-estimation dans HTF-36 et HTF-38 est particulièrement prononcée, le 75e percentile prédit n'atteignant pas la médiane réelle.

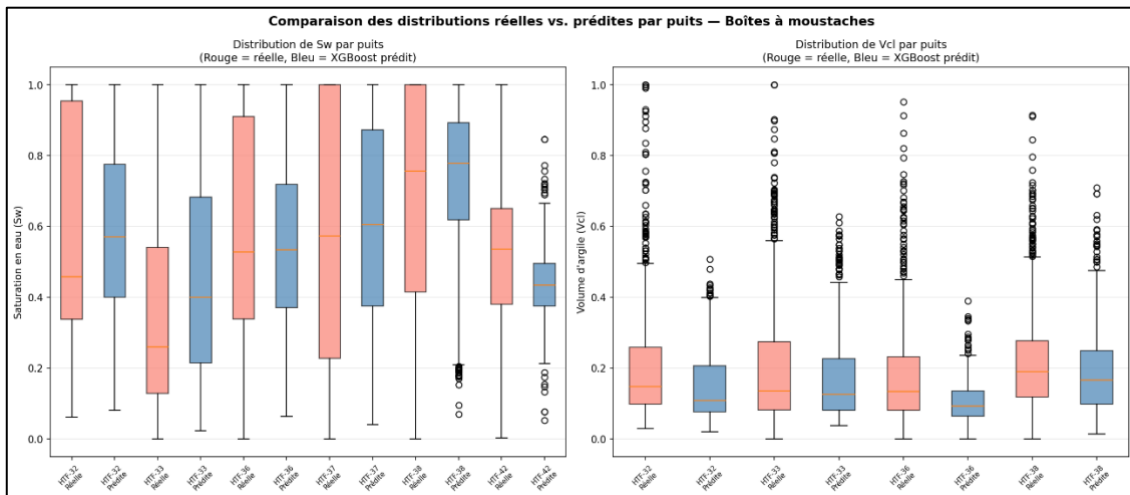


Figure 13: Boîtes à moustaches- Sw et Vcl réelles vs. prédites par puits

La Figure 14 synthétise les R^2 XGBoost par puits avec un code couleur (vert $\geq 0,5$; orange $0,25-0,5$; rouge $< 0,25$ ou négatif). Pour Sw, quatre puits sur six atteignent le niveau acceptable ou fiable. Pour Vcl, HTF-33 est le seul puits fiable. Cette carte de performance oriente directement la stratégie d'utilisation opérationnelle (Ahmed T, 2018).

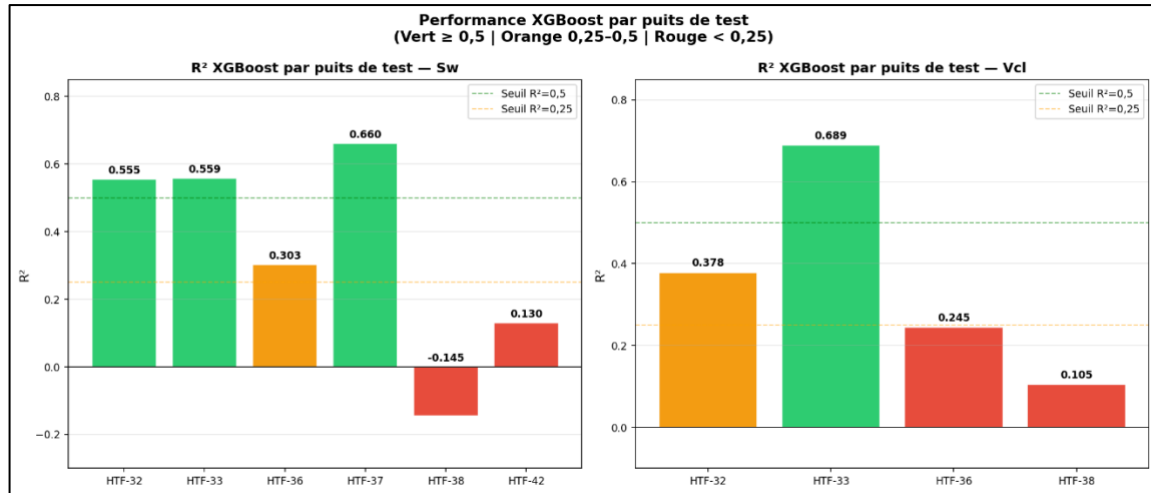


Figure 14: R^2 XGBoost par puits de test pour Sw et Vcl avec code couleur

IV.7. Discussion Générale

IV.7.1. Performances Globales et Benchmark Littérature

Les R^2 de test sur puits aveugles (0,44 pour Sw, 0,41 pour Vcl) sont inférieurs aux valeurs de la littérature (0,75–0,95). Cet écart s'explique principalement par le protocole d'évaluation : la plupart des études publiées utilisent un découpage aléatoire des lignes qui surestime les performances en testant l'interpolation intra-puits (Wood, 2021). Notre protocole sur puits aveugles est scientifiquement plus rigoureux et mesure la véritable généralisation spatiale. L'ensemble d'attributs restreint (9 vs 15–20 dans les meilleures études) constitue la seconde limitation principale ; des gains de R^2 de 0,15–0,25 sont attendus avec l'ajout de NPHI et RHOB.

IV.7.2. Cohérence Physique

Plusieurs éléments attestent positivement de la cohérence physique des modèles. La hiérarchie d'importance est physiquement interprétable : RT60 domine pour Sw (physique d'Archie), GR domine pour Vcl (index gamma-ray, Kamayou et al., 2021). Les corrélations ont les signes attendus : négatif pour la résistivité vs Sw, positif pour GR vs Sw. Les valeurs prédites restent systématiquement dans la plage physique $[0, 1]$ et aucune dérive résiduelle avec la profondeur n'est observée.

IV.7.3. Recommandations

Sur la base des résultats obtenus, les recommandations suivantes sont formulées pour améliorer les performances :

- Interpréter Vcl dans les 7 puits où elle est absente : mesure la plus impactante à court terme, doublant le jeu d'apprentissage Vcl.
- Acquérir NPHI et RHOB systématiquement dans les futurs puits : prédicteurs manquants les plus critiques
- Mener une étude pétrographique de HTF-38 : comprendre la particularité lithologique avant intégration dans un futur apprentissage.
- Optimisation des hyperparamètres (Optuna ou GridSearch) : gain estimé de 3–7 % sur R^2 .
- Estimation d'incertitude (forêts quantiles, intervalles de confiance XGBoost) pour une utilisation opérationnelle sécurisée ([Amjad et al., 2025](#)).

IV.8. Synthèse Finale

Le Tableau 8 synthétise les performances et la fiabilité opérationnelle pour chaque paramètre cible. XGBoost est retenu pour Sw et la porosité ; ExtraTree est préféré pour Vcl. Les prédictions sont directement utilisables pour HTF-32, HTF-33 et HTF-37 (Sw) et HTF-33 (Vcl).

Tableau 8: Synthèse des performances et niveaux de fiabilité opérationnelle

Cible	Meilleur modèle	R2 test	EAM	Fiabilité opérationnelle
Sw	XGBoost	0,4433	0,1840	Modérée : HTF-32, 33, 37 fiables ; HTF-38, 42 à valider
Vcl	ExtraTree	0,4112	0,0808	Modérée : HTF-33 fiable ; HTF-36, 38 à valider
Porosité	ExtraTree	0,956 (VC)	0,00694	Haute intra-puits (HTF-19) : non validée spatialement

Conclusion Générale

Cette étude a permis d'évaluer rigoureusement la performance des modèles d'apprentissage automatique pour la prédiction des paramètres pétrophysiques clés, porosité (ϕ), volume d'argile (Vcl) et saturation en eau (Sw), à partir des données de diagraphies du champ de Hassi Tarfa.

L'analyse comparative montre que les modèles d'ensemble, en particulier XGBoost ([Chen & Guestrin, 2016](#)) et les arbres extrêmement aléatoires ([Geurts et al., 2006](#)), surclassent systématiquement les approches linéaires (régression Ridge). Cette supériorité s'explique par la capacité intrinsèque de ces modèles à capter les relations hautement non linéaires et à s'adapter à la forte hétérogénéité lithologique des réservoirs gréseux du Cambro-Ordovicien.

Le protocole d'évaluation sur puits aveugles, plus rigoureux que le découpage aléatoire classique ([Wood, 2021](#)), a produit des R^2 de test de 0,44 pour Sw et 0,41 pour Vcl. Bien qu'inférieurs aux valeurs de la littérature utilisant des protocoles moins stricts, ces résultats reflètent une véritable capacité de généralisation spatiale. La porosité, limitée à un seul puits d'apprentissage (HTF-19), affiche un excellent R^2 en validation croisée (0,96), mais nécessite une validation sur des puits indépendants ([Isah et al., 2025](#)).

Sur le plan pratique, les prédictions sont immédiatement utilisables pour HTF-32, HTF-33 et HTF-37 (Sw) et HTF-33 (Vcl). Le puits HTF-38 présente un comportement atypique nécessitant une étude pétrographique approfondie avant intégration.

Pour pérenniser et enrichir cette démarche, les travaux futurs devraient :

- Étendre la méthodologie à un plus grand nombre de puits régionaux pour tester le pouvoir de généralisation des modèles.
- Intégrer des variables diagraphiques manquantes (NPHI, RHOB) pour améliorer les performances de 15 à 25 points de R^2
- Incorporer des données dynamiques de carottage (SCAL) et les pressions capillaires pour une meilleure contrainte des paramètres cibles.
- Explorer des modèles hybrides guidés par la physique (Physics-Informed Machine Learning) pour allier la puissance du calcul statistique à la rigueur des lois de la physique des roches.
- Mettre en oeuvre des méthodes d'estimation d'incertitude (forêts quantiles, intervalles de prédiction) pour sécuriser l'utilisation opérationnelle des prédictions

Cette étude contribue ainsi au développement des méthodes d'interprétation pétrophysique automatisée, offrant des perspectives prometteuses pour la caractérisation des réservoirs algériens et la valorisation du capital de données diagraphiques disponibles.

Références Bibliographiques

- Ahmed, T. (2018). Reservoir engineering handbook. Gulf professional publishing.
- Al-Mudhafar, W. J. (2017). Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. *Journal of Petroleum Exploration and Production Technology*, 7(4), 1023-1033.
- Amjad, M. R., Varghese, R. B., & Amjad, T. (2025). Machine Learning Models for Subsurface Pressure Prediction: A Data Mining Approach. *Computers*, 14(11), 499.
- Anifowose, F. A., Labadin, J., & Abdulraheem, A. (2017). Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. *Journal of Petroleum Science and Engineering*, 151, 480-487.
- Archie, G. E. (2003). The electrical resistivity log as an aid in determining some reservoir characteristics. *SPE reprint series*, 9-18.
- Bhatt, A., & Helle, H. B. (2002). Committee neural networks for porosity and permeability prediction from well logs. *Geophysical prospecting*, 50(6), 645-660.
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Elkatatny, S., Tariq, Z., Mahmoud, M., & Abdulraheem, A. (2018). New insights into porosity determination using artificial intelligence techniques for carbonate reservoirs. *Petroleum*, 4(4), 408-418.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1), 3-42.
- Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.
- Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions*, 2022, 1-10.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Isah, A., Tariq, Z., Mustafa, A., Mahmoud, M., & Okoroafor, E. R. (2025). A review of data-driven machine learning applications in reservoir petrophysics. *Arabian Journal for Science and Engineering*, 50(24), 20343-20377.
- Kamayou, V. M., Ehirim, C. N., & Ikiensikimama, S. S. (2021). Estimating volume of Shale in a Clastic Niger Delta reservoir from well Logs: a comparative study. *International Journal of Geosciences*, 12(10), 949-959.

- Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).
- Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International journal of computer science*, 1(2), 111-117.
- Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). Big data preprocessing. *Cham: Springer*, 1, 1-186.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peters, E. J. (2012). *Advanced Petrophysics: Geology, porosity, absolute permeability, heterogeneity, and geostatistics* (Vol. 1). Greenleaf Book Group.
- Slatt, R. M. (2006). *Stratigraphic reservoir characterization for petroleum geologists, geophysicists, and engineers* (Vol. 61). Elsevier.
- Stadtmüller, M., & Jarzyna, J. A. (2023). Estimation of petrophysical parameters of carbonates based on well logs and laboratory measurements, a review. *Energies*, 16(10), 4215.
- Tariq, Z., Elkatatny, S., Mahmoud, M., Ali, A. Z., & Abdulraheem, A. (2017, April). A new approach to predict failure parameters of carbonate rocks using artificial intelligence tools. In *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition* (p. D043S041R003). SPE.
- Valentín, M. B., Bom, C. R., Coelho, J. M., Correia, M. D., De Albuquerque, M. P., de Albuquerque, M. P., & Faria, E. L. (2019). A deep residual convolutional neural network for automatic lithological facies identification in Brazilian pre-salt oilfield wellbore image logs. *Journal of Petroleum Science and Engineering*, 179, 474-503.
- Vallabhaneni, S., Saraf, R., & Priyadarshy, S. (2019, June). Machine-learning-based petrophysical property modeling. In *SPE Europec featured at EAGE Conference and Exhibition* (p. D032S013R007). SPE.
- Wood, D. A. (2025). Extracting useful information from sparsely logged wellbores for improved rock typing of heterogeneous reservoir characterization using well-log attributes, feature influence and optimization. *Petroleum Science*, 22(6), 2307-2311.