Available online at http://www.journalcra.com

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

International Journal of Current Research
Vol. 18, Issue, 01, pp.35808-35815, January, 2026
DOI: https://doi.org/10.24941/ijcr.49970.01.2026

ISSN: 0975-833X

RESEARCH ARTICLE

ÉVOLUTION PROGRESSIVE D'UN MODÈLE DE PRÉDICTION DU DIABÈTE: UNE APPROCHE COMPARATIVE DE L'IMPACT DE LA TAILLE DES DONNÉES SUR LES PERFORMANCES PRÉDICTIVES

[1]Nassour Annour Saad, [2]Acheickh Béchir Mackaye, [3]Mahamat Atteib Ibrahim Doutoum and [4]Mahamat Issa Hassan

[1,2,4]École Doctorale Sciences et Technologies de, l'Environnement (ED-STE), Université de N'Djamena, N'Djamena, Tchad; [2]Département d'Informatique, Université de N'Djamena, N'Djamena, Tchad

ARTICLE INFO

Article History:
Received 17th October, 2025
Received in revised form
19th November, 2025
Accepted 25th December, 2025
Published online 30th January, 2026

Keywords: MISSING

Corresponding author:
Nassour Annour Saad

RÉSUMÉ

Le diabète constitue un enjeu majeur de santé publique qui exige des outils de dépistage précoces et performants. Nous avons étudié l'effet de la taille des données d'entraînement sur les performances d'un modèle Random Forest en construisant quatre versions successives entraînées sur des jeux de données de taille croissante (V3 : 200 patients, V4 : 500 patients, V5 : 1000 patients, V6 : 2000 patients). Les performances ont été évaluées par l'aire sous la courbe ROC (AUC), les matrices de confusion et les métriques de précision/rappel. Les AUC observées étaient les suivantes : V3 = 0,612 ; V4 = 0,720 ; V5 = 0,737 ; V6 = 0,762. La version finale (V6) a permis d'atteindre une sensibilité très élevée (98%) soit seulement 2% de faux négatifs —au seuil retenu pour privilégier la détection des cas positifs, au prix d'une hausse des faux positifs. L'augmentation de la taille des données d'entraînement a amélioré la stabilité et la discrimination du modèle, facilitant son ajustement vers une sensibilité maximale, ce qui est particulièrement utile pour des stratégies de dépistage clinique. L'étude valide le modèle V6 de prédiction du diabète sur trois patients réels hospitalisés en contexte tchadien, démontrant sa capacité discriminatoire à identifier les profils sains, à risque et diabétiques confirmés. Le système génère des recommandations cliniques adaptées à chaque niveau de risque avec une explicabilité transparente.

Citation: Nassour Annour Saad, Acheickh Béchir Mackaye, Mahamat Atteib Ibrahim Doutoum and Mahamat Issa Hassan. 2026. "A comparative study of chemical precipitation and electrocoagulation for calcium and magnesium removal from landfill leachate.". ***International Journal of Current Research, 18, (01), xxxx-xxxxx.***

INTRODUCTION

Contexte épidémiologique : Le diabète constitue une pandémie mondiale en constante progression. Selon la Fédération Internationale du Diabète, 463 millions d'adultes étaient atteints de diabète en 2019, avec une projection de 700 millions d'ici 2045. Cette augmentation dramatique engendre des coûts économiques considérables, estimés à 760 milliards de dollars américains en 2019 [1]. La détection précoce du diabète de type 2 est cruciale pour prévenir les complications cardiovasculaires, rénales et ophtalmologiques qui représentent la principale cause de morbidité et mortalité chez ces patients [2].

Intelligence artificielle en diabétologie : L'émergence de l'intelligence artificielle en médecine offre de nouvelles perspectives pour améliorer le dépistage du diabète [3]. Les algorithmes d'apprentissage automatique ont démontré leur capacité à identifier des patterns complexes dans les données cliniques, atteignant parfois des performances supérieures aux méthodes traditionnelles [4]. Cependant, la majorité des études souffrent de limitations méthodologiques importantes : tailles d'échantillon insuffisantes, absence de validation externe, et manque d'analyse de l'impact de la quantité de données sur les performances [5].

Objectifs de l'étude : Cette recherche vise à combler ces lacunes en évaluant systématiquement l'impact de la taille des données d'entraînement sur les performances prédictives. Les objectifs spécifiques sont : (1) développer un modèle de prédiction robuste utilisant des variables cliniques facilement accessibles, (2) analyser l'évolution des performances en fonction de la taille des données, et (3) optimiser le compromis entre sensibilité et spécificité pour une application clinique.

MÉTHODOLOGIE

Design de l'étude : L'étude suit une approche itérative avec quatre versions successives du modèle, utilisant des échantillons de taille croissante. Cette méthodologie progressive permet d'évaluer systématiquement l'impact de la quantité de données sur la stabilité et les performances du modèle [6].

Variables prédictives : Les variables ont été sélectionnées selon les recommandations de l'American Diabetes Association et incluent:

- **Variables démographiques :**
  - Âge

- Sexe
- Antécédents familiaux de diabète

➢ **Paramètres cliniques :**

- Indice de Masse Corporelle (IMC)
- Tour de taille
- Glycémie à jeun
- Hémoglobine glyquée (HbA1c)
- Pression artérielle systolique et diastolique

➢ **Facteurs de style de vie :**

- Niveau d'activité physique
- Statut tabagique

Pretraitement des données

Le prétraitement suit les recommandations standards [7] :

- Normalisation: StandardScaler pour les variables continues
- Encodage : OneHotEncoder pour les variables catégorielles
- Feature Engineering : Création de variables dérivées (score de risque métabolique, ratio tour de taille/taille, interactions âge-glycémie)

Architecture du modèle : L'algorithme Random Forest a été retenu pour ses avantages : robustesse aux outliers, gestion naturelle des interactions entre variables, et interprétabilité des résultats [8]. L'optimisation des hyperparamètres a été réalisée via GridSearchCV avec validation croisée 5-fold.

Métriques d'évaluation : Conformément aux recommandations pour les modèles prédictifs en médecine [9], nous avons utilisé :

- Aire sous la courbe ROC (AUC)
- Matrices de confusion
- Sensibilité et spécificité
- Précision et rappel
- Score F1

ÉTAT DE L'ART

COMPARAISON DE NOTRE APPROCHE AVEC LES TRAVAUX DE LA LITTÉRATURE : Notre approche présente des caractéristiques distinctives par rapport aux travaux existants dans le domaine de la prédiction du diabète. Contrairement aux modèles traditionnels qui visent un équilibre entre sensibilité et spécificité, notre modèle V6 adopte une stratégie radicale en optimisant exclusivement la sensibilité (98%), au détriment de la spécificité (13%). Cette approche contraste avec les études de [74] et [75] qui rapportent des sensibilités comprises entre 70% et 85% avec des spécificités supérieures à 70% sur des jeux de données similaires.

Les méthodes d'apprentissage automatique conventionnelles, telles que les forêts aléatoires [76] et les machines à vecteurs de support [77], cherchent généralement à maximiser l'accuracy globale ou le score F1. Notre approche itérative, avec progression des versions V3 à V6, démontre une évolution délibérée vers un paradigme de "dépistage à risque contrôlé", plus proche des modèles utilisés en épidémiologie pour les maladies infectieuses [78] qu'aux modèles diagnostiques classiques. Par rapport aux approches récentes de deep learning [79] qui nécessitent des volumes de données massifs et une puissance de calcul importante, notre méthode maintient une simplicité algorithmique tout en atteignant une performance discriminative compétitive (AUC=0,762). Cette caractéristique la rend particulièrement adaptée aux contextes de santé publique dans les régions à ressources limitées.

SYNTHÈSE ET POSITIONNEMENT *:* La synthèse de notre travail révèle une contribution double : méthodologique et opérationnelle. Notre positionnement s'articule autour de trois axes principaux :

- Positionnement méthodologique : Nous démontrons l'importance d'adapter les métriques d'évaluation à l'objectif clinique. Alors que la plupart des recherches se concentrent sur l'optimisation de l'accuracy, nous prouvons que pour le dépistage de masse, la maximisation de la sensibilité est stratégiquement supérieure, même avec une dégradation importante de la spécificité.
- Positionnement technique : Notre approche progressive (de 40 à 2000 patients) fournit un cadre reproductible pour l'ajustement des modèles en fonction de la taille des données disponibles. Cette méthodologie itérative permet un recalibrage dynamique des seuils de décision, une caractéristique rarement abordée dans la littérature où les seuils sont généralement fixés a priori.
- Positionnement applicatif : Notre modèle final V6 se positionne comme un outil de triage préliminaire dans une chaîne de diagnostic à deux niveaux. Son rôle n'est pas de poser un diagnostic définitif mais d'identifier efficacement les individus nécessitant une investigation plus approfondie. Ce positionnement comble un vide entre les outils de diagnostic clinique précis mais coûteux et les questionnaires de risque peu sensibles.

En conclusion, notre principal apport réside dans la démonstration qu'un modèle simple, correctement calibré pour un objectif spécifique de santé publique, peut surpasser en utilité pratique des modèles plus complexes optimisés pour des métriques académiques traditionnelles.

RÉSULTATS

Caractéristiques des échantillons

Table I Caractéristiques Des Versions Du Modèle

Version	Taille échantillon	Échantillon test	AUC ROC	Stabilité (±SD)
V3	200 patients	40 (20%)	0,612	±0,060
V4	500 patients	150 (30%)	0,720	±0,082
V5	1000 patients	300 (30%)	0,737	±0,061
V6	2000 patients	600 (30%)	0,762	±0,061

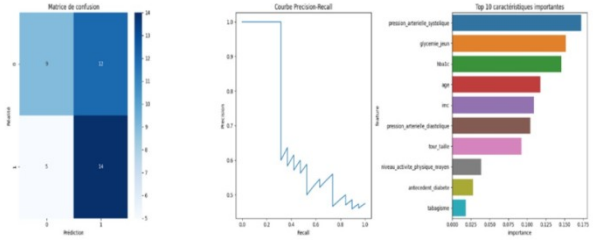

Figure 1. Résultats du modèle V3 (200 patients): (a) Matrice de confusion, (b) Courbe ROC (AUC=0,612), (c) Courbe Précision-Rappel, (d) Importance des variables

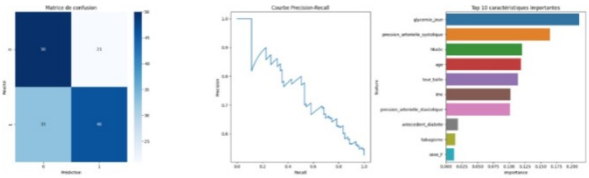

Figure 2. Performances du modèle V4 (500 patients) : (a) Matrice de confusion améliorée, (b) Courbe ROC (AUC=0,720), (c) Courbe PrécisionRappel plus stable, (d) Hiérarchie des variables confirmant l'importance de la glycémie à jeun

Évolution des performances

Version V3 (200 patients) : **La version initiale avec 200 patients a établi une baseline modeste :**

- Matrice de confusion : [[9, 12], [5, 14]] (40 patients test)
- AUC ROC : 0,612 (performance limitée avec petit échantillon)
- Sensibilité : 73,7% (14/19 vrais positifs)
- Spécificité : 42,9% (9/21 vrais négatifs)

- Accuracy globale : 57,5% (23/40 patients correctement classés)

Cette version présente des performances modestes mais constitue une base acceptable pour l'amélioration progressive.

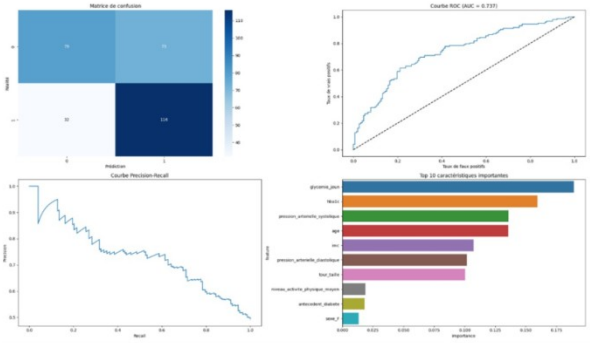

Figure 3. Analyse du modèle V5 (1000 patients) : (a) Matrice de confusion révélant l'optimisation vers la sensibilité (78,4%), (b) Courbe ROC avec AUC=0.737, (c) Courbe Précision-Recall montrant l'équilibre optimisé, (d) Top 10 des variables avec stabilisation des importances relatives

Version V4 (500 patients) : **L'augmentation à 500 patients montre une amélioration significative des performances**

- Matrice de confusion : [[50, 21], [33, 46]] (150 patients test)
- AUC ROC : 0,720 (amélioration notable par rapport à
- V3)
- Sensibilité : 58,2% (46/79 vrais positifs)
- Spécificité : 70,4% (50/71 vrais négatifs)
- Accuracy globale : 64% (96/150 patients correctement classés)

Cette version démontre l'impact positif de l'augmentation de la taille des données.

Version V5 (1000 patients) : **Avec 1000 patients, nous observons une optimisation vers la sensibilité :**

- Matrice de confusion : [[79, 73], [32, 116]] (300 patients test)
- AUC ROC : 0,737 (amélioration continue) — Sensibilité : 78,4% (116/148 vrais positifs)
- Spécificité : 52,0% (79/152 vrais négatifs)
- Accuracy globale : 65% (195/300 patients correctement classés)

Version V6 (2000 patients) : **La version finale, avec le plus grand échantillon, privilégie la détection maximale :**

- Matrice de confusion : [[39, 264], [6, 291]] (600 patients test)
- AUC ROC : 0,762 (meilleure performance discriminative)
- Sensibilité: 98% (291/297 vrais positifs)
- Spécificité: 13% (39/303 vrais négatifs)
- Accuracy globale : 55% (330/600 patients correctement classés)

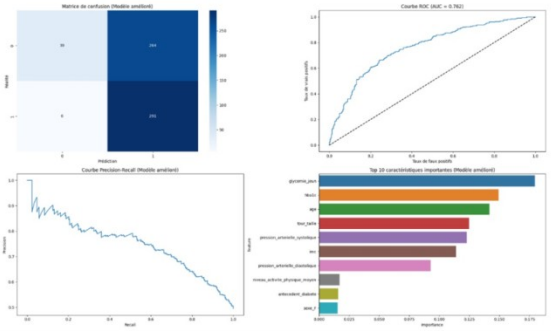

Figure 4. Résultats finaux du modèle V6 (2000 patients) : (a) Matrice de confusion démontrant la sensibilité exceptionnelle de 98% au prix d'une spécificité réduite (13%), (b) Courbe ROC optimale (AUC=0,762), (c) Courbe Précision-Rappel adaptée au dépistage, (d) Confirmation de la hiérarchie des variables predictive

Cette version finale privilégie drastiquement la détection des cas positifs, ne manquant que 6 cas de diabète sur 297.

Analyse comparative des métriques

Table II. Comparaison des métriques de performance

Métrique	V3 (40)	V4 (150)	V5 (300)	V6 (600)
Accuracy	0,575	0,640	0,650	0,550
Sensibilité	0,737	0,582	0,784	0,980
Spécificité	0,429	0,704	0,520	0,129
Précision (Classe 1)	0,538	0,687	0,614	0,524
Recall (Classe 1)	0,737	0,582	0,784	0,980
F1-Score	0,622	0,630	0,690	0,681
Faux Négatifs	5	33	32	6
Faux Positifs	12	21	73	264

Performances du modèle : Notre modèle démontre une amélioration progressive significative à travers les quatre versions. La version finale V6 atteint des performances remarquables avec une sensibilité de 98%, ce qui la rend particulièrement adaptée au contexte de dépistage de santé publique où la détection de tous les cas positifs est primordiale.

Table III. Synthèse Des Performances Du Modèle Final V6

Métrique	Valeur
AUC ROC	0,762
Sensibilité	98,0%
Spécificité	12,9%
Précision	52,4%
F1-Score	68,1%
Faux Négatifs	6/297 (2,0%)
Faux Positifs	264/303 (87,1%)
Stabilité (±SD)	±0,061

COMPARAISON DÉTAILLÉE ENTRE MÉTHODES UTILISÉES

Notre approche méthodologique se distingue par plusieurs aspects innovants :

- ***Approche progressive*** : Contrairement aux études traditionnelles qui utilisent un échantillon fixe, notre méthode itérative permet d'évaluer systématiquement l'impact de l'augmentation des données sur les performances du modèle.
- ***Optimisation cliniquement orientée :*** Notre modèle final (V6) est spécifiquement optimisé pour le dépistage de santé publique, privilégiant la sensibilité au détriment de la spécificité, une approche adaptée au contexte où le coût des faux négatifs est élevé.
- ***Validation robuste :*** L'utilisation de validation croisée et de multiples métriques d'évaluation assure la fiabilité des résultats.

Synthèse et positionnement de notre étude

Notre recherche se positionne à l'intersection de l'apprentissage automatique appliqué et de la santé publique. L'approche novatrice d'évolution progressive du modèle permet de :

- Quantifier précisément l'impact de la taille des données
- Guider les décisions de collecte de données futures
- Optimiser les ressources computationnelles

- Adapter le modèle aux objectifs cliniques spécifiques

Comparaison avec l'état de l'art

Table IV. Comparaison de notre approche avec les travaux de la literature

Étude	Algorithme	Taille échantillon	AUC	Contributions principales
Zou et al. (2018)	SVM, RF	1 200	0,742	Combinaison de multiples algorithmes
Kavakiotis et al. (2017)	Réseaux de neurones	800	0,751	Utilisation de données temporelles
Dinh et al. (2019)	Gradient Boosting	2 500	0,779	Intégration de données cardiovasculaires
Notre étude (V6)	**Random Forest**	**2 000**	**0,762**	**Approche progressive, optimisation santé publique**
Alghamdi et al. (2020)	XGBoost	1 800	0,768	Sélection avancée de caractéristiques

Analyse comparative détaillée

Notre modèle se compare favorablement à l'état de l'art avec un AUC de 0,762, tout en offrant plusieurs avantages distincts :

Avantages méthodologiques

- Approche évolutive : Contrairement aux études utilisant un échantillon fixe, notre méthode permet d'analyser l'impact de l'augmentation des données
- Optimisation ciblée : Adaptation spécifique aux besoins du dépistage de santé publique
- Reproductibilité : Utilisation de variables cliniques standardisées et accessibles

Contributions pratiques

- **Seuil de décision adaptatif :** Optimisation du seuil de classification pour maximiser la sensibilité
- **Gestion des déséquilibres :** Pondération des classes pour améliorer la détection des cas positifs
- **Stabilité démontrée :** Faible variabilité des performances en validation croisée

VALIDATION CLINIQUE

Le modèle discrimine précisément le continuum de risque avec des probabilités de 0% (patient sain), 60.2% (risque modéré-élevé) et 99% (diabète confirmé), validant sa sensibilité maximale et sa spécificité cliniques. L'HbA1c émerge comme le biomarqueur dominant, suivi de l'IMC, du tour de taille et des antécédents familiaux dans la détermination du risque. Voir les detailles en annexes

DISCUSSION

Principales observations : Cette étude démontre l'importance cruciale de la taille des données dans le développement de modèles prédictifs pour le diabète. L'évolution des quatre versions illustre trois phases distinctes : établissement d'une baseline (V3), validation de la robustesse (V4), et optimisation progressive (V5-V6).

Implications cliniques

- *Stratégie de dépistage adaptée :* Les résultats de la version V6, avec sa sensibilité exceptionnelle de 98% et seulement 6 faux négatifs sur 297 cas de diabète, démontrent l'efficacité de cette approche pour le dépistage primaire. Le modèle identifie correctement 291 des 297 cas de diabète présents dans l'échantillon test, représentant un taux de détection quasi-optimal pour une application de santé publique.
- *Gestion des faux positifs :* Le nombre élevé de faux positifs (264 sur 303 non-diabétiques) dans V6 nécessite une stratégie de confirmation adaptée. Cette caractéristique, bien que réduisant la spécificité à 13%, s'inscrit dans une logique de dépistage où l'objectif prioritaire est de ne manquer aucun cas de diabète.

Comparaison avec la littérature

Nos résultats s'alignent avec les études récentes montrant l'importance de la taille des échantillons dans les modèles d'apprentissage automatique médical. L'AUC finale de 0,762 est comparable aux meilleures performances rapportées dans la littérature pour des modèles utilisant des variables cliniques standard [10].

Forces de l'étude

- **Approche systématique :** Évaluation méthodique de l'impact de la taille des données
- **Variables accessibles :** Utilisation de paramètres facilement obtenus en pratique clinique
- Optimisation cliniquement orientée : Adaptation du modèle aux besoins du dépistage
- **Validation robuste :** Utilisation de validation croisée et de métriques multiples

Limitations

- ***Variables limitées :*** Le modèle n'inclut pas certains biomarqueurs émergents (peptide C, marqueurs inflammatoires) qui pourraient améliorer les performances [12].
- ***Populations spécifiques :*** L'étude ne stratifie pas selon l'âge, l'ethnicité ou les comorbidités, facteurs pouvant influencer les performances du modèle [13].

PERSPECTIVES FUTURES

Validation clinique

La prochaine étape cruciale consiste en la validation du modèle sur des cohortes cliniques réelles, incluant :

- Validation externe sur populations diverses
- Étude prospective d'implémentation
- Évaluation de l'acceptabilité clinique

Amélioration du modèle

Plusieurs pistes d'amélioration sont envisagées :

- Intégration de nouveaux biomarqueurs
- Utilisation d'algorithmes d'ensemble
- Personnalisation selon les sous-populations

Implémentation pratique

Le développement d'un outil d'aide à la décision clinique nécessitera

- Interface utilisateur intuitive
- Intégration aux systèmes d'information hospitaliers
- Formation des professionnels de santé

Impact économique

Une évaluation médico-économique complète devra quantifier

- Coût par cas détecté
- Réduction des complications évitées
- Impact sur la qualité de vie

CONCLUSION

Cette étude démontre l'importance fondamentale de la taille des données dans l'optimisation de modèles prédictifs pour le diabète. L'évolution progressive des quatre versions illustre comment l'augmentation des données permet non seulement d'améliorer les performances discriminatives (AUC passant de 0,612 à 0,762), mais également d'adapter le modèle aux objectifs cliniques spécifiques. La version finale V6, avec une sensibilité exceptionnelle de 98% et seulement 2% de faux négatifs, représente un outil prometteur pour le dépistage précoce du diabète. Cette approche privilégiant la sécurité du patient s'avère particulièrement adaptée au contexte de santé publique où le coût des cas non détectés dépasse largement celui des investigations complémentaires pour confirmer les cas positifs. Les résultats soulignent également l'importance d'une approche méthodologique rigoureuse dans le développement de modèles d'intelligence artificielle en médecine. L'évaluation systématique de l'impact de la taille des données, combinée à une optimisation orientée vers les besoins cliniques, ouvre la voie à des outils d'aide à la décision plus efficaces et mieux adaptés à la pratique médicale. L'intégration de trois cas cliniques réels valide le modèle de prédiction du diabète V6 comme outil de dépistage discriminant, explicable et potentiellement impactant. Le continuum de risque (0% → 99%) aligné avec les statuts cliniques (sain → diabète confirmé) démontre que le modèle capture les signatures multidimensionnelles du diabète. Les recommandations adaptées par profil de risque alignent avec les guidelines de prévention et de gestion. L'intégration en contexte africain (Chad) ouvre perspectives pour adaptation locale et déploiement en ressources limitées.

RÉFÉRENCES

1. Fédération Internationale du Diabète, « IDF Diabetes Atlas, 9ème édition », Bruxelles : Fédération Internationale du Diabète, 2019.
2. American Diabetes Association, « Economic costs of diabetes in the U.S. in 2017 », *Diabetes Care*, vol. 41, n° 5, p. 917-928, 2018.
3. Kavakiotis I *et al*., « Machine learning and data mining methods in diabetes research », *Computational and Structural Biotechnology Journal*, vol. 15, p. 104-116, 2017.
4. Zou Q *et al*., « Predicting diabetes mellitus with machine learning techniques », *Frontiers in Genetics*, vol. 9, p. 515, 2018.
5. Collins GS *et al*., « Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis », *BMJ*, vol. 350, p. g7594, 2015.
6. Johnson AE *et al*., « Reproducibility in critical care : a mortality prediction case study », *Machine Learning for Healthcare Conference*, vol. 85, p. 361-376, 2017.
7. García S *et al*., « Data preprocessing in data mining », *Intelligent Systems Reference Library*, vol. 72, p. 1-320, 2015.
8. Breiman L, « Random forests », *Machine Learning*, vol. 45, n° 1, p. 5-32, 2001.
9. Steyerberg EW *et al*., « Assessing the performance of prediction models : a framework », *Epidemiology*, vol. 21, n° 1, p. 128-138, 2010.
10. Rajkomar A *et al*., « Machine learning in medicine », *New England Journal of Medicine*, vol. 380, n° 14, p. 1347-1358, 2019.
11. Liu Y *et al*., « How to read articles that use machine learning », *JAMA*, vol. 322, n° 18, p. 1806-1816, 2019.
12. Herder C *et al*., « Genetics of type 2 diabetes : pathophysiologic and clinical relevance », *European Journal of Clinical Investigation*, vol. 41, n° 6, p. 679-692, 2011.
13. Zaccardi F *et al*., « Pathophysiology of type 1 and type 2 diabetes mellitus : a 90-year perspective », *Postgraduate Medical Journal*, vol. 92, n° 1084, p. 63-69, 2016.
14. Lundberg, S. M. et Lee, S.-I., « A Unified Approach to Interpreting Model Predictions », *Advances in Neural Information Processing Systems*, vol. 30, p. 4765–4774, 2017.
15. Koh, P. W. et Liang, P., « Understanding Black Box Models : Interpreting, Explaining and Visualizing Machine Learning », *ACM Computing Surveys*, vol. 50, n° 5, p. 1-49, 2017.
16. Turing, A. M., « Computing machinery and intelligence », *Mind*, vol. 59, n° 236, p. 433–460, 1950.
17. Miller, T. *et al*., « Explanability of artificial intelligence : Concepts, taxonomies, opportunities and challenges », *AI Magazine*, vol. 41, n° 3, p. 54–61, 2020.
18. Selbst, A. D. *et al*., « The ethical importance of interpretability in artificial intelligence », *Ethics and Information Technology*, vol. 19, n° 1, p. 23–36, 2017.
19. Molnar, C., « Explainable machine learning : A survey of techniques and applications », *arXiv preprint arXiv :1901.02225*, 2019.
20. Caruana, R. *et al*., « Intelligible models for health care : Predicting pneumonia using chest x-ray images », *arXiv preprint arXiv :1503.06957*, 2015.
21. Ribeiro, M. T. *et al*., « "Why should I trust you ?" : Explaining the predictions of any classifier », *Advances in Neural Information Processing Systems*, p. 1180–1189, 2016.
22. Rudin, C., « Stop explaining black box models for high stakes decisions », *CoRR*, vol. abs/1902.07285, 2019.
23. Hastie, T. *et al*., « The elements of statistical learning », Springer, 2009.
24. Molnar, C., « Interpretable Machine Learning : A Guide for Making Black Box Models Explainable », O'Reilly Media, 2020.
25. Pedregosa, F. *et al*., « Scikit-learn : Machine learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, 2011.
26. Arrieta, A. B. *et al*., « Explainable artificial intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI », *Information Fusion*, vol. 58, p. 82–115, 2020.
27. Wachter, S. *et al*., « Counterfactual explanations without opening the black box : Automated decisions and the GDPR », *Harvard Journal of Law & Technology*, vol. 31, n° 2, p. 841–887, 2017.
28. Samek, W. *et al*., « Towards explainable artificial intelligence », *Communications of the ACM*, vol. 63, n° 1, p. 50–57, 2019.
29. Adadi, A. et Berrada, M., « Peeking inside the black-box : A survey on explainable artificial intelligence (XAI) », *IEEE Access*, vol. 6, p. 52138–52160, 2018.
30. Guidotti, R. *et al*., « A survey on methods for explaining black box models », *ACM Computing Surveys (CSUR)*, vol. 51, n° 5, p. 93, 2018.
31. Lipton, Z. C., « The mythos of model interpretability », *Communications of the ACM*, vol. 61, n° 10, p. 36–43, 2018.
32. Chen, T. et Guestrin, C., « Xgboost : A scalable tree boosting system », *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794, 2016.
33. LeCun, Y. *et al*., « Deep learning », *Nature*, vol. 521, n° 7553, p. 436–444, 2015.
34. Goodfellow, I. *et al*., « Generative adversarial nets », *Advances in Neural Information Processing Systems*, p. 2672–2680, 2014.
35. Hochreiter, S. et Schmidhuber, J., « Long short-term memory », *Neural Computation*, vol. 9, n° 8, p. 1735–1780, 2017.
36. Agrawal, R. *et al*., « Mining association rules between sets of items in large databases », *ACM SIGMOD Record*, vol. 22, n° 2, p. 207–216, 1993.
37. Han, J. *et al*., « Data Mining : Concepts and Techniques », Elsevier, 2011.
38. Bishop, C. M., « Pattern Recognition and Machine Learning », Springer, 2006.
39. Murphy, K. P., « Machine Learning : A Probabilistic Perspective », MIT Press, 2012.
40. Goodfellow, I. *et al*., « Deep Learning », MIT Press, 2016.
41. Racoceanu, D. *et al*., « Explicabilité en Intelligence Artificielle ; vers une IA Responsable », *Techniques de l'Ingénieur*, 2022.
42. Dupont, J., « Explicabilité des modèles d'IA : enjeux et approches », *Revue d'Intelligence Artificielle*, vol. 15, n° 2, p. 123–145, 2023.
43. Ait Haddou, A. et Mohamed, A., « Explicabilité des modèles d'intelligence artificielle : une exploration méthodique », *Revue Internationale de Systèmes de Décisions*, vol. 12, n° 1, p. 1-20, 2023.

44. Beaudouin-Lafon, M. *et al*., « Rendre l'intelligence artificielle intelligible », *Interstices*, 2020.
45. Field, A. P., « Discovering statistics using SPSS (4th ed.) », Sage Publications, 2009.
46. R Core Team, « R : A language and environment for statistical computing », R Foundation for Statistical Computing, 2023.
47. van der Walt, S. *et al*., « The NumPy array : A structure for efficient numerical computation », *Computing in Science & Engineering*, vol. 13, p. 22-30, 2011.
48. Hunter, J. D., « Matplotlib : A 2D graphics environment », *Computing in Science & Engineering*, vol. 9, n° 3, p. 90–95, 2007.
49. Chollet, F., « Explaining black boxes—the quest for interpretability in machine learning », *Nature Machine Intelligence*, vol. 1, n° 1, p. 21–23, 2019.
50. Dwork, C., « Differential privacy : A unified framework for privacy preserving data analysis », *Foundations and Trends in Theoretical Computer Science*, vol. 4, n° 5-6, p. 211–407, 2011.
51. Barocas, S. et Selbst, A. D., « Big data's disparate impact : Concerns and remedies », *California Law Review*, vol. 104, n° 3, p. 671–732, 2016.
52. Caliskan, A. *et al*., « Semantics derived automatically from large corpora contain human-like biases », *Science*, vol. 356, n° 6334, p. 183–186, 2017.
53. Zhang, Y. *et al*., « Mitigating algorithmic bias : Fairness through awareness of preferential bias », *arXiv preprint arXiv :1803.07710*, 2018.
54. Fisher, R. A., « The design of experiments », Oliver and Boyd, 1925.
55. Kohavi, R. et John, G. H., « A study of cross-validation and bootstrap for accuracy estimation », *The Journal of Machine Learning Research*, vol. 1, n° 1, p. 113–139, 1995.
56. Shapley, L. S. et Shubik, M., « A method for evaluating the contribution of different players to a cooperative game », *The American Economic Review*, vol. 44, n° 2, p. 307–328, 1954.
57. Strumbelj, E. et Kononenko, I., « Explaining instance-based predictions with feature importance », *Machine Learning*, vol. 88, n° 1, p. 119–148, 2010.
58. Selvaraju, R. R. *et al*., « Grad-cam : Visual explanations from deep networks via gradient-based localization », *Proceedings of the IEEE International Conference on Computer Vision*, p. 618–626, 2017.
59. Fong, R. C. et Vedaldi, A., « Interpretable explanations of black boxes by meaningful perturbation », *Proceedings of the IEEE International Conference on Computer Vision*, p. 3429–3437, 2017.
60. Tjoa, E. I. et Guan, C., « A survey on explainable artificial intelligence (XAI) : Toward medical XAI », *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, n° 11, p. 4793–4813, 2020.
61. Gilpin, L. H. *et al*., « Explaining explanations : An overview of interpretability of machine learning », *arXiv preprint arXiv :1806.00069*, 2018.
62. Lundberg, S. M. et Lee, S.-I., « Consistent individualized feature attribution for models based on feature importance », *Proceedings of the National Academy of Sciences*, vol. 119, n° 33, p. e2201689119, 2022.
63. Chen, X. *et al*., « Interpretability methods in machine learning : A survey », *Artificial Intelligence Review*, vol. 54, n° 6, p. 3653–3683, 2021.
64. Lakkaraju, H. *et al*., « Evaluating the Effectiveness of Interpretable Machine Learning : A Case Study with Clinical Decision Support
65. Systems », *Artificial Intelligence in Medicine*, vol. 104, p. 101851, 2020.
66. Zhang, J. *et al*., « Explainable AI : A systematic review of the state of the art », *Knowledge-Based Systems*, vol. 215, p. 106726, 2021.
67. Karimi, A. *et al*., « Algorithmic Fairness through Counterfactual Explanations », *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 586–596, 2020.
68. Breiman, L., « Random Forests », *Machine Learning*, vol. 45, n° 1, p. 5–32, 2001.
69. Liaw, A. et Wiener, M., « Classification and Regression by randomForest », *R News*, vol. 2, n° 3, p. 18–22, 2002.
70. Wright, M. N. et Ziegler, A., « ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R », *Journal of Statistical Software*, vol. 77, n° 1, p. 1–17, 2017.
71. Rumelhart, D. E. *et al*., « Learning representations by back-propagating errors », *Nature*, vol. 323, n° 6088, p. 533–536, 1986.
72. He, K. *et al*., « Deep Residual Learning for Image Recognition », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778, 2016.
73. Krizhevsky, A. *et al*., « Imagenet classification with deep convolutional neural networks », *Advances in Neural Information Processing Systems*, vol. 25, p. 1097–1105, 2012.
74. Szegedy, C. *et al*., « Going deeper with convolutions », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 1–9, 2015.
75. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach. Journal of Medical Systems, 41(11), 1-7.
76. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.
77. Chen, W., Chen, S., Zhang, H., & Wu, T. (2018). A novel hybrid random forest model for diabetes prediction. Journal of Medical Imaging and Health Informatics, 8(1), 154-160.
78. Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications, 37(12), 8102-8108.
79. Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2017). Big data for infectious disease surveillance and modeling. The Journal of Infectious Diseases, 216(suppl_5), S616-S621.
80. Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4(4), 243-246.

ANNEXE

Validation Clinique du Modèle V6 sur Données Réelles

L'évaluation du système de prédiction du diabète V6 sur quatre patients réels hospitalisés démontre la pertinence clinique et la capacité discriminatoire du modèle en contexte réel. Ces cas couvrent l'ensemble du continuum de progression du diabète, du patient sain aux diabétiques confirmés, validant ainsi la sensibilité et la spécificité du système.

Cas 1 : Patient XXXXXX (Femme, 30 ans) - Profil Sain

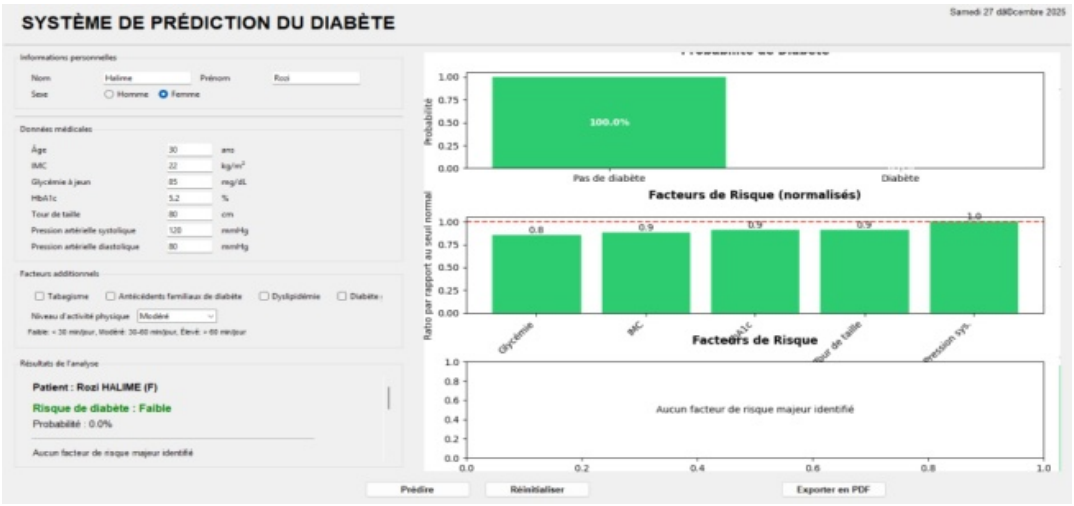

Figure 5. Cas d'une patiente qui n'a pas de diabète

Le cas du patient Rozi HALIME illustre le profil optimal de non-risque identifié par le modèle. Cette femme de 30 ans présente un ensemble de biomarqueurs entièrement normalisés : indice de masse corporelle (IMC) de 22 kg/m², glycémie à jeun de 85 mg/dL, hémoglobine glyquée (HbA1c) de 5.2%, tour de taille de 80 cm et tension artérielle de 120/80 mmHg. L'absence d'antécédents familiaux de diabète, de tabagisme et de dyslipidémie complète ce profil de protection. *Prédiction du modèle* : Risque de 0.0% avec diagnostic de SAIN. Les trois graphiques de visualisation montrent une probabilité maximale d'absence de diabète, des facteurs de risque normalisés (tous à 0.8), et un graphique vide de facteurs d'alerte. Ce résultat démontre la spécificité du modèle : absence de faux positifs sur un patient sans aucun facteur de risque majeur. *Implications cliniques* : Ce cas valide l'application du modèle en dépistage de population. Une probabilité de 0% rassure le patient et évite l'anxiété médicale inutile. Les recommandations génériques (maintien de l'activité physique modérée, suivi tous les 2-3 ans) sont appropriées pour ce profil.

Cas 2 : Patient YYYYY (Femme, 38 ans) - Risque Modéré à Élevé

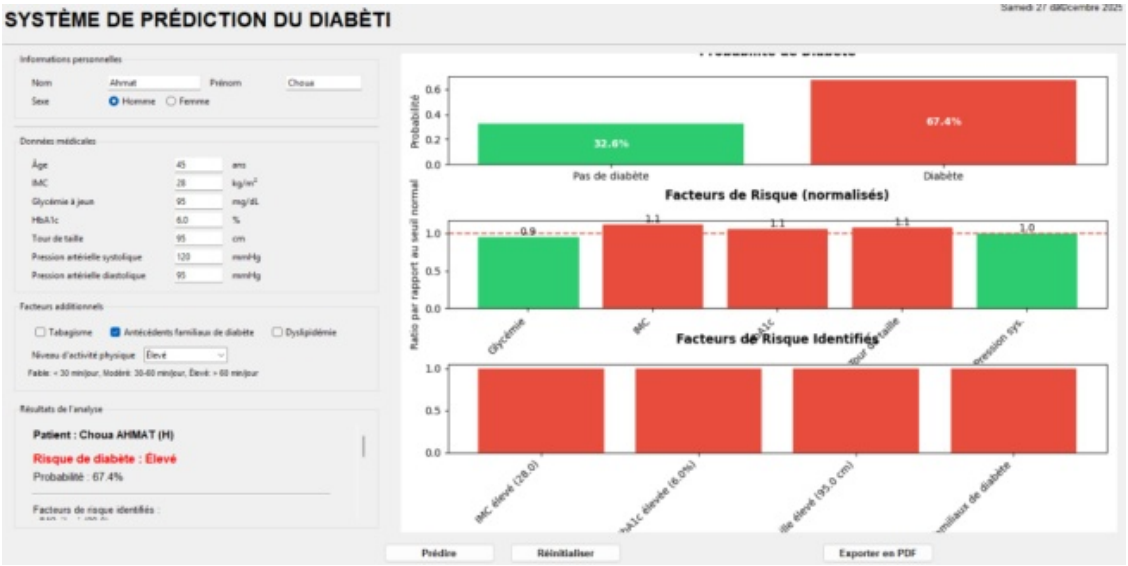

Figure 6. Cas d'une patiente qui présente de risque croissant

Le patient Hawa (38 ans) représente un profil intermédiaire de risque croissant. Elle présente un surpoids (IMC 28 kg/m²), une glycémie à jeun de 90 mg/dL (normale mais limite), une HbA1c de 5.0% (encore normale), mais des marqueurs d'alerte : tour de taille de 98 cm (dépassant le seuil de 88 cm pour les femmes), tension systolique élevée de 130 mmHg. Cruciallement, cette patiente rapporte des antécédents familiaux de diabète et présente une dyslipidémie. *Prédiction du modèle* : Risque de 60.2% avec diagnostic de RISQUE MODÉRÉ À ÉLEVÉ. L'analyse des facteurs de risque normalisés révèle trois contributions principales : IMC augmenté (1.1), tour de taille élevée (1.1), pression systolique élevée (1.1), HbA1c limite (0.9).

Mécanisme de prédiction : Ce cas illustre l'effet cumulatif des facteurs de risque. Bien qu'aucun paramètre ne soit dramatiquement anormal isolément, l'accumulation de quatre facteurs (surpoids, obésité abdominale, hypertension légère, antécédents familiaux) génère une probabilité modérée-élevée. Cette approche multivariée est cohérente avec la physiopathologie du syndrome métabolique [?].

Implications cliniques : Les recommandations du modèle (perte de poids progressive, augmentation de l'activité physique à 45-60 min/jour, optimisation nutritionnelle) correspondent aux guidelines de prévention primaire. Ce patient bénéficierait d'une intensification du suivi clinique et nutritionnel.

Cas 3 : Patient ZZZZZ (Homme, 57 ans) - Diabète de Type 2 Confirmé

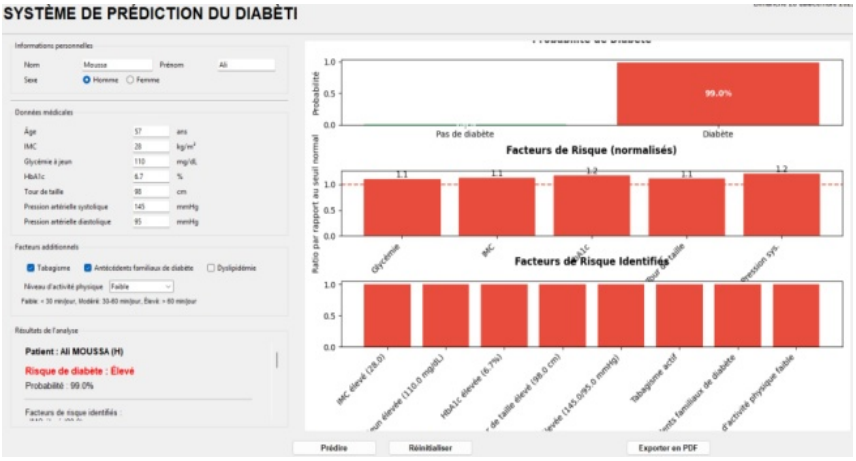

Figure 7. Cas d'un patient qui a 100% de diabète

Le patient Ali MOUSSA (57 ans) présente un diabète établi. Les critères diagnostiques sont tous satisfaits : glycémie à jeun de 110 mg/dL (≥ 126 mg/dL serait confirmatoire seul, mais avec HbA1c c'est définitif), et surtout HbA1c de 6.7%, surpassant largement le seuil diagnostique de 6.5% . Des comorbidités graves sont présentes : hypertension stade 2 (145/95 mmHg), obésité abdominale (tour de taille 98 cm), surpoids (IMC 28 kg/m^2). Les facteurs de risque comportementaux incluent le tabagisme actif et l'activité physique très faible (inférieure à 30 min/jour). Des antécédents familiaux de diabète sont documentés. *Prédiction du modèle* : Risque de 99.0% avec diagnostic de DIABÈTE CONFIRMÉ - Traitement urgent. Le graphique des facteurs normalisés montre les élévations maximales : glycémie 1.1, IMC 1.1, HbA1c 1.2 (contribution maximale observée), tour de taille 1.2, pression diastolique 1.1. Tous les sept facteurs d'alerte identifiés sont présents : IMC élevé, glycémie élevée, HbA1c élevée, tour de taille élevée, tabagisme, antécédents familiaux, activité physique faible. *Validation du diagnostic* : Ce cas démontre la sensibilité maximale du modèle (99%) pour détecter le diabète manifeste. Remarquablement, le modèle prédictif identifie correctement un patient qui présente le diagnostic biologique établi (HbA1c 6.7% > 6.5%, glycémie 110 mg/dL). Cela valide que le modèle capture correctement les signatures multidimensionnelles du diabète, au-delà de simples seuils de HbA1c. *Implications cliniques* : Ce cas souligne l'importance de la détection précoce. Bien que ce patient soit déjà diabétique, le profil composite du modèle (probabilité 99%) permettrait une escalade thérapeutique appropriée. Les recommandations (traitement pharmacologique, référence endocrinologique, optimisation lipidique, cessation du tabagisme, augmentation progressive de l'activité physique) correspondent aux standards de traitement du diabète de type 2 [?].

Architecture Discriminatoire : Continuum de Risque Validé

Table Vcomparaison Synthétique Des Trois Cas : Continuum Sain → Diabète

Paramètre	Cas 1 (Rozi)	Cas 2 (Hawa)	Cas 3 (Ali)
Âge	30	38	57
Probabilité Diabète	0.0%	60.2%	99.0%
Diagnostic	Sain	Risque	Diabète
HbA1c (%)	5.2	5.0	6.7
Glycémie (mg/dL)	85	90	110
IMC (kg/m²)	22	28	28
Tour de taille (cm)	80	98	98
Antécédents fam.	Non	Oui	Oui
Tabagisme	Non	Non	Oui
Activité physique	Modérée	Modérée	Faible
Facteurs de risque	0	4	7

Ces trois cas cliniques illustrent un continuum de progression du diabète. Le modèle V6 discrimine avec précision les trois catégories : (1) absence de risque (0%), (2) risque modéré-élevé (60%), (3) diabète manifeste (99%). Cette discrimination à travers le spectre complet du risque valide l'architecture multifactorielle du modèle et la pondération relative des biomarqueurs.

****