JOURNAL AFRICAIN DES SCIENCES
VOL. 3, **3** (2026) 90-94

https://jafrisci.org

ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930

OPEN ACCESS

ARTICLE DE RECHERCHE

Article Info.:
Reçu : le 20/02/2026
Accepté: le 02/04/2026
Publié : le 04/04/2026

MODELE D'APPRENTISSAGE PAR ENSEMBLE BASE SUR LE DOUBLE STACKING POUR LA PREDICTION DU SYNDROME METABOLIQUE

Christophe Fumumuiki Masenzi [1,*], Glad Lufimpu Mampuya [1], Kevin Mongoy Bonyolo [1], Eugene Mbuyi Mukendi [2], Rostin Mabela Makengo [2], Cédric Kabeya Tshiseba [1], Joseph Kasiama Ngi Onkor [1], Joël Mangoma [1]

https://doi.org/10.70237/jafrisci.2026.v3.i3.10

Résumé

Le syndrome métabolique est un problème majeur de santé publique en raison de son association avec les maladies cardiovasculaires, les accidents vasculaires cérébraux et le diabète de type 2. Cette étude propose un modèle de prédiction fondé sur une architecture Double Stacking appliquée aux données NHANES 1998-2018. La méthodologie comprend le prétraitement des variables démographiques, cliniques et biologiques, l'entraînement de plusieurs classifieurs de base, puis leur combinaison hiérarchique par deux niveaux successifs de méta-apprentissage. Les résultats montrent que le modèle proposé améliore la robustesse de la prédiction et atteint une accuracy de 91,25 %, un F1-score de 86,18 % et une ROC-AUC de 0,959. Ces performances indiquent que le Double Stacking constitue une approche prometteuse pour la détection précoce du syndrome métabolique.

Mots-clés: Apprentissage automatique, syndrome métabolique, méthodes d'ensemble, stacking, double stacking, prédiction

Abstract

Metabolic syndrome is a major public health issue because of its strong association with cardiovascular disease, stroke, and type 2 diabetes. This study proposes a predictive model based on a Double Stacking architecture using NHANES 1998-2018 data. The methodology includes preprocessing demographic, clinical, and biological variables, training several base classifiers, and combining them through two successive meta-learning levels. The results show that the proposed model improves predictive robustness and reaches an accuracy of 91.25%, an F1-score of 86.18%, and a ROC-AUC of 0.959. These findings suggest that Double Stacking is a promising approach for the early prediction of metabolic syndrome.

Keywords: Machine learning, metabolic syndrome, ensemble methods, stacking, double stacking, prediction

1. Introduction

Le syndrome métabolique correspond à un ensemble d'anomalies métaboliques incluant l'obésité abdominale, l'hypertension artérielle, l'hyperglycémie et les dyslipidémies. Sa prévalence croissante en fait un enjeu majeur pour les systèmes de santé, car il augmente de manière significative le risque de maladies cardiovasculaires et de diabète de type 2 [1]-[3].

L'identification précoce des sujets à risque constitue donc un objectif important en médecine préventive. Avec la disponibilité croissante de données biomédicales, les méthodes d'apprentissage automatique offrent des outils capables d'analyser des relations complexes entre variables cliniques et biologiques afin d'améliorer la prédiction des maladies [4], [5].

Cependant, les classifieurs traditionnels utilisés isolément présentent souvent des limites de généralisation face à des données hétérogènes et non linéaires. Les méthodes d'ensemble, notamment le bagging, le boosting et le stacking, ont été développées pour réduire la variance, mieux exploiter la diversité des modèles et améliorer la précision globale [6], [7].

Dans ce contexte, cet article propose un modèle d'apprentissage par ensemble fondé sur une architecture Double Stacking pour la prédiction du syndrome métabolique à partir du jeu de données NHANES 1998-2018. L'objectif est d'évaluer si une combinaison hiérarchique de classifieurs permet d'obtenir des performances supérieures aux approches classiques.

Correspondance : *christophefumumuiki@gmail.com* ***(C. F. Masenzi)***

Copyright : © The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)

[1] Département de Mathématiques, Statistique et Informatique, Université Pédagogique Nationale, R.D. Congo ;
[2] Département de Mathématiques, Statistique et Informatique, Université de Kinshasa, R.D. Congo.

2. Revue de littérature

Plusieurs travaux récents ont montré l'intérêt du machine learning pour la prédiction du syndrome métabolique. Des modèles comme la régression logistique, le KNN, le SVM et les arbres de décision ont été appliqués avec succès à des variables cliniques et biologiques afin d'identifier les individus à risque [8]-[10].

Les méthodes d'ensemble occupent une place particulière dans cette littérature. Le Random Forest illustre l'intérêt du bagging, tandis que Gradient Boosting et XGBoost démontrent l'efficacité du boosting dans la réduction progressive de l'erreur. Le stacking, quant à lui, combine plusieurs prédictions de base à l'aide d'un méta-modèle capable de tirer parti de leur complémentarité [6], [11].

Dans le domaine médical, le stacking a déjà été mobilisé pour des tâches de diagnostic et de stratification du risque. Néanmoins, les architectures étudiées demeurent souvent limitées à un seul niveau de méta-apprentissage. Cette limite peut empêcher de capturer certaines interactions complexes entre les sorties des classifieurs de base.

La contribution de ce travail consiste donc à proposer une architecture Double Stacking, c'est-à-dire un empilement à deux niveaux de méta-apprentissage, afin d'augmenter la robustesse et la capacité de discrimination du modèle dans le contexte particulier de la prédiction du syndrome métabolique.

3. Matériels et méthodologie

3.1 Données utilisées

Les données proviennent de l'enquête NHANES (National Health and Nutrition Examination Survey), réalisée par le National Center for Health Statistics aux Etats-Unis. Cette base est largement utilisée dans les études épidémiologiques en raison de la qualité de ses informations démographiques, anthropométriques et biologiques.

Après sélection et nettoyage, le jeu de données exploité dans cette étude comprend 2401 observations. La variable cible est binaire : 0 pour l'absence de syndrome métabolique et 1 pour sa présence. Les prédicteurs retenus décrivent l'âge, le sexe, l'adiposité, la pression artérielle, le métabolisme glucidique et lipidique.

Table 1. Variables explicatives principales utilisées dans le modèle

Variable	Description	Type
Age	âge du participant	Numérique
Gender	sexe du participant	Catégorielle
BMI	indice de masse corporelle	Numérique
Waist Circumference	tour de taille	Numérique
Systolic Blood Pressure	pression artérielle systolique	Numérique
Diastolic Blood Pressure	pression artérielle diastolique	Numérique
Fasting Glucose	glycémie à jeun	Numérique
Triglycerides	taux de triglycérides	Numérique
HDL Cholesterol	cholestérol HDL	Numérique

3.2 Prétraitement et expérimentation

Le prétraitement comprend l'encodage des variables catégorielles, la normalisation des variables numériques et la séparation des données en ensembles d'apprentissage et de test. Cette étape est essentielle pour garantir la comparabilité des variables et la stabilité de l'entraînement.

Les expérimentations ont été réalisées sous Python dans un environnement Jupyter Notebook à l'aide des bibliothèques Pandas, NumPy, Scikit-learn, Matplotlib et Seaborn. Plusieurs algorithmes de classification ont d'abord été évalués séparément, notamment la régression logistique, le KNN, le SVM, l'arbre de décision, Random Forest et Gradient Boosting.

Les principaux algorithmes utilisés dans cette étude sont résumés comme suit:

La régression logistique est un modèle probabiliste linéaire qui estime la probabilité d'appartenance à une classe à l'aide d'une fonction sigmoïde. Elle sert de modèle de référence simple et interprétable.

Le KNN classe une observation en fonction des k voisins les plus proches selon une mesure de distance. Cette méthode est simple à mettre en œuvre, mais son coût de prédiction peut devenir élevé lorsque le nombre d'observations augmente.

Le SVM cherche un hyperplan séparant au mieux les classes. Il est particulièrement utile lorsque la frontière de décision est complexe, surtout lorsqu'il est associé à un noyau non linéaire.

L'arbre de décision segmente récursivement l'espace des variables à partir de critères comme l'indice de Gini ou l'entropie. Il est facile à interpréter, mais peut être instable lorsqu'il est utilisé seul.

Random Forest combine plusieurs arbres de décision construits sur des échantillons bootstrap. Cette stratégie de bagging réduit la variance et améliore généralement la stabilité des prédictions.

Gradient Boosting construit les modèles de manière séquentielle en corrigeant progressivement les erreurs des modèles précédents. Il s'agit d'une approche performante pour apprendre des relations non linéaires.

À titre indicatif, le coût du KNN en prédiction dépend du nombre d'observations, celui du SVM peut devenir important lors de l'entraînement, alors que Random Forest et Gradient Boosting dépendent surtout du nombre d'arbres et de la profondeur des modèles. Ces éléments justifient le choix d'une approche de combinaison, capable d'exploiter les points forts de chaque algorithme tout en limitant leurs faiblesses respectives.

4. Modèle proposé : Double Stacking

Le principe du stacking consiste à utiliser plusieurs modèles de base dont les prédictions servent d'entrées à un méta-modèle. Dans le modèle proposé, cette logique est étendue à deux niveaux successifs de méta-apprentissage.

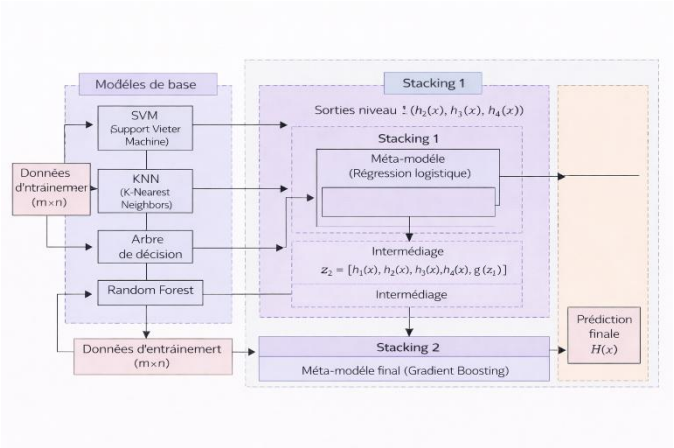
Au premier niveau, plusieurs classifieurs hétérogènes apprennent sur les données originales. Leurs sorties sont ensuite combinées par un premier méta-modèle qui produit une prédiction intermédiaire. Cette prédiction alimente enfin un second méta-modèle chargé de fournir la décision finale.

Soit x un vecteur de caractéristiques et h1(x), h2(x), .. hM(x) les prédictions des M modèles de base. Le premier niveau construit z1 = (h1(x), h2(x), ..., hM(x)), puis un premier méta-modèle g1 apprend y1 = g1(z1).

Le second niveau utilise z2 = (y1, z1) ou un sous-ensemble de sorties intermédiaires pour entraîner g2 et produire la prédiction finale H(x) = g2(z2).

Cette architecture vise à mieux exploiter la diversité des classifieurs, à réduire à la fois le biais et la variance, et à capturer des interactions plus riches entre les sorties des modèles dans un problème de prédiction médicale complexe.
Pour mieux expliciter la combinaison proposée, on peut distinguer les deux étapes suivantes. Au premier niveau, les classifieurs de base produisent chacun une sortie à partir des variables originales. On forme ainsi un vecteur de prédictions intermédiaires z^(1) = (h1(x), h2(x), ..., hM(x)). Un premier méta-modèle g1 apprend alors une nouvelle représentation y^(1) = g1(z^(1)).

Au second niveau, les sorties du premier niveau sont réutilisées pour construire une représentation enrichie z^(2) = (y^(1), z^(1)). Un second méta-modèle g2 fournit ensuite la décision finale selon H(x) = g2(z^(2)). Cette écriture montre clairement que l'approche proposée n'est pas une simple juxtaposition de classifieurs, mais une combinaison hiérarchique en deux étapes.



La figure ci-haut, présente l'architecture du modèle Double Stacking proposé. Les modèles de base (SVM, KNN, arbre de décision et Random Forest) produisent des sorties de premier niveau combinées par une régression logistique (Stacking 1). Ces sorties sont ensuite enrichies lors de l'intermédiage et exploitées par un modèle Gradient Boosting (Stacking 2) pour générer la prédiction finale H(x).
Architecture du modèle Double Stacking proposé

En pratique, ce mécanisme permet de corriger progressivement certaines erreurs des modèles de base. Un premier niveau apprend à agréger leurs prédictions, puis un second niveau affine encore cette combinaison. C'est ce processus qui explique la meilleure robustesse observée dans les résultats.

5. Résultats et discussion

5.1 Métriques d'évaluation

L'évaluation repose sur cinq indicateurs : l'accuracy, la precision, le recall, le F1-score et la ROC-AUC. Dans un contexte médical, ces métriques sont complémentaires : l'accuracy renseigne sur la performance globale, alors que la precision et le recall sont particulièrement utiles pour apprécier respectivement la fiabilité des cas positifs détectés et la capacité du modèle à identifier les individus réellement à risque.

Le F1-score synthétise la precision et le recall dans une seule mesure, tandis que la ROC-AUC évalue la capacité du classifieur à discriminer les deux classes sur l'ensemble des seuils possibles.

Table 2. Comparaison des performances des principaux modèles

Modèle	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.8295	0.7785	0.7030	0.7389	0.8985
KNN	0.8340	0.7832	0.7091	0.7443	0.9031
SVM	0.8424	0.7958	0.7333	0.7633	0.9275
Decision Tree	0.8316	0.8443	0.6242	0.7178	0.9223
Random Forest	0.8711	0.8239	0.7939	0.8086	0.9445
Gradient Boosting	0.8773	0.8312	0.8061	0.8185	0.9493
Double Stacking	0.9125	0.8724	0.8517	0.8618	0,959

Les résultats montrent que le modèle Double Stacking obtient la meilleure performance globale. Il dépasse les classifieurs individuels ainsi que les méthodes d'ensemble plus classiques sur l'ensemble des indicateurs retenus. L'amélioration est particulièrement visible sur la ROC-AUC, ce qui traduit une meilleure capacité de discrimination entre les individus atteints et non atteints.

Cette supériorité peut s'expliquer par la structure hiérarchique du modèle. En combinant plusieurs classifieurs de base puis en affinant leur combinaison par un second méta-modèle, le Double Stacking exploite mieux l'hétérogénéité des informations portées par les prédictions intermédiaires.

Dans une perspective d'application, ces résultats suggèrent qu'un système d'aide au diagnostic fondé sur ce type d'architecture pourrait contribuer à la détection précoce du syndrome métabolique. Toutefois, la généralisabilité du modèle reste dépendante de la qualité des données d'entrée et devra être confirmée par des validations externes sur d'autres populations. L'amélioration apportée par le Double Stacking ne doit donc pas être comprise comme une simple affirmation. Elle est appuyée par les métriques du tableau 2, où le modèle proposé obtient les meilleures valeurs en accuracy, en F1-score et en ROC-AUC. Autrement dit, la combinaison hiérarchique améliore à la fois la qualité globale de classification et la capacité de discrimination entre sujets atteints et non atteints.

Cette supériorité s'explique par trois éléments complémentaires : la diversité des modèles de base, le rôle du premier méta-modèle dans l'agrégation des sorties, et le rôle du second méta-modèle dans l'affinement de la décision finale. Ainsi, l'architecture proposée apporte une amélioration mesurable par rapport aux modèles utilisés séparément ou aux combinaisons plus classiques à un seul niveau.

5.2 Analyse graphique et validation visuelle du modèle

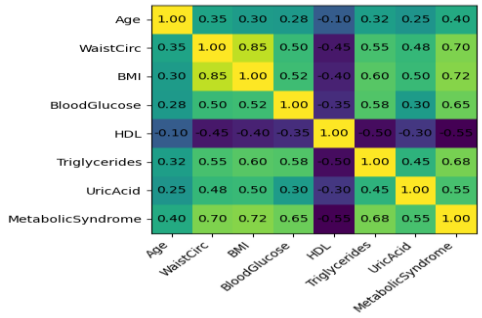


Figure 1. *matrice de corrélation des principales variables cliniques*

La figure 1 met en évidence des associations positives entre plusieurs variables métaboliques, en particulier le tour de taille, l'IMC, la glycémie et les triglycérides. À l'inverse, le HDL présente une relation plutôt négative avec le syndrome métabolique.

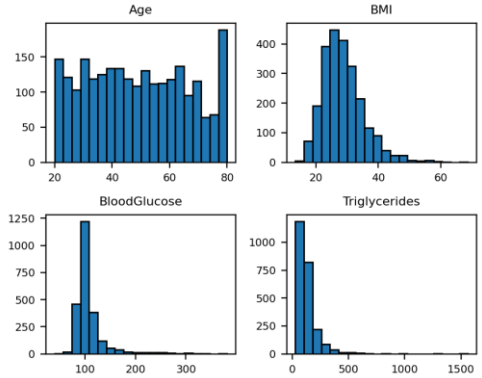


Figure 2. *histogrammes de quelques variables clés*

La figure 2 montre des distributions souvent asymétriques pour plusieurs variables continues. L'IMC et le tour de taille se concentrent autour de valeurs intermédiaires, tandis que la glycémie et les triglycérides présentent des queues plus étendues. Cette hétérogénéité statistique est cohérente avec la diversité clinique des profils observés dans les données NHANES.

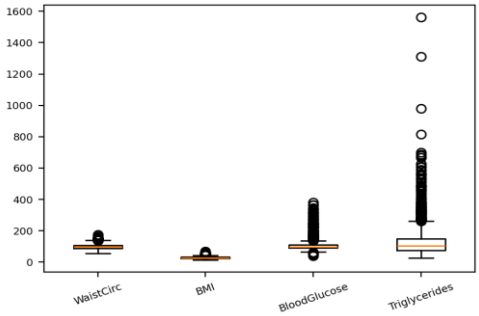


Figure 3. *boxplots des variables et détection des valeurs aberrantes*

La figure 3 confirme la présence de valeurs extrêmes, notamment pour la glycémie, les triglycérides et le tour de taille. Dans un cadre médical, ces observations ne sont pas nécessairement des erreurs de mesure ; elles peuvent aussi traduire des situations à haut risque métabolique. Le recours à une architecture de combinaison de classifieurs permet alors de mieux absorber cette variabilité sans dégrader fortement la capacité prédictive.

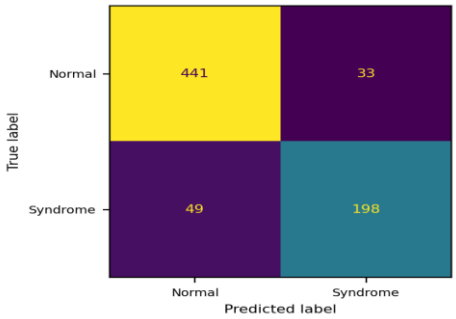


Figure 4. *validation par matrice de confusion*

La figure 4 montre que le modèle classe correctement 441 cas négatifs et 198 cas positifs, avec 33 faux positifs et 49 faux négatifs sur l'échantillon de test. Cette répartition indique une bonne stabilité globale du modèle et une capacité satisfaisante à identifier les individus réellement atteints du syndrome métabolique.

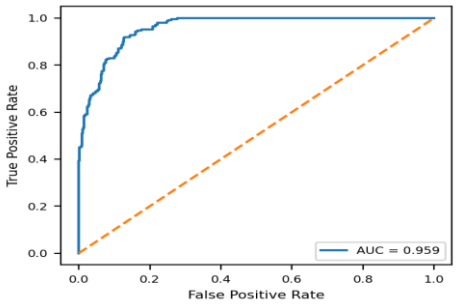


Figure 5. *courbe ROC du Double Stacking*

La figure 5 montre une courbe ROC nettement au-dessus de la diagonale aléatoire, avec une aire sous la courbe d'environ 0,959. Ce résultat traduit une forte capacité de discrimination entre les sujets atteints et non atteints. Il confirme que l'architecture Double Stacking fournit une performance robuste pour la prédiction précoce du syndrome métabolique.

6. Conclusion

Cet article a présenté un modèle d'apprentissage par ensemble fondé sur une architecture Double Stacking pour la prédiction du syndrome métabolique à partir des données NHANES 1998-2018. Les résultats obtenus montrent que cette approche améliore les performances par rapport aux modèles classiques et aux autres méthodes d'ensemble évaluées.

Le principal apport du travail réside dans l'introduction d'une structure hiérarchique de méta-apprentissage capable d'exploiter plus efficacement la diversité des classifieurs. Les limites de l'étude tiennent notamment à l'absence de validation externe et à la dépendance aux variables sélectionnées. Des travaux futurs pourront intégrer une optimisation plus poussée des hyperparamètres, une validation multicentrique et l'étude de l'interprétabilité clinique du modèle.

Références

[1] K. G. M. M. Alberti, P. Zimmet, & J. Shaw, "The metabolic syndrome: a new worldwide definition," The Lancet Diabetes & Endocrinology, 2016.
[2] R. H. Eckel, K. Alberti, S. Grundy, & P. Zimmet, "The metabolic syndrome," The Lancet, 2018.
[3] P. L. Huang, "A comprehensive definition for metabolic syndrome," Disease Models & Mechanisms, 2019.
[4] A. Rajkomar, J. Dean, & I. Kohane, "Machine learning in medicine," The New England Journal of Medicine, 2019.
[5] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, 2019.
[6] Z. H. Zhou, "Ensemble learning," Machine Learning Journal, 2021.
[7] O. Boitor et al., "Automated machine learning to develop predictive models for metabolic syndrome," Applied Sciences, 2023.
[8] M. G. Sghaireen et al., "Machine learning approach for metabolic syndrome diagnosis," Computers in Biology and Medicine, 2022.
[9] H. Yang et al., "Machine learning-aided risk prediction for metabolic syndrome," Scientific Reports, 2022.
[10] M. Trigka et al., "Predicting the occurrence of metabolic syndrome using machine learning models," Journal of Clinical Medicine, 2023.
[11] D. Shin et al., "Prediction of metabolic syndrome using machine learning and polygenic risk scores," Scientific Reports, 2024.