Analyse de l'Influence des Caractéristiques sur la Performance du Modèle Prédictif de la Rétinopathie Diabétique

Tinhinane Bessaa, Ferroudja Bellal, Samira Ait Kaci Azzou, Djamila Boukredera, Rafik Tafoukt

Analyse de l'Influence des Caractéristiques sur la Performance du Modèle Prédictif de la Rétinopathie Diabétique

Bellal Ferroudja and Bessaa Tinhinane

Department of Informatic, Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria

ferroudja.bellal@se.univ-bejaia.dz
tinhinane.bessaa@se.univ-bejaia.dz

Ait Kaci Azzou Samira[(1)], Boukredera Djamila[(2)] and Dr Tafoukt Rafik

[(1)] *LIMED Laboratory of Informatic and Medical, Faculty of Exact Sciences, University of Bejaia, Algeria*
[(2)] *Laboratory of Applied Mathematics, Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria*

Samira.aitkaciazzou@univ-bejaia.dz
Djamila.boukredera@univ-bejaia.dz
Rafiktafoukt6@gmail.com

***Résumé*– La rétinopathie diabétique (RD) est une complication fréquente chez les patients atteints de diabète de type 2, entraînant une perte de vision significative voire la cécité. La prédiction précoce de la RD joue un rôle crucial dans sa prévention et dans la réduction de sa progression.**
L'objectif de cette étude est de mettre en exergue l'influence des caractéristiques du dataset sur la qualité du modèle prédictif.
Nous avons utilisé deux datasets publics différents et expérimenté plusieurs techniques d'apprentissage automatique pour développer notre modèle prédictif.
Le premier dataset comprend les données de 844 patients avec 11 caractéristiques, tandis que le deuxième contient les données de 133 patients avec 21 caractéristiques. En appliquant diverses méthodes d'apprentissage telles que Adaboost, Catboost, Decision Tree, etc., à ces deux datasets, nous avons constaté que plus le nombre de caractéristiques représentatives est élevé, meilleur est le modèle prédictif.
De plus, nous avons comparé les différentes techniques utilisées afin de déterminer le modèle le plus performant en termes d'AUC-ROC (Area Under the Receiver Operating Characteristic curve). Les résultats obtenus ont montré que le modèle CatBoost a donné les meilleures performances, avec une AUC de 0.615 pour le premier dataset et une AUC de 0.855 pour le deuxième dataset. Dans le cadre de cette étude, CatBoost représente ainsi le modèle prédictif optimal de la RD.

Mots Clés : Caracteristiques, Dataset, Apprentissage automatique, Rétinopathie diabétique, prédiction

I. Introduction

La rétinopathie diabétique (RD) est une complication grave qui se développe généralement chez les personnes âgées atteintes de diabète de type 2. Elle est caractérisée par une détérioration des vaisseaux sanguins de la rétine, ce qui en fait l'une des principales causes de cécité. La détection précoce de cette maladie est essentielle pour éviter les complications visuelles et réduire les coûts élevés de traitement, tout en allégeant la charge de travail des professionnels de la santé [1, 2, 3]. Il est important de noter que la RD est souvent asymptomatique aux premiers stades, soulignant l'importance cruciale d'une prédiction précoce.

Dans ce contexte, plusieurs méthodes d'apprentissage automatique ont été développées pour la prédiction de la rétinopathie diabétique (RD). Ces méthodes permettent d'établir un processus de prédiction automatisé, offrant ainsi une solution à faible coût pour fournir un traitement en temps opportun aux patients. La Figure Fig.1 illustre l'évolution temporelle des méthodes d'apprentissage dans ce domaine [8].

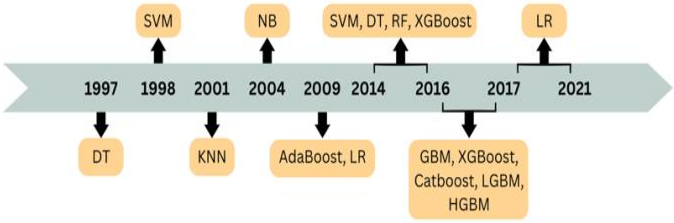


Fig.1 L'utilisation des modèles d'apprentissage automatique au fil du temps pour la prédiction de la RD

DT: Decision Tree,
SVM: Support Vector Machine,
KNN: K plus proche voisins,
NB: Naïve Bayes,
Adaboost :Adaptive Boosting Classifier
LR: Régression Logistique,
RF: Random Forest,
XGBoost: Extreme Gradient Boosting,
GBM: Gradient Boosting Machine,
CatBoost: CatBoost Classifier,
LGBM: Light Gradient Boosting Machine,
Histogram GB: Histogram Gradient Boosting

De nombreuses études ont été menées avec des résultats prometteurs dans ce domaine. En effet, Hsin-Yi Tsao et al. [4] ont appliqué différents algorithmes d'apprentissage automatique tels que SVM, les arbres de décision, la régression logistique et les réseaux de neurones artificiels pour prédire la rétinopathie diabétique (RD) et identifier de nouvelles caractéristiques associées à cette maladie. Leur objectif était d'obtenir la meilleure prédiction possible. Les

résultats de leur étude ont révélé que le déclenchement de la RD était lié non seulement à un taux élevé de HbA1C, mais également à d'autres caractéristiques telles que l'utilisation d'insuline et la durée du diabète. Le meilleur modèle obtenu est SVM avec une précision de 83.9%.

Wanyue Li et al. [7] ont étudié les facteurs de risque de la RD en utilisant les modèles d'apprentissage automatique à savoir XGBoost, LR, RF et SVM et ce, à l'aide d'une très large base de données de patients hospitalisés atteints de diabète de type 2 de l'hôpital général chinois. XGBoost a donné le meilleur résultat de prédiction à savoir une précision de 90%.

D'autre part, Yazan Jian et al. [5] ont également mené une étude sur la prédiction et la classification de huit complications du diabète, dont la rétinopathie diabétique, en utilisant une base de données recueillie par le Rashid Center for Diabetes and Research (RCDR) situé à Ajman, aux Émirats arabes unis [6]. Plusieurs algorithmes de classification supervisée, tels que la régression logistique, SVM, les arbres de décision (DT), la forêt aléatoire (RF), AdaBoost et XGBoost, ont été appliqués dans cette étude. Les résultats ont montré que le modèle XGBoost a donné les meilleurs résultats de prédiction pour la RD, avec une précision de 87,2% et un score F1 de 86,7%.

Il convient de noter que la comparaison directe entre différentes méthodes de prédiction de la RD peut être difficile en raison des différences dans les datasets utilisés. Certains datasets peuvent être petits mais comportent de nombreuses caractéristiques, tandis que d'autres peuvent être plus grands mais avec moins de caractéristiques, comme illustré dans le Tableau 1. Ces variations dans les datasets peuvent influencer les performances des modèles de prédiction.

TABLE 1 Modèles et résultats

Article	Dataset (P/C)	Meilleur modèle	AUC	ACC	Limitation
[4]	536/10	SVM	0.839	0.795	Peu de ressource et absence de validation
[5]	844/87	XGBoost	/	0.872	Dataset déséquilibré
[7]	32452/17	XGBoost	0.90	0.90	Absence de validation externe

P : nombre de patients, *C :* nombre de caractéristiques, *AUC :* area under the receiver operating characteristic (ROC) curve, *ACC :* Accuracy, *SVM :* Support Vector Machine, *XGBoost :* Extreme Gradient Boosting

L'objectif principal de cette étude est d'analyser l'impact des caractéristiques du dataset sur la qualité du modèle prédictif de la RD. Pour atteindre cet objectif, nous avons utilisé deux datasets publics distincts et entraîné diverses techniques d'apprentissage automatique afin de développer notre modèle prédictif. En utilisant ces deux datasets, nous avons pu évaluer comment le nombre des caractéristiques pertinentes influence les performances du modèle.

En outre, nous avons effectué une comparaison des différentes techniques pour identifier le modèle le plus performant en termes d'AUC-ROC (aire sous la courbe ROC). Les résultats de cette comparaison ont indiqué que le modèle CatBoost a montré les performances les plus élevées pour les deux datasets.

La suite de ce papier est structurée comme suit : dans la section II, nous présenterons les datasets utilisés. La section III détaillera notre méthodologie de travail, mettant en évidence les différentes phases du processus de développement du modèle prédictif ainsi que la comparaison des différentes méthodes. La section IV sera dédiée à la présentation des résultats obtenus, tandis que la section V présentera notre conclusion.

II. Choix du dataset

Pour cette étude, nous avons utilisé deux datasets accessibles au public. Le premier dataset, appelé dataset1 contient un total de 844 patients dont 368 atteints de la RD et 476 sains. Ce dataset contient 11 caractéristiques en entrée dont l'âge, l'urea, la créatine, l'hémoglobine glyquée, le cholestérol,etc., et une seule sortie (Class). Les détails des caractéristiques du dataset1 sont résumés dans la Table 2.

Le deuxième dataset, appelé dataset2, est moins large en nombre de patients et plus riche en caractéristiques. Il contient 133 patients dont 42 atteints de la RD et 91 sains. Il fournit 23 caractéristiques en entrée dont le sexe, l'âge, l'IMC, le type de diabète et sa durée, l'hémoglobine glyqué, etc., et une sortie (Ret_pathy) (voir Table 3).

TABLE 2 Caractéristiques des patients (Dataset 1)

	Description
Gender	Sexe du patient
AGE	Âge du patient
Urea	Déchet azoté éliminé dans l'urine
Cr	Créatine, acide organique azote
HbA1c	Hémoglobine glyquée
Chol	Cholestérol
TG	Triglycérides
BMI	Indice de masse corporelle
HDL	Lipoprotéine de haute densité
LDL	Lipoprotéines de basse densité
VLDL	Lipoprotéine de très basse densité

TABLE 3 Quelques caractéristiques des patients (Dataset 2)

	Description
Sexe	Sexe du patient
Age	Âge du patient
BMI	Indice de masse corporelle
DM_type	Type du diabète (1 ou 2)
DM_duration	Durée du diabète
FBS	Syndrome du rachis opéré
A1C	Hémoglobine glyquée
LDL	Lipoprotéines de basse densité
HDL	Lipoprotéine de haute densité
TG	Triglycérides

Dose_Í2_if_BID	La dose de traitement.
Neu_pathy	Neuropathie
Neph_pathy	Néphropathie

III. Méthodologie

La méthodologie adoptée dans cette étude repose sur plusieurs étapes clés dans le processus de développement du modèle prédictif de la rétinopathie diabétique.

Le modèle est composé de trois étapes principales :

- Le prétraitement des données.
- La construction du modèle de prédiction de la RD en utilisant les techniques d'apprentissage automatique.
- Comparaison des différentes techniques.

Le processus global est décrit dans la figure Fig.2.

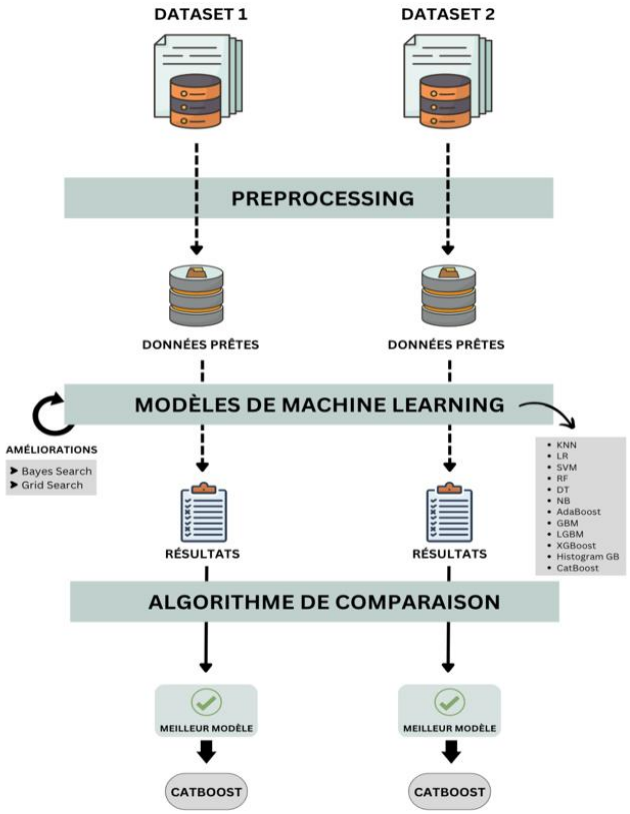


Fig.2 Schéma résumant la méthodologie

A. *Prétraitement des données*

- *Prétraitement du Dataset 1*

Nous avons d'abord procédé à une suppression des entrées à valeurs manquantes du dataset de manière à assurer la qualité des données utilisées dans notre modèle prédictif. Ensuite, nous avons effectué un encodage pour les caractéristiques catégorielles pour les préparer à l'apprentissage automatique. Enfin, nous avons divisé nos données en des données d'entrainement représentant 80%, et des données de tests représentant 20%. De plus, nous avons réservé 10 patients pour une validation interne afin d'évaluer les performances de notre modèle de manière indépendante. La Fig. 3 illustre le prétraitement du dataset 1.

- *Prétraitement du Dataset 2*

Comme pour le dataset1, les entrées contenant des valeurs manquantes ont été supprimées ainsi que des caractéristiques non pertinentes. De même, les caractéristiques catégorielles ont été encodées. Pour ce dataset, nous avons réservé 5 patients pour une validation interne. Les données d'entrainement représentent 70% et celles du test représente 30% (voir Fig. 4).

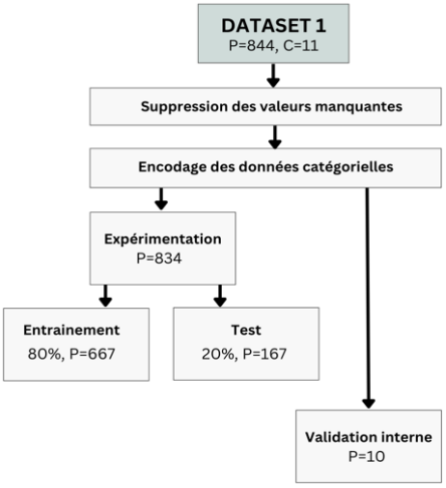
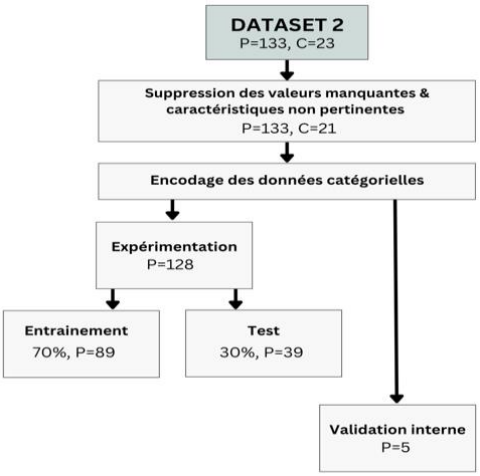


Fig.3 Prétraitement du dataset1



P : nombre de patients, *C :* nombre de caractéristiques

Fig.4 Prétraitement du dataset 2

B. *Méthodes d'apprentissage*

Dans le cadre de cette étude, nous avons appliqué, sur les deux datasets, 12 modèles d'apprentissage automatique : Adaboost, Catboost, DT, RF, GBM, Histogram GB, KNN, LGBM, Naive bayes, Régression Logistic, SVM et XGBoost. L'objectif principal est de mettre en évidence l'impact du nombre de caractéristiques pertinentes sur les performances du modèle prédictif.

Chaque modèle a été entraîné et évalué indépendamment des autres sur les deux datasets. Pour garantir les meilleurs résultats possibles des différents modèles, nous avons eu recours à deux techniques d'optimisation d'hyperparamètres à savoir GridSearch et BayesSearch. Ces techniques nous ont permis d'explorer différentes combinaisons d'hyperparamètres pour chaque modèle afin de trouver les valeurs optimales qui maximisent les performances de prédiction.

Afin de comparer les performances des différents modèles, nous avons utilisé un algorithme de comparaison qui nous a permis d'observer leurs performances respectives. Les résultats numériques obtenus seront présentés dans la section suivante.

IV. Résultats et Discussion

Dans cette section, nous présenterons les résultats de notre étude comparative des différents modèles d'apprentissage automatique appliqués sur les deux datasets pour la prédiction de la rétinopathie diabétique.

Tout d'abord, pour chaque modèle nous avons trouvé les meilleures combinaisons d'hyperparamètres qui maximisent les performances de prédiction. Lors de l'evaluation, nous avons considéré la métrique AUC pour estimer la performance globale du modèle. De plus, nous avons calculé la précision et l'accuracy (ACC) pour avoir une mesure plus détaillée des performances.

Comme le montrent les tables 4 et 5, les 12 méthodes utilisées ont donné de meilleurs résultats sur le dataset2 que sur le dataset1 en termes des trois métriques suscitées. Rappelons que le dataset2 est celui qui a le plus de carctéristiques. Ceci montre clairement l'influence du nombre de carctéristiques pertinentes sur les performances du modèle.

TABLE 4 Résultats du dataset 1

Modèle	AUC	ACC	Prec
AdaBoost	0.521	0.497	0.423
CatBoost	**0.615**	0.580	0.8
DT	0.590	0.622	0.782
GB	0.537	0.592	0.833
Histogram GB	0.516	0.550	0.521
KNN	0.450	0.532	0.458
LGBM	0.576	0.580	0.875
NB	0.513	0.520	0.447
RF	0.552	0.610	0.823
RL	0.478	0.526	0.44
SVM	0.563	0.556	1
XGBoost	0.529	0.598	0.714

TABLE 5 Résultats du dataset 2

Modèle	AUC	ACC	Prec
AdaBoost	0.755	0.717	0.692
CatBoost	**0.855**	0.794	0.9
DT	0.699	0.743	0.8
GB	0.739	0.692	0.642
Histogram GB	0.835	0.743	0.75
KNN	0.565	0.615	0.6
LGBM	0.853	0.794	0.9
NB	0.808	0.769	0.733
RF	0.841	0.769	0.888
RL	0.796	0.692	0.75
SVM	0.703	0.589	0
XGBoost	0.817	0.717	0.727

La précision et l'accuracy (ACC) ont également été calculées et la courbe ROC a été tracée comme illustré dans les figures Fig.5 et Fig.6.

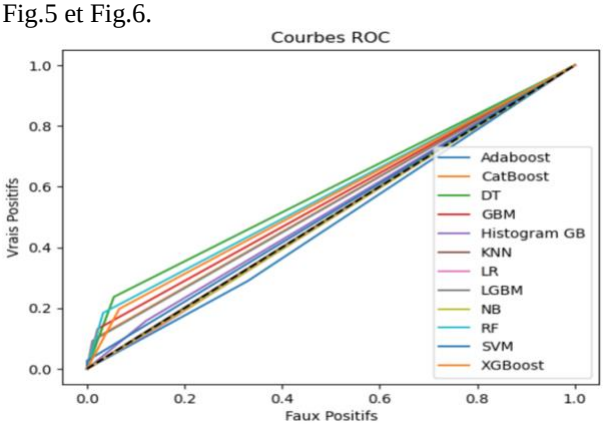


Fig.5 La courbe ROC de tous les modèles en utilisant le dataset 1

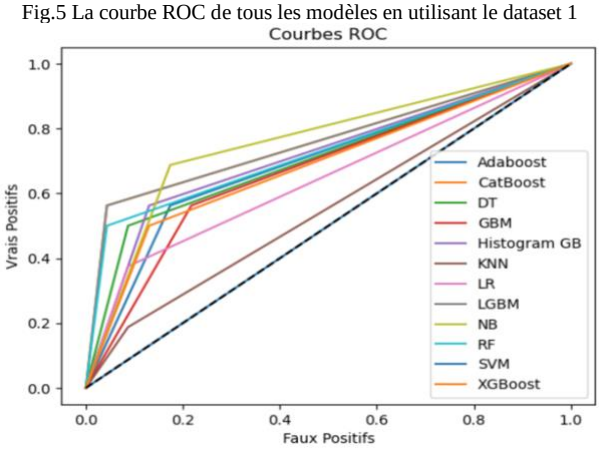


Fig.6 La courbe ROC de tous les modèles en utilisant le dataset 2

En ce qui concerne le meilleur modèle prédictif, dans cette étude, les résultats de comparaison ont montré que le modèle CatBoost appliqué au dataset1 a obtenu la valeur la plus élevée de l'AUC (AUC=0,615), dépassant ainsi les autres méthodes.

De même pour le dataset2, CatBoost reste le plus performant (AUC=0.855) suivi de très près par le modèle LGBM avec une AUC de 0.853.

Les performances des autres modèles de prédiction sont indiquées dans Table 4 et Table 5 et leur comparaison est résumée dans Fig. 8 et Fig. 9 pour le dataset1 et le dataset2, respectivement.

Lors de la validation interne, nous avons utilisé une portion des données, spécifiquement 10 patients pour le dataset1 et 5 patients pour le dataset2, pour évaluer les performances du modèle CatBoost. Les résultats obtenus ont montré une capacité élevée du modèle à discriminer entre les deux classes, ce qui signifie qu'il est capable de différencier avec précision les patients sains des patients atteints de la RD.

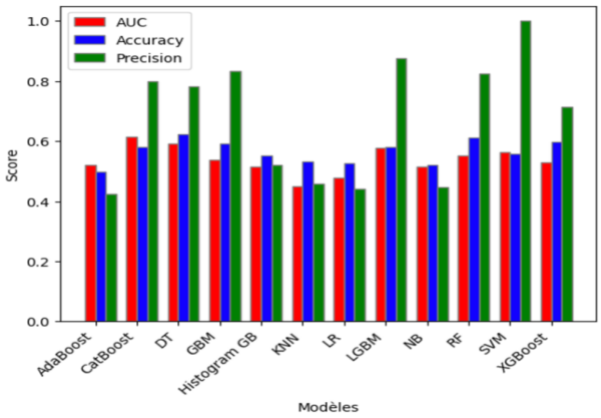


Fig.8 Comparaison des performaces des modèles sur le dataset 1

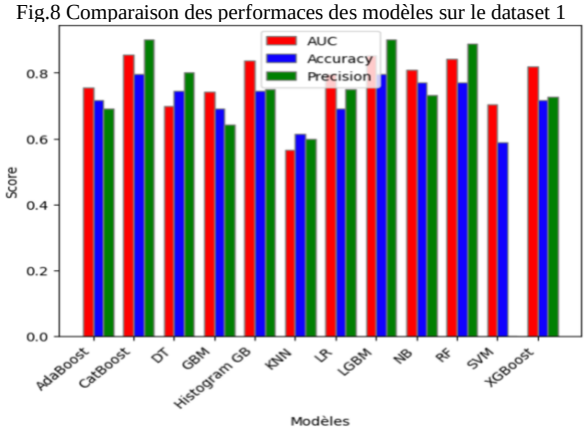


Fig.9 Comparaison des performaces des modèles sur le dataset 2

Les graphiques présentés dans les figures Fig.8 et Fig.9, mettent en évidence l'impact significatif des différences entre les datasets sur les résultats obtenus. Malgré l'utilisation exacte des mêmes modèles d'apprentissage automatique, les résultats obtenus à partir du dataset2 se sont révélés plus élevés et plus prometteurs pour la prédiction de la rétinopathie diabétique, malgré le nombre de patients moins élevé par rapport au dataset1. Nous avons observé que de nouvelles caractéristiques présentes dans le dataset2 tels que la durée du diabète et la présence de la nephropathy ont eu un impact significatif sur la prédiction. Cela indique que le nombre et le type de caractéristiques sont des facteurs clés et influants sur les performances des méthodes de prédiction de la rétinopathie diabétique. La Table 6 résume les meilleurs modèles de prédiction pour chacun des deux datasets.

TABLE 6 Résultats des deux datasets par rapport au nombre de patients et de caractéristiques

Modèle	Dataset	AUC %
CatBoost	Dataset 1	61.596
	P=844 C=**11**	
CatBoost	Dataset 2	85.597
	P=133 C=**21**	

V.Conclusion

Dans cette étude nous nous sommes concentrés sur l'impact du nombre de caractéristiques pertinentes sur les performances du modèle prédictif.

En utilisant deux datasets différents en termes de nombre de caractéristiques importantes, nous avons pu mettre en évidence que le choix du modèle de prédiction du risque de rétinopathie diabétique peut être considérablement influencé par le type et le nombre de caractéristiques représentatives. En effet, les résultats obtenus, après expérimentation de 12 techniques d'apprentissage sur les deux datasets, ont montré que les performances sont nettement plus élevées avec le dataset2 ayant le plus de caractéristiques influentes.

Par conséquent, avant d'établir le facteur de risque, il est essentiel d'identifier le maximum de caractéristiques pertinentes qui auront un impact sur la prise de décision. La validation sur un large dataset renforcera la fiabilité et l'applicabilité du modèle, offrant ainsi un outil précieux pour les professionnels de la santé dans la prédiction du risque de rétinopathie diabétique.

Acknowledgment

Ce travail a été sponsorisé par la Direction Générale de la Recherche Scientifique et du Developpement Technologique, Ministere de l'Enseignement Superieur et de la Recherche Scientifique (DGRSDT), Algerie.

REFERENCES

[1] Findings on Teaching Machine Learning in High School:
A Ten - Year ystematic Literature Review, Ramon Mayor MARTINS, Christiane GRESSE VON WANGENHEIM, March 2022
https://www.infedu.vu.lt/journal/INFEDU/article/742/info

[2] B. Wolff, P. Baudouin, J.-F. Girmens, G. Quentel, J.-A. Sahel, P. Massin. La rétinopathie diabétique non proliférante. EMC - Ophtalmologie 2018;15(4):1-12 [Article 22-018-G-10], Rétine et viré (consulté le 01/12/2022) https://www.em-consulte.com

[3] Mazari, Fettouma, and Karim Ait Idir. "Rétinopathie diabétique entre le diagnostic, la classification et le rythme de surveillance." *Med Sci* 4.1 (2017): 5-9.

[4] Tsao, Hsin-Yi, Pei-Ying Chan, and Emily Chia-Yu Su. "Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms." BMC bioinformatics 19 (2018): 111-121.

[5] Jian, Yazan, et al. "A machine learning Approach to predicting diabetes complications." Healthcare. Vol. 9. No. 12. MDPI, 2021.

[6] SKMCA. Rashid Centre for Diabetes & Research. Available online: https://www.skmca.ae/rashid-centre-for-diabetes-research/ (accessed on 21 January 2021).

[7] Li, Wanyue, et al. "Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China." BMJ open 11.11 (2021): e050989.

[8] van der Heijden, Amber A., et al. "Prediction models for development of retinopathy in people with type 2 diabetes: systematic review and external validation in a Dutch primary care setting." *Diabetologia* 63 (2020): 1110-1119.