Application de l'Intelligence Artificielle pour la Prédiction de Résultats Biomédicaux à Partir de Données Cliniques Incomplètes : Une Étude Basée sur la Base MIMIC-II

By Keke Mahuvivi Turibio[1,2], Houessouvo C. Roland[1,2], Jossou R. Thierry[1,2], Medenou Daton[1,2]

[1] Department of Biomedical Engineering (GBM), École Polytechnique d'Abomey-Calavi (EPAC), Université d'Abomey-Calavi, Abomey-Calavi, 01 BP 2009 Cotonou, Benin
[2] Laboratoire d'Electrotechnique, de Télécommunication et d'Informatique Appliquée (LETIA), Université d'Abomey-Calavi, Abomey-Calavi, 01 BP 2009 Cotonou, Benin
[3]École Supérieure Des Techniques Administratives et de Gestion "Saint Christophe" (ESTAG - SC), Cotonou, 03 BP 1662 Cotonou, Bénin.

Auteur Correspondant : KEKE Mahuvivi Turibio (mahuvivituribiok@gmail.com)

Résumé

Dans de nombreux contextes médicaux à faibles ressources, l'accès aux analyses biomédicales complètes est sévèrement limité par des contraintes financières, infrastructurelles et logistiques. Cette situation contraint les professionnels de la santé à prendre des décisions diagnostiques critiques avec des données partielles. Cette étude explore l'application de l'intelligence artificielle (IA) pour surmonter ce défi en prédisant les valeurs de laboratoire manquantes à partir d'un ensemble réduit de paramètres disponibles. En utilisant la base de données cliniques MIMIC-II, qui contient des informations sur plus de 25 000 patients en unités de soins intensifs, nous avons entraîné et évalué plusieurs modèles d'apprentissage automatique. L'objectif était d'estimer avec précision des paramètres biochimiques essentiels tels que le potassium, le phosphate, la créatinine, l'urée et le sodium. Après un processus rigoureux de prétraitement des données, de sélection des variables et d'optimisation des modèles, l'algorithme Random Forest Regressor a démontré les performances les plus robustes, mesurées par l'erreur quadratique moyenne (RMSE). Les résultats confirment qu'il est possible de prédire des biomarqueurs avec une précision suffisante pour une utilisation clinique de première intention. Ce système offre une voie prometteuse pour réduire les coûts de santé, accélérer le processus de diagnostic et améliorer l'équité d'accès aux soins, notamment dans les environnements isolés ou mal équipés. Ce travail met en évidence le potentiel transformateur de l'IA pour combler les lacunes diagnostiques et soutenir une médecine prédictive et personnalisée.

Mots-clés : Intelligence artificielle, apprentissage automatique, prédiction biomédicale, analyse de laboratoire, soins intensifs, santé en milieu à ressources limitées, Random Forest, prétraitement des données, MIMIC-II.

1. Introduction

L'accès à des soins de santé de qualité demeure l'un des défis les plus pressants à l'échelle mondiale, particulièrement dans les pays à faibles et moyens revenus[1]. Dans ces régions, les systèmes de santé sont souvent confrontés à un manque criant d'infrastructures, de personnel qualifié et de ressources financières[2]. Les zones rurales et isolées sont les plus touchées, les centres de santé y étant fréquemment dépourvus d'équipements de diagnostic modernes[3]. Les analyses biomédicales, qui sont pourtant fondamentales pour le diagnostic, le suivi et la prise en charge des patients, restent inaccessibles pour une grande partie de la population[4]. Les obstacles sont multiples : le coût élevé des tests, la rupture de stock des réactifs, la complexité logistique de l'acheminement des prélèvements et le manque de laboratoires spécialisés[5].

Face à ce constat, les cliniciens sont souvent contraints de poser des diagnostics et d'initier des traitements sur la base d'informations cliniques incomplètes, ce qui peut compromettre la qualité des soins et la sécurité des patients[6]. Cette situation crée une disparité inacceptable dans l'accès à la santé et entrave les efforts de médecine préventive.

L'émergence de l'intelligence artificielle (IA) et de l'apprentissage automatique (*machine learning*) offre des perspectives nouvelles et prometteuses pour transformer la prestation de soins dans ces contextes difficiles[7]. En exploitant la puissance des algorithmes pour analyser de vastes ensembles de données, l'IA peut aider à identifier des schémas et des corrélations invisibles à l'œil humain[8]. Une application particulièrement intéressante est la prédiction de paramètres biologiques manquants à partir d'un ensemble limité de données disponibles et peu coûteuses[9]. Par exemple, serait-il possible d'estimer avec une précision clinique acceptable le taux de créatinine ou de potassium d'un patient en se basant uniquement sur des paramètres plus simples comme l'urée, le poids et le genre ?

C'est l'hypothèse centrale de cette étude.

L'objectif principal est de concevoir, d'entraîner et d'évaluer des modèles d'apprentissage automatique capables de prédire les résultats de tests biomédicaux clés à partir d'un sous-ensemble de données cliniques et démographiques incomplètes[10]. Pour ce faire, nous nous appuierons sur la base de données publique et anonymisée MIMIC-II, une ressource riche en informations cliniques issues d'unités de soins intensifs. En démontrant la faisabilité d'une telle approche, nous espérons ouvrir la voie au développement d'outils d'aide à la décision clinique, peu coûteux et déployables sur des plateformes mobiles, afin de soutenir les professionnels de santé dans les environnements à ressources limitées et de promouvoir une médecine plus équitable et préventive[11]. Cette recherche s'inscrit dans un effort global visant à rendre la médecine de précision plus accessible, en transformant les données existantes en connaissances actionnables pour le bien de tous les patients, où qu'ils se trouvent[12].

2. État de l’art

2.1. Intelligence Artificielle et Médecine

L'intégration de l'intelligence artificielle en médecine a catalysé des avancées remarquables au cours de la dernière décennie[13]. Au-delà de l'analyse d'images médicales (radiologie, pathologie), où les algorithmes de *deep learning* atteignent et parfois surpassent les performances humaines, l'IA s'est

étendue à de nombreux autres domaines. Elle est aujourd'hui utilisée pour la prédiction des risques de maladies, l'optimisation des plans de traitement, la surveillance en temps réel des patients en soins intensifs et l'aide au diagnostic[14]. Des études ont montré son efficacité dans la prédiction de la septicémie, de l'insuffisance rénale aiguë ou encore de la mortalité en USI[15]. Ces succès reposent sur la capacité des modèles à traiter et à synthétiser des données hétérogènes et massives (signaux physiologiques, résultats de laboratoire, notes cliniques) pour en extraire des informations prédictives[16].

La prédiction de biomarqueurs spécifiques est un domaine particulièrement actif. En analysant les corrélations entre différents paramètres biologiques, les modèles d'IA peuvent "inférer" la valeur d'un test manquant ou coûteux à partir de tests plus courants. Cette approche est prometteuse pour le dépistage précoce et le suivi des maladies chroniques dans des contextes où des panels de tests complets ne sont pas réalisables[17].

2.2. Données Cliniques Ouvertes : La Révolution MIMIC

Le développement de modèles d'IA performants et fiables nécessite l'accès à de vastes ensembles de données cliniques de haute qualité. Cependant, la sensibilité et la confidentialité des données de santé ont longtemps constitué un frein majeur à la recherche. L'initiative MIMIC (*Multiparameter Intelligent Monitoring in Intensive Care*), développée par le MIT, a changé la donne[18]. En fournissant un accès libre et gratuit à des bases de données riches et anonymisées (MIMIC-II, MIMIC-III, et plus récemment MIMIC-IV), elle a démocratisé la recherche en informatique médicale[19].

Ces bases de données contiennent des dizaines de milliers de dossiers de patients admis en unités de soins intensifs, incluant des données démographiques, des signes vitaux, des résultats de laboratoire, des prescriptions médicamenteuses et des notes cliniques[20]. L'anonymisation est réalisée via des processus rigoureux, notamment le décalage des dates et la suppression des identifiants directs, garantissant la protection de la vie privée des patients tout en préservant l'intégrité temporelle et clinique des données[21]. Grâce à MIMIC, des centaines d'études ont pu être menées pour développer et valider de nouveaux algorithmes prédictifs, faisant progresser notre compréhension des maladies critiques[22].

2.3. Algorithmes de Prédiction en Contexte Biomédical

Plusieurs familles d'algorithmes d'apprentissage automatique ont été appliquées avec succès à la prédiction de variables biomédicales.

- **Modèles linéaires (ex: Régression Linéaire, Lasso) :** Simples et interprétables, ils sont souvent utilisés comme modèles de référence. Cependant, leur hypothèse de linéarité est une limite dans le domaine médical où les relations entre variables sont souvent complexes et non linéaires.
- **Approches à base d'arbres (ex: Arbres de Décision, Random Forest, XGBoost) :** Ces algorithmes sont particulièrement populaires et performants pour les données tabulaires cliniques[23]. Le
- **Random Forest**, en agrégeant les prédictions de multiples arbres de décision, est robuste au surapprentissage et gère bien les interactions complexes entre variables[24].
- **XGBoost** (*Extreme Gradient Boosting*) est une version optimisée du *Gradient Boosting* qui a souvent montré des performances de pointe dans les compétitions de *data science* et les études médicales, notamment pour la prédiction de la mortalité[25].

- **Réseaux de Neurones Profonds (*Deep Learning*) :** Ces modèles sont extrêmement puissants pour capturer des motifs complexes, en particulier dans les données non structurées (images, texte) ou les séries temporelles (ECG, EEG)[26]. Leur utilisation sur des données cliniques tabulaires est également explorée, bien qu'ils nécessitent de très grandes quantités de données et soient plus difficiles à interpréter (effet "boîte noire")[27].

Le choix de l'algorithme dépend de la nature des données, de la taille de l'échantillon et du besoin d'interprétabilité du modèle. Pour notre étude, axée sur des données tabulaires et un besoin de robustesse, les approches à base d'arbres comme Random Forest et XGBoost constituent des candidats de choix[28].

3. Matériel et Méthodes

3.1. Source des Données : La Base MIMIC-II

Cette étude a été menée en utilisant la base de données publique MIMIC-II (version 2.6)[29]. Cette base contient des données cliniques anonymisées de plus de 25 000 patients admis dans les unités de soins intensifs (USI) du *Beth Israel Deaconess Medical Center* à Boston, entre 2001 et 2008[30]. Les données sont structurées dans une base de données relationnelle de 38 tables. Pour notre travail, nous avons principalement utilisé les trois tables suivantes :

1. labevents : Contient les résultats de tous les tests de laboratoire, avec un horodatage (charttime) pour chaque mesure[31].
2. d_labitems : Fournit les métadonnées de chaque test (nom, fluide corporel, catégorie)[32].
3. icustay_detail : Contient des informations démographiques et anthropométriques sur les patients pour chaque séjour en USI (sexe, taille, poids)[33].

À partir des 713 types de tests de laboratoire disponibles, nous avons sélectionné un sous-ensemble de 10 biomarqueurs biochimiques sanguins pertinents pour le suivi en soins intensifs[34]. Les variables démographiques (sexe, taille, poids) ont également été incluses. Le tableau ci-dessous résume les variables initialement sélectionnées pour constituer notre jeu de données.

Tableau 1 : Variables sélectionnées pour la construction du jeu de données

Catégorie	Variable	Description
Biomarqueurs	Calcium	Calcémie
	Chlorure	Chlorémie
	Glucose	Glycémie

Catégorie	Variable	Description
	Créatinine	Créatininémie
	Magnésium	Magnésémie
	Phosphate	Phosphatémie
	Potassium	Kaliémie
	Sodium	Natrémie
	Urée	Urémie
	Lithium	Lithiémie (exclu par la suite)
Démographie	gender	Sexe du patient (M/F)
	height	Taille du patient (cm)
	weight_avg	Poids moyen du patient durant le séjour (kg)

Le jeu de données final a été construit en fusionnant ces tables. Chaque ligne du dataset représente un ensemble de mesures de laboratoire pour un patient donné, à un instant charttime précis, enrichi des données démographiques correspondantes. Le dataset final comprenait

323 699 enregistrements et 15 colonnes avant le nettoyage final[35].

3.2. Prétraitement des Données

Un prétraitement rigoureux a été appliqué pour garantir la qualité et la cohérence des données avant la modélisation[36].

La figure suivante présente une heatmap pour la visualisation des valeurs manquantes, les zones non en noir représentent des valeurs manquantes.

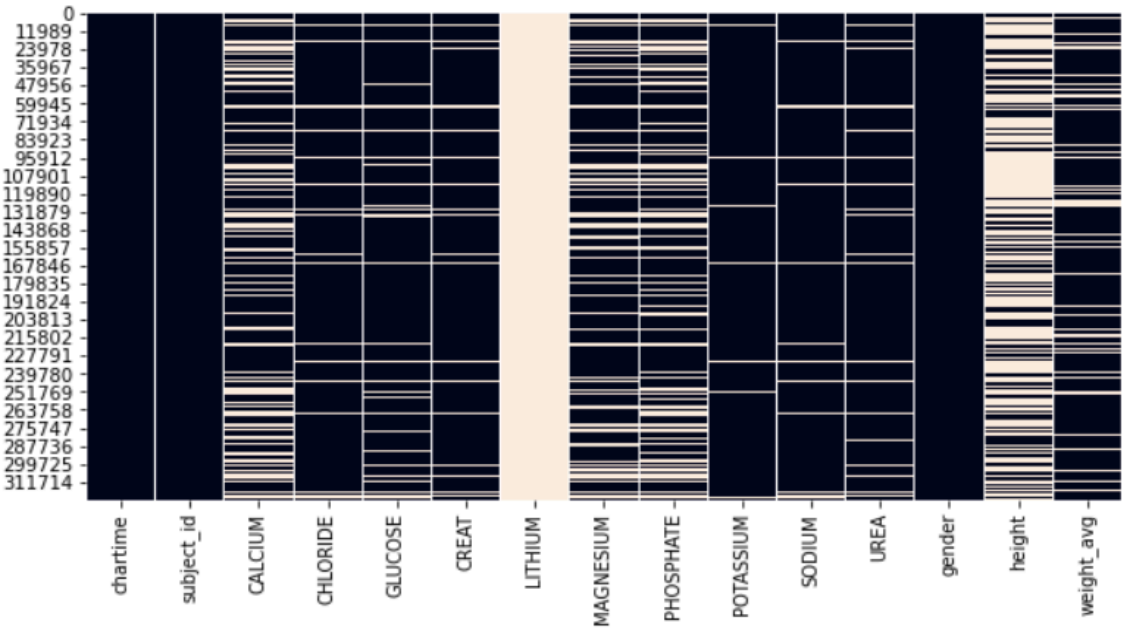


Figure 3: Représentation graphique des données statistiques montrant les valeurs manquantes

1. **Gestion des valeurs manquantes :** Une analyse initiale a révélé un taux élevé de données manquantes[37].
    - **Suppression de colonnes :** Les variables avec plus de 70% de valeurs manquantes ont été supprimées. Ce fut le cas de la lithiémie, qui présentait plus de 90% de données manquantes et a donc été exclue de l'analyse[38].
    - **Imputation :** Pour les valeurs manquantes restantes :
        - Les variables **numériques** (biomarqueurs) ont été imputées en utilisant la **médiane** de chaque colonne. La médiane a été préférée à la moyenne en raison de la distribution souvent asymétrique des valeurs biologiques[39].
        - La variable **catégorielle** `gender` a été imputée avec le **mode** (la valeur la plus fréquente, 'M' dans ce cas)[40].
2. **Traitement des valeurs aberrantes (*outliers*) :** Pour limiter l'influence des valeurs extrêmes potentiellement erronées, une technique de *clipping* basée sur les quantiles a été utilisée. Les valeurs inférieures au 5ème percentile et supérieures au 95ème percentile de chaque variable ont été ramenées à ces seuils respectifs[41].
3. **Encodage et Normalisation :**
    - La variable catégorielle `gender` a été transformée en variables numériques via un encodage *one-hot*[42].
    - Toutes les variables numériques ont été mises à l'échelle entre 0 et 1 à l'aide d'un `MinMaxScaler` pour garantir que tous les prédicteurs aient une importance égale au début du processus d'entraînement[43].

3.3. Sélection des Variables (*Feature Selection*)

Pour comprendre les relations entre les variables et sélectionner les prédicteurs les plus pertinents, deux approches ont été utilisées :

1. **Analyse de corrélation :** La matrice de corrélation de Pearson a été calculée et visualisée sous forme de carte de chaleur (*heatmap*) pour identifier les relations linéaires entre les variables[44]. Une corrélation, même modérée, entre les biomarqueurs justifie l'approche prédictive (voir
2. **Annexe B**).
3. **Sélection par tests univariés :** La méthode SelectKBest de la bibliothèque Scikit-learn a été employée[45]. Cette technique évalue la pertinence de chaque variable prédictive par rapport à la variable cible à l'aide de tests statistiques (tests F pour la régression). Elle a permis, pour certains modèles, de ne conserver que les
4. k variables les plus informatives, réduisant ainsi la complexité et le risque de surapprentissage[46].

La figure suivante présente une heatmap pour la visualisation des corrélation entre les variables.

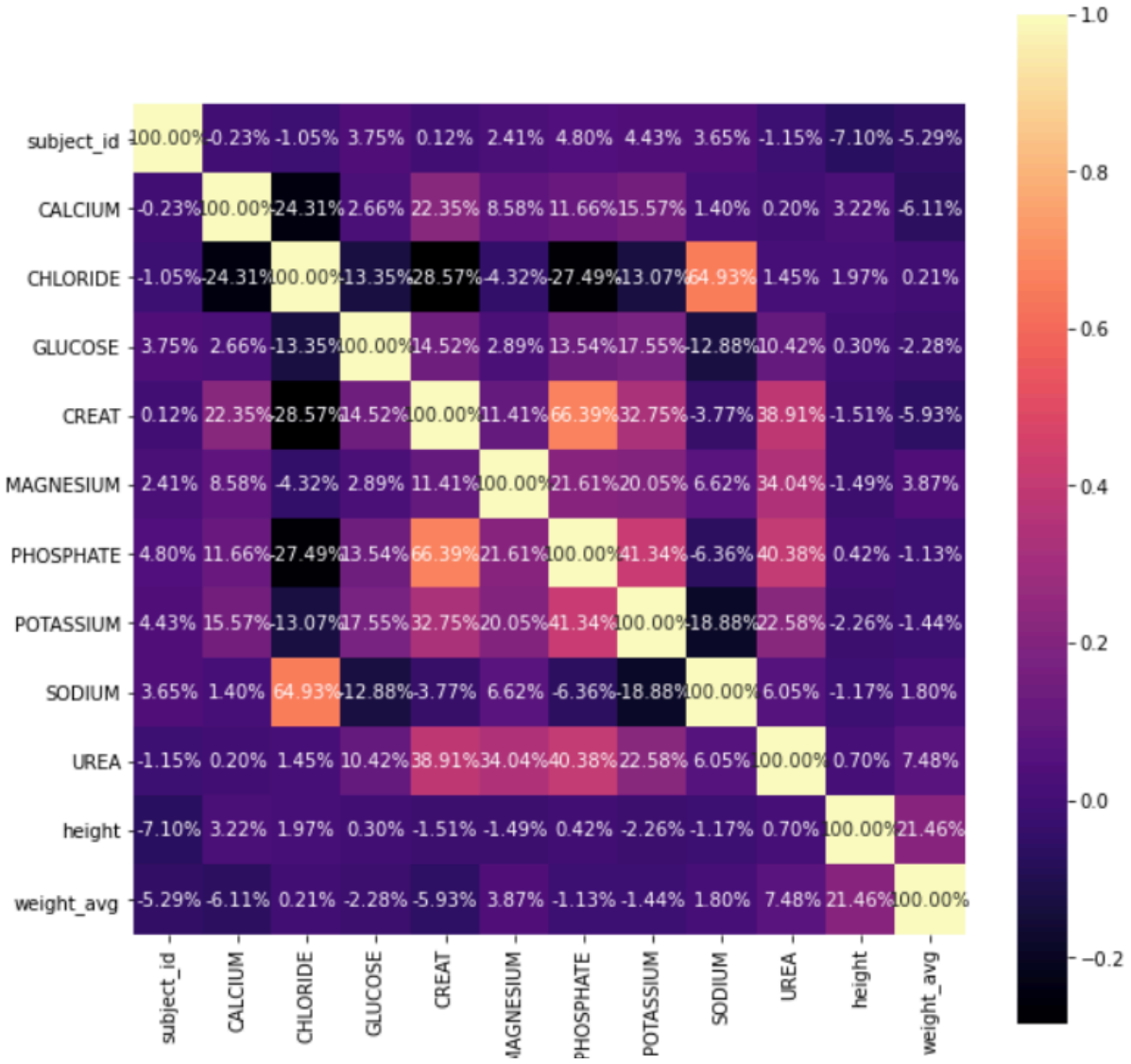


Figure 4: Heatmap de la corrélation linéaire entre variables

3.4. Modélisation

Cinq biomarqueurs clés ont été définis comme variables cibles à prédire :

Phosphate, Créatinine, Urée, Sodium et Potassium [47]. Pour chaque cible, un modèle de régression distinct a été développé.

Six algorithmes d'apprentissage automatique ont été testés et comparés [48] :

- Régression Linéaire
- Régression Lasso
- Arbre de Décision (*Decision Tree Regressor*)
- Forêt Aléatoire (*Random Forest Regressor*)
- *Gradient Boosting Regressor*

- *XGBoost Regressor*

Le jeu de données a été divisé en un ensemble d'entraînement (80%) et un ensemble de test (20%)[49]. Des pipelines de prétraitement et de modélisation ont été construits pour systématiser les étapes. Pour les modèles les plus performants, une optimisation des hyperparamètres a été réalisée à l'aide de

RandomizedSearchCV, une technique qui explore aléatoirement une grille de paramètres pour trouver la meilleure combinaison en se basant sur la validation croisée[50].

3.5. Stratégie d'Évaluation des Modèles

La performance des modèles a été évaluée à l'aide de plusieurs métriques de régression, calculées sur l'ensemble de test (données non vues pendant l'entraînement)[51].

- **RMSE (Root Mean Squared Error - Erreur Quadratique Moyenne) :** C'est la métrique principale utilisée. Elle pénalise fortement les grandes erreurs et est exprimée dans la même unité que la variable cible, ce qui la rend plus interprétable[52]. Un RMSE plus faible indique un meilleur modèle.
- **R^2 (Coefficient de Détermination) :** Indique la proportion de la variance de la variable cible qui est prévisible à partir des variables indépendantes. Une valeur proche de 1 indique un bon ajustement du modèle.
- **MAE (Mean Absolute Error - Erreur Absolue Moyenne) :** Mesure la moyenne des erreurs absolues entre les valeurs prédites et réelles.

Une **validation croisée à 10 plis** a été utilisée durant la phase d'optimisation pour assurer la robustesse et la capacité de généralisation des modèles[53]. Enfin, l'intervalle de confiance à 95% pour le RMSE a été calculé sur l'ensemble de test pour estimer la stabilité des performances du modèle[54].

4. Résultats

4.1. Performances Comparées des Modèles

Après entraînement et évaluation des six algorithmes sur les cinq variables cibles, une tendance claire s'est dégagée : le **Random Forest Regressor a systématiquement surpassé les autres modèles** sur l'ensemble des cibles, en se basant sur la métrique RMSE[55]. Les modèles linéaires se sont montrés insuffisants pour capturer la complexité des relations, tandis que les modèles de

boosting (Gradient Boosting, XGBoost) ont montré des performances compétitives mais légèrement inférieures à celles du Random Forest dans ce cas précis.

Le tableau ci-dessous présente les performances finales des modèles Random Forest optimisés sur l'ensemble de test pour chaque biomarqueur cible.

Tableau 2 : Performances des modèles Random Forest optimisés sur l'ensemble de test

Variable Cible	RMSE (entraînement)	**RMSE (test)**	Intervalle de Confiance à 95% (RMSE test)

Potassium	0.156	**0.288**	[0.284, 0.291]
Phosphate	0.164	**0.398**	[0.390, 0.404]
Urée	0.164	**0.415**	[0.408, 0.421]
Créatinine	0.217	**0.577**	[0.567, 0.588]
Sodium	0.694	**1.575**	[1.558, 1.595]

Les valeurs de RMSE sont exprimées dans les unités respectives de chaque biomarqueur (ex: mEq/L pour le potassium).

Les résultats montrent des erreurs de prédiction relativement faibles pour le potassium, le phosphate et l'urée, indiquant une bonne capacité du modèle à estimer ces valeurs. Les erreurs pour la créatinine et surtout le sodium sont plus élevées, suggérant que leur prédiction est plus complexe ou que les variables disponibles sont moins informatives pour ces cibles spécifiques. L'intervalle de confiance étroit pour chaque RMSE témoigne de la robustesse et de la stabilité des prédictions du modèle[56].

4.2. Analyse des Erreurs et Importance des Variables

L'analyse de l'importance des variables (*feature importance*) du modèle Random Forest a révélé que les prédicteurs les plus influents n'étaient pas toujours les mêmes pour toutes les cibles. Cependant, un schéma général est apparu :

- Les variables anthropométriques,
  le poids (weight_avg) et la taille (height), se sont révélées être des prédicteurs très importants pour la plupart des cibles[57].
- D'autres biomarqueurs se sont également avérés prédictifs, confirmant l'existence de corrélations physiologiques exploitées par le modèle.
- Le genre (gender) a également contribué de manière significative aux prédictions.

Le graphique ci-dessous illustre l'importance relative des variables pour la prédiction d'un des biomarqueurs.

*Légende : Importance relative des variables (prédicteurs) selon le modèle Random Forest. Une valeur plus élevée indique une plus grande contribution de la variable à la prédiction. Voir une version détaillée en **Annexe C**.*

Ces résultats confirment que les modèles d'IA, et en particulier le Random Forest, sont capables de prédire des valeurs d'analyses manquantes avec une précision qui, pour plusieurs biomarqueurs, pourrait être suffisante pour des usages cliniques d'orientation ou de première intention[58].

5. Discussion

5.1. Interprétation des Résultats et Implications Cliniques

Les résultats de cette étude démontrent de manière convaincante la faisabilité de la prédiction de résultats biomédicaux à partir d'un jeu de données cliniques limité. La performance supérieure de l'algorithme Random Forest s'explique par sa capacité à modéliser des relations non linéaires complexes et des interactions entre les variables, ce qui est typique en biologie humaine[59]. Les corrélations observées entre les différents paramètres (voir

Annexe B) ne sont pas surprenantes d'un point de vue physiopathologique (par exemple, la relation entre la fonction rénale, l'urée et la créatinine) et justifient l'approche prédictive : l'information sur un paramètre est partiellement redondante avec d'autres[60].

Les implications cliniques de ces résultats sont potentiellement considérables, surtout pour les systèmes de santé à ressources limitées[61]. Un outil basé sur ces modèles pourrait :

1. **Réduire les coûts :** En estimant la valeur de tests coûteux ou non disponibles à partir de tests de routine, il permettrait de réduire les dépenses pour le patient et le système de santé[62].
2. **Accélérer le diagnostic :** Dans des situations d'urgence ou dans des zones isolées où le rendu des résultats de laboratoire peut prendre plusieurs heures ou jours, une estimation immédiate pourrait guider la prise de décision clinique initiale[63].
3. **Améliorer l'accès aux soins :** Un tel outil, potentiellement déployé sur un smartphone ou un ordinateur portable, pourrait équiper les agents de santé communautaires et les cliniques rurales, leur fournissant des capacités de diagnostic améliorées même en l'absence de laboratoire[64].

Par exemple, une estimation fiable du potassium (kaliémie) est cruciale, car des anomalies peuvent entraîner des arythmies cardiaques potentiellement mortelles. L'obtention rapide d'une valeur estimée pourrait alerter le clinicien et l'inciter à prendre des mesures conservatoires en attendant une confirmation.

5.2. Comparaison avec la Littérature Existante

Nos travaux s'inscrivent dans une lignée croissante d'études utilisant les bases de données MIMIC pour développer des modèles prédictifs[65]. Plusieurs études ont utilisé des approches similaires pour prédire la mortalité, la durée de séjour ou l'apparition de complications[66]. En ce qui concerne la prédiction de biomarqueurs spécifiques, nos performances (mesurées en RMSE) sont comparables, voire meilleures pour certaines variables, que celles rapportées dans d'autres travaux utilisant MIMIC-II ou MIMIC-III avec des algorithmes similaires comme Random Forest[67]. La nouveauté de notre approche réside principalement dans son objectif final : développer un outil spécifiquement pensé pour les contextes à faibles ressources, en se concentrant sur un ensemble de biomarqueurs de base et en visant un déploiement sur des plateformes légères.

5.3. Limites de l'Étude

Il est essentiel de reconnaître les limites de notre étude avant d'envisager une application clinique :

- **Origine des données :** La base MIMIC-II contient des données de patients d'unités de soins intensifs d'un seul centre hospitalier aux États-Unis[68]. Le profil démographique et pathologique de cette population peut être très différent de celui des populations africaines ou d'autres régions à faibles ressources. Les modèles entraînés sur ces données pourraient ne pas se généraliser correctement. Une
- **validation clinique locale est donc une étape indispensable** avant tout déploiement[69].
- **Qualité des données d'entrée :** La performance du modèle dépend entièrement de la précision des données d'entrée[70]. Des erreurs de mesure dans les quelques tests disponibles se répercuteront inévitablement sur la qualité des prédictions.
- **Contexte des soins intensifs :** Les corrélations apprises par le modèle sont spécifiques à une population de patients gravement malades. L'application de ces modèles à des patients en consultation externe ou moins sévèrement atteints nécessiterait un ré-entraînement sur des données plus appropriées.
- **Imputation des données :** Bien que nécessaire, l'imputation des valeurs manquantes introduit un biais potentiel. Des stratégies d'imputation plus sophistiquées pourraient améliorer les performances.

5.4. Perspectives et Travaux Futurs

Cette étude ouvre plusieurs voies de recherche passionnantes[71].

1. **Extension à d'autres biomarqueurs :** Le pipeline développé peut être étendu pour prédire d'autres paramètres importants, tels que les enzymes hépatiques (ALAT, ASAT), la bilirubine ou les marqueurs de l'inflammation (CRP).
2. **Intégration de données multimodales :** L'ajout de données cliniques textuelles (extraites des notes des médecins) ou de signes vitaux pourrait enrichir le modèle et améliorer la précision des prédictions.
3. **Développement et déploiement :** La prochaine étape concrète est de développer une interface utilisateur simple (API et application mobile) pour rendre le modèle utilisable par les cliniciens sur le terrain[72].
4. **Validation prospective et locale :** La perspective la plus importante est la mise en place d'une étude prospective en collaboration avec des centres de santé locaux pour collecter des données et valider les performances du modèle dans un contexte réel, avant d'envisager une intégration dans la pratique clinique[73].

6. Conclusion

Cette étude a démontré avec succès la faisabilité de l'utilisation de modèles d'apprentissage automatique, en particulier l'algorithme Random Forest, pour prédire avec une précision prometteuse les résultats de tests biomédicaux à partir de données cliniques incomplètes. En nous appuyant sur la vaste base de données MIMIC-II, nous avons développé des modèles capables d'estimer des paramètres vitaux comme le potassium, le phosphate ou l'urée, en se basant sur un ensemble limité d'informations accessibles.

Ces résultats valident le potentiel de l'intelligence artificielle comme un outil puissant pour surmonter certains des obstacles à l'accès aux soins dans les milieux à ressources limitées. En fournissant une estimation rapide et peu coûteuse des biomarqueurs manquants, cette technologie peut aider à réduire

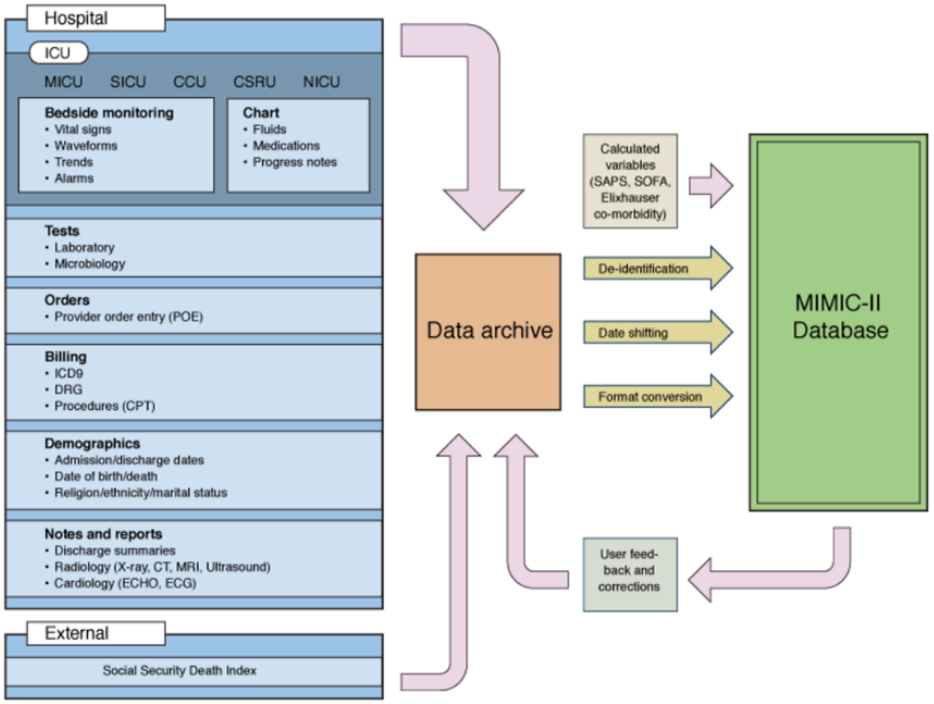
les coûts et les délais de diagnostic, à améliorer la prise de décision clinique et, finalement, à lutter contre les disparités en matière de santé.

Bien que ce travail ne soit qu'une preuve de concept et que des recherches supplémentaires soient nécessaires pour affiner et valider cliniquement ces modèles dans des contextes locaux, il jette les bases d'une nouvelle génération d'outils d'aide au diagnostic. Le projet ouvre la voie à une médecine plus accessible, préventive et personnalisée, en transformant intelligemment les données fragmentaires disponibles en informations cliniques actionnables. Nous lançons un appel à la collaboration avec les institutions cliniques et de recherche pour faire avancer ce projet de la validation à l'implémentation, avec l'espoir de contribuer à une offre de soins plus équitable pour tous.

7. Annexes

Annexe A : Schéma du Pipeline de Traitement des Données

Le schéma ci-dessous illustre les quatre étapes principales de notre pipeline IA, de la collecte des données à la modélisation.



Légende : Pipeline IA montrant les étapes de collecte des données (MIMIC-II), le prétraitement (nettoyage, imputation), la sélection des variables (Pearson, SelectKBest) et la modélisation (ex: Random Forest, XGBoost).

8. Références Bibliographiques

1. Syed, M., et al. (2021). Application of machine learning in ICU using MIMIC dataset: systematic review. *Informatics, 8*(1), 16.

2. Zhou, H., et al. (2023). A survey of large language models in medicine. *arXiv:2311.05112*.
3. Schinkel, M., et al. (2019). Clinical applications of AI in sepsis: narrative review. *Computers in Biology and Medicine, 115*, 103488.
4. Khope, S. R., & Elias, S. (2023). Strategies of predictive schemes using MIMIC-III: systematic review. *Healthcare, 11*(5), 710.
5. Sadeghi, S., et al. (2024). SICdb: Comparative analysis with MIMIC-IV. *Scientific Reports*.
6. Lim, L., & Lee, H. C. (2023). Open datasets in perioperative medicine: a review. *Anesthesia and Pain Medicine*.
7. Liu, S., et al. (2020). Analysis of disease characteristics using MIMIC-III. *PLOS ONE, 15*(4), e0232176.
8. Ranade, M. D., & Deshpande, A. (2021). Neonatal ICU data analysis using MIMIC-IV. *IEEE*.
9. Kataria, A., et al. (2022). Blood screening prediction using neural networks. In *Predictive Modeling in Biomedical Data Mining*.
10. Edin, J., et al. (2023). Automated medical coding on MIMIC: replicability study. *ACM*.
11. Mishra, R. (2010). Information transmission in the MIMIC II clinical database. *MIT*.
12. Li, Q., et al. (2019). BiLSTM-based severity prediction in ICU. *IEEE BIBM*.
13. Hadweh, P., et al. (2025). AI in Intensive Care Medicine. *Journal of Clinical Informatics*.
14. Yang, S., et al. (2023). Predicting ICU mortality with XGBoost on MIMIC. *JMDH*.
15. Liu, S., Wu, J., & Li, M. (2020). Predictive analysis for diseases based on MIMIC-II. *PLOS ONE*.