Comment gérer les données manquantes ?

Oulad Ali Ayoub[1] Mabrouk Mohy[1]

[1] ML pour la santé

1er avril 2026

Résumé

Les données manquantes constituent l'un des défis les plus répandus et pourtant les plus souvent sous-estimés du Machine Learning clinique. Que l'on construise un modèle pronostique en oncologie, un prédicteur de réadmission ou une étude en vie réelle, des lacunes apparaissent inévitablement dans les données. La question n'est pas *« ai-je des données manquantes ? »* mais plutôt *« quelle est leur proportion, pourquoi manquent-elles, et quelle est la manière la plus rigoureuse de les traiter ? »*

On parcourt l'ensemble des méthodes d'imputation : des approches classiques (suppression, imputation par la moyenne, KNN, NIPALS, LOESS, SVD, MICE) aux approches génératives basées sur l'apprentissage profond (GAIN, VAE, VAEAC). Nous évaluons ces méthodes selon quatre critères RMSE, AUC, temps de calcul et stabilité des valeurs de Shapley sur deux datasets médicaux réels et un dataset simulé haute dimension ($n = 800$, $p = 120$).

Table des matières

1 Introduction : contexte et problématique **4**

2 Mécanismes de données manquantes **4**

2.1 Formalisation mathématique 5
2.2 MCAR Missing Completely At Random 5
2.3 MAR Missing At Random 5
2.4 MNAR Missing Not At Random 5
2.5 Que faire face à MNAR ? 6

3 Structures spatiales des données manquantes **6**

3.1 Manque univarié 6
3.2 Manque monotone 6
3.3 Manque non monotone (arbitraire) 7

4 Méthodes d'imputation classiques **7**
4.1 Suppression des observations . . . 7
4.2 Complétions stationnaires . . . 7
4.3 Imputation par combinaison linéaire . . . 8
4.4 Méthode des plus proches voisins (KNN) . . . 8
4.5 Algorithme NIPALS . . . 9
4.6 Régression locale (LOESS) . . . 10
4.7 Décomposition en valeurs singulières (SVD) . . . 10
4.7.1 SVD directe (cas avec suffisamment de données complètes) . . . 11
4.7.2 SVD itérative (fort taux de manque) . . . 11
4.8 MICE — *Multivariate Imputation by Chained Equations* . . . 11

5 Méthodes génératives **12**
5.1 Fondements bayésiens . . . 12
5.2 GAIN *Generative Adversarial Imputation Network* . . . 13
5.3 VAE — *Variational Autoencoder* : imputation itérative . . . 13
5.4 Universal Marginalizer imputation en un seul passage . . . 14
5.5 VAEAC *Variational Autoencoder with Arbitrary Conditioning* . . . 15
5.6 Comparaison synthétique VAE vs VAEAC . . . 15

6 Protocole expérimental **16**
6.1 Données . . . 16
6.2 Simulation des données manquantes . . . 16
6.3 Méthodes comparées et métriques . . . 16

7 Analyse des résultats **16**
7.1 Qualité d'imputation RMSE . . . 16
7.2 Performance prédictive AUC . . . 17
7.3 Scalabilité temps de calcul . . . 17
7.4 Compromis global qualité/temps . . . 17

8 Analyse des valeurs de Shapley **18**
8.1 Rappel valeurs de Shapley . . . 18
8.2 Résultats sous MCAR . . . 18
8.3 Résultats sous MNAR . . . 19

9 Conséquences d'une mauvaise imputation **19**

10 Guide décisionnel pratique **20**
10.1 Quelle méthode choisir ? . . . 20
10.2 Bonnes pratiques incontournables . . . 20

11 Conclusion **21**

1. Introduction : contexte et problématique

Pourquoi les données manquantes constituent-elles un enjeu aussi critique en recherche clinique ? Quatre raisons essentielles permettent d'en mesurer l'importance.

Un facteur de risque méthodologique. Sans traitement approprié, les données manquantes introduisent un biais systématique dans les estimations, les intervalles de confiance, et in fine les décisions cliniques. Un médicament peut paraître plus sûr qu'il ne l'est, ou un biomarqueur plus prédictif qu'il ne l'est réellement, simplement parce que les données relatives aux cas les plus graves sont absentes.

Une vulnérabilité accrue des études en vie réelle. Contrairement aux essais contrôlés randomisés, les études observationnelles reflètent le désordre de la réalité clinique : des patients manquent des visites de suivi, des examens biologiques ne sont pas prescrits de façon systématique, et les dossiers médicaux électroniques ont été conçus pour la facturation, pas pour la recherche. La seule perte au suivi peut invalider une étude pourtant bien conçue.

Une propagation à tout le pipeline analytique. L'imputation est au cœur du pipeline d'analyse. Une mauvaise décision d'imputation se propage dans chaque étape en aval : ingénierie des variables, entraînement du modèle, calibration et interprétation.

L'apport du Machine Learning. Le Machine Learning offre aujourd'hui des outils d'imputation bien plus riches que ceux dont disposait la communauté statistique il y a vingt ans. Cette revue évalue ces outils en les comparant tête-à-tête, avec une attention particulière portée à leurs effets sur l'explicabilité des modèles via les valeurs de Shapley.

Message fondamental. Il convient de considérer les données manquantes comme un problème de validité scientifique, et non comme un simple inconvénient de nettoyage de données. Le biais introduit lors de l'imputation est silencieux : il ne génère pas de message d'erreur.

2. Mécanismes de données manquantes

Avant de choisir une méthode d'imputation, il est indispensable de diagnostiquer *pourquoi* les données manquent. La taxonomie de Rubin [1] distingue trois mécanismes, dont le remède approprié dépend entièrement.

2.1. Formalisation mathématique

Soit $Y = (y_{ij}) \in \mathbb{R}^{n\times p}$ la matrice des données et $M = (m_{ij}) \in \{0,1\}^{n\times p}$ la matrice indicatrice de manque binaire, où $m_{ij} = 1$ signifie que la valeur est manquante. Les trois mécanismes se distinguent par la structure de la probabilité $p(M \mid Y)$.

2.2. MCAR Missing Completely At Random

$$p(M \mid Y) = p(M)$$

La probabilité qu'une valeur manque n'a aucun lien avec les autres variables, observées ou non. L'exemple canonique est celui d'un technicien de laboratoire qui fait accidentellement tomber un lot de tubes de sang : le manque est purement aléatoire. Sous MCAR, une analyse sur cas complets reste sans biais, même si elle perd en efficacité statistique.

2.3. MAR Missing At Random

$$p(M \mid Y) = p(M \mid Y_{\text{obs}})$$

La probabilité de manque dépend des données observées, mais pas de la valeur manquante elle-même. Par exemple, les patients plus âgés peuvent être moins susceptibles d'avoir un certain biomarqueur mesuré ; l'âge étant observé, on peut conditionner dessus. Malgré son nom, MAR n'est pas vraiment aléatoire : cela signifie que le mécanisme est *ignorable* une fois que l'on conditionne sur les covariables observées. La grande majorité des méthodes d'imputation modernes sont conçues pour ce cas.

2.4. MNAR Missing Not At Random

$$p(M \mid Y) \text{ dépend de } Y_{\text{mis}}$$

La probabilité de manque dépend de la valeur elle-même qui est manquante. C'est le cas le plus difficile et, malheureusement, souvent le plus réaliste en contexte clinique. L'exemple frappant est celui des patients présentant les pires résultats—la maladie la plus sévère, les effets secondaires les plus invalidants—qui sont précisément ceux ayant le plus de chances d'abandonner une étude ou de refuser le suivi. Les données manquantes contiennent alors de l'information sur exactement ce que l'on cherche le plus à mesurer. Ce mécanisme est *non ignorable* : quelle que soit la sophistication de la méthode d'imputation, il est impossible de s'en affranchir entièrement sans hypothèses externes fortes ou données supplémentaires.

2.5. Que faire face à MNAR ?

Il n'existe pas de solution miracle pour MNAR. Les approches envisageables sont les suivantes.

— **Analyse de sensibilité :** ajuster l'analyse sous plusieurs hypothèses plausibles sur les valeurs manquantes (ex. supposer que les patients manquants avaient des résultats 20 %, 50 % moins bons) et rapporter comment les conclusions changent.
— **Modèles de sélection et *pattern-mixture models* :** modéliser explicitement le mécanisme de manque via des copules ou des distributions en mélange. Ces approches nécessitent des hypothèses paramétriques fortes, mais peuvent au moins quantifier l'incertitude.
— **Variables instrumentales :** si l'on dispose d'une variable externe qui prédit le manque mais pas directement le résultat, l'utiliser pour identifier le biais.
— **Documentation honnête :** reconnaître MNAR et ne pas prétendre que l'imputation a résolu le problème. Tout article prétendant avoir entièrement corrigé des données MNAR sans hypothèses externes doit être lu avec un scepticisme extrême.

3. Structures spatiales des données manquantes

Au-delà du mécanisme (pourquoi les données manquent), il importe de s'intéresser à la structure spatiale du manque (où il se situe), qui influence directement les algorithmes pouvant traiter les données efficacement.

3.1. Manque univarié

Les valeurs manquantes concernent une seule variable Y_k. C'est le cas le plus simple : n'importe quelle méthode convient.

3.2. Manque monotone

Si la variable Y_j est manquante pour un individu, alors toutes les variables suivantes $\{Y_k\}_{k>j}$ le sont également :

$$m_{ij} = 1 \implies m_{ik} = 1, \quad \forall\, k > j.$$

Ce schéma est fréquent dans les études temporelles avec abandon : une fois qu'un patient quitte l'étude, toutes les mesures futures sont manquantes.

3.3. Manque non monotone (arbitraire)

Les valeurs manquantes n'ont pas de structure d'ordre particulière dans M. C'est le cas général, le plus courant en pratique clinique, et le plus difficile computationnellement. Des méthodes comme MICE et VAEAC sont spécifiquement conçues pour le gérer.

4. Méthodes d'imputation classiques

Le spectre des méthodes d'imputation va du simple-et-rapide au complexe-et-puissant. Cette section présente les approches classiques, dont certaines restent des références compétitives malgré leur ancienneté.

4.1. Suppression des observations

La suppression est l'instrument le plus brut. Rapide et interprétable, elle n'est valide que sous MCAR et avec un faible taux de manque.

Suppression par liste (*listwise deletion*). On supprime les individus présentant au moins une valeur manquante. La perte d'information peut être dramatique : avec 20 variables présentant chacune 5 % de manque, on peut perdre 60 à 70 % des lignes sous MCAR. Plus dangereusement, si les données ne sont pas MCAR—ce qui est le cas usuel en pratique clinique—la suppression par liste biaise l'échantillon : on analyse uniquement les patients « faciles », ceux qui sont venus à tous les rendez-vous et ont un suivi complet, rarement un groupe représentatif.

Suppression par paire (*pairwise deletion*). Chaque analyse utilise les observations disponibles pour les variables concernées. Cette approche préserve plus de données, mais crée un problème d'incohérence : différentes parties de l'analyse sont estimées sur des sous-échantillons différents, ce qui peut produire des résultats logiquement incohérents.

La règle empirique de Little & Rubin [2] est instructive : la suppression n'est acceptable que si moins de 5 % des valeurs manquent *et* si l'on a de fortes preuves que les données sont MCAR. En pratique, c'est une rareté dans les données de santé.

4.2. Complétions stationnaires

CMCF — *Most Common Attribute Value Fitting.* Pour les variables catégorielles, on impute avec le mode (la catégorie la plus fréquente) ; pour les variables continues, cela correspond à l'imputation par la moyenne ou la médiane.

LOCF — *Last Observation Carried Forward.* Une norme dans l'analyse des essais cliniques longitudinaux. Si la mesure d'un patient est manquante au temps t, on reporte la dernière mesure connue au temps t' :

$$(y_{ij})_{\text{mis}} = y_{ij'}, \quad m_{ij'} = 0,\ j' < j.$$

LOCF suppose que l'état du patient est resté stable depuis sa dernière mesure une hypothèse presque jamais vraie pour des patients malades ayant abandonné l'essai précisément parce que leur état s'est aggravé. LOCF sous-estime systématiquement la progression de la maladie et biaise fréquemment les estimations d'effet en faveur du bras expérimental. Malgré ces limitations, LOCF persiste dans les soumissions réglementaires pour des raisons historiques ; les autorités modernes exigent de plus en plus des analyses de sensibilité complémentaires.

4.3. Imputation par combinaison linéaire

L'imputation par la moyenne et la médiane sont les approches paramétriques les plus simples :

$$(y_{ij})_{\text{mis}} = \bar{Y}_j \qquad \text{ou} \qquad (y_{ij})_{\text{mis}} = \tilde{Y}_j.$$

Lorsqu'on impute chaque valeur manquante par la moyenne de la colonne, on affirme implicitement que ce patient est parfaitement dans la moyenne pour cette variable. Pour un faible taux de manque sous MCAR, c'est globalement inoffensif. Au fur et à mesure que la proportion de manque augmente, on concentre artificiellement la distribution des données au centre, réduisant la variance et distordant les corrélations entre variables. L'imputation par la médiane est plus robuste aux valeurs aberrantes, mais souffre du même défaut fondamental : elle ignore les dépendances entre variables.

Ces méthodes sont acceptables comme référence de comparaison rapide ou pour une analyse exploratoire, mais elles ne doivent jamais servir d'imputation finale pour des statistiques inférentielles.

4.4. Méthode des plus proches voisins (KNN)

L'imputation KNN est conceptuellement élégante : pour remplir une valeur manquante pour l'individu i, on trouve les k individus les plus similaires sur toutes les variables observées, puis on impute par la moyenne de leurs valeurs pour la variable manquante.

Algorithm 1 Imputation KNN

1: **Entrée :** entier k avec $1 \leq k \leq n$.
2: Calculer les distances $d(Y_i, Y_{i'})$ pour $i = 1, \ldots, n$ (sur variables observées).
3: Retenir les k observations $Y_{(i_1)}, \ldots, Y_{(i_k)}$ de plus petite distance.
4: Imputer :

$$(y_{ij})_{\text{mis}} = \frac{1}{k} \sum_{r=1}^{k} Y_{(i_r)}$$

KNN est meilleur que l'imputation globale par la moyenne parce qu'il est localement adaptatif : au lieu de dire « ce patient est dans la moyenne », on dit « ce patient ressemble à ces individus spécifiques, donc sa valeur est probablement similaire à la leur ». Cela préserve la structure locale des données.

Compromis.

— Le choix de k est important : un k petit est sensible au bruit ; un k grand ressemble à une imputation par la moyenne globale. La validation croisée sur des données artificiellement masquées peut aider.
— Le coût de calcul est $O(n^2)$—faisable pour quelques milliers de patients, coûteux à grande échelle.
— La qualité dépend directement de la métrique de distance. La distance euclidienne convient pour des variables continues homogènes.

4.5. Algorithme NIPALS

NIPALS (*Nonlinear Iterative Partial Least Squares*) est un algorithme de réduction dimensionnelle initialement conçu pour la régression PLS, qui traite naturellement les données manquantes comme sous-produit de sa structure itérative.

L'idée centrale est la suivante : si les données ont une structure de faible rang—c'est-à-dire que les variables sont très corrélées et peuvent être approximées par un petit nombre de facteurs latents—on peut reconstruire les valeurs manquantes à partir de ces facteurs.

Algorithm 2 NIPALS

1: Initialiser $Y_0 = Y$.
2: **for** $h = 1, \ldots, q$ **do** Fixer $\xi_h = Y^1_{h-1}$.
3: Itérer jusqu'à convergence de u_h : mise à jour de u_h, normalisation, mise à jour de ξ_h.
4: Mettre à jour la matrice résiduelle : $Y_h = Y_{h-1} - \xi_h u_h^\top$.
5: **end for**
6: Imputer :

$$(\hat{y}_{ij})_{\text{mis}} = \sum_{h=1}^{q} \hat{\xi}_h(i)\, \hat{u}_h(j)$$

NIPALS excelle quand les données sont hautement multicolinéaires (données génomiques, panels omiques, ou jeux de données cliniques où de nombreuses variables sont des proxies du même processus biologique). Il est moins adapté quand la structure est véritablement non linéaire, ou quand le taux de manque est très élevé, car la procédure itérative peut devenir instable.

4.6. Régression locale (LOESS)

LOESS (*Locally Estimated Scatterplot Smoothing*) reprend l'idée de KNN et la combine avec une régression linéaire locale. Pour chaque valeur manquante, on trouve les k plus proches voisins dans l'espace des variables observées, puis on ajuste une régression locale par moindres carrés pondérés.

Algorithm 3 LOESS pour imputation

1: Obtenir les k plus proches voisins $Y_{(i_1)}, \ldots, Y_{(i_k)}$.
2: Construire $A \in \mathbb{R}^{k\times(n-q)}$, $B \in \mathbb{R}^{k\times q}$, $w \in \mathbb{R}^{(n-q)\times 1}$.
3: Résoudre :

$$\min_{x\in\mathbb{R}^k} \|Ax - w\|^2$$

4: Imputer :

$$u = Bx = B\,(A^+)\,w$$

LOESS capture la structure non linéaire locale plus précisément que le KNN simple : il ne fait pas que moyenner, il modélise. Il est véritablement non paramétrique au sens où il fait des hypothèses minimales sur la forme fonctionnelle globale. La limitation principale réside dans le coût de calcul similaire à KNN, augmenté de la charge d'ajuster une régression locale pour chaque entrée manquante.

4.7. Décomposition en valeurs singulières (SVD)

L'imputation basée sur la SVD exploite le fait mathématique que les matrices de données réelles sont souvent approximativement de faible rang elles peuvent être bien approximées par un produit de quelques vecteurs.

4.7.1. SVD directe (cas avec suffisamment de données complètes)

Algorithm 4 SVD directe

1: Séparer les données en Y_c (observations complètes) et Y_m (observations incomplètes).
2: Calculer la SVD tronquée de rang J sur Y_c :

$$\hat{Y}_c^{(J)} = U_J D_J V_J^\top$$

3: Estimer les coefficients par régression sur les composantes observées :

$$\hat{\beta} = (V_J^\top V_J)^{-1} V_J^\top Y_i$$

4: Prédire les valeurs manquantes :

$$\hat{Y}_i = V_J^{(c)} \hat{\beta}$$

4.7.2. SVD itérative (fort taux de manque)

Quand le taux de manque est élevé, il est impossible de calculer de façon fiable une SVD sur un sous-ensemble d'observations complètes. L'algorithme de complétion itérative répond à ce problème.

Algorithm 5 SVD itérative

1: Initialiser $Y^{(0)}$ en remplaçant les valeurs manquantes par la moyenne de colonne.
2: **repeat**
3: Résoudre :

$$\min_{U_J, V_J, D_J} \left\| Y - m - U_J D_J V_J^\top \right\|^2$$

(en ignorant les valeurs manquantes), puis construire $Y^{(t+1)}$.
4: **until** $\dfrac{\|Y^{(t+1)} - Y^{(t)}\|}{\|Y^{(t)}\|} < \varepsilon$, typiquement $\varepsilon = 10^{-6}$.

La SVD itérative est une forme de minimisation alternée : on optimise conjointement la structure de faible rang de la matrice et les valeurs imputées. Même à 50–60 % de manque, l'algorithme peut récupérer le signal si les données sont suffisamment de faible rang. Si les données ne le sont pas, l'imputation par SVD se dégrade gracieusement vers l'imputation par la moyenne.

4.8. MICE — *Multivariate Imputation by Chained Equations*

MICE est sans doute la référence absolue pour l'imputation en recherche clinique. L'idée centrale est la suivante : au lieu d'imputer toutes les valeurs manquantes d'un coup avec un modèle global, on impute une variable à la fois à l'aide d'un modèle de régression qui conditionne sur toutes les autres variables, en enchaînant ces imputations univariées de façon tournante.

Algorithm 6 MICE

1: Initialiser toutes les valeurs manquantes par la moyenne de colonne.
2: **for** chaque variable Y_j avec valeurs manquantes **do**
3: Régresser Y_j sur Y_{-j} en utilisant les valeurs imputées courantes.
4: Tirer des imputations de la distribution prédictive (avec bruit aléatoire).
5: **end for**
6: Répéter ce cycle 5 à 10 fois.
7: Produire $m \geq 5$ jeux de données imputés en exécutant la procédure m fois avec des graines aléatoires différentes.
8: Analyser chaque jeu séparément et combiner les résultats via les règles de Rubin.

L'étape de production de multiples jeux de données imputés est la plus importante et la plus souvent omise en pratique. En produisant de multiples imputations, MICE quantifie l'incertitude due aux données manquantes. Les règles de Rubin permettent ensuite de combiner correctement les estimations et d'en gonfler les erreurs standard pour tenir compte de l'incertitude d'imputation.

MICE est meilleur que l'imputation par la moyenne car il préserve la structure des corrélations entre variables. Il est meilleur que KNN car il gère naturellement les types de données mixtes (différents modèles de régression pour les variables continues, binaires et catégorielles) et produit des estimations d'incertitude appropriées.

5. Méthodes génératives

Nous passons ici du paradigme statistique classique au paradigme de la modélisation générative. La différence philosophique est subtile mais importante : les méthodes classiques modélisent les distributions conditionnelles $p(Y_j \mid Y_{-j})$ une variable à la fois, tandis que les méthodes génératives visent à modéliser la distribution jointe complète $p(Y)$.

5.1. Fondements bayésiens

Le cadre bayésien est élégant. La distribution a posteriori des paramètres du modèle sachant les données observées est :

$$p(\theta \mid Y_{\text{obs}}) \propto p(\theta)\, f(Y_{\text{obs}}; \theta).$$

L'algorithme de *data augmentation* de Tanner & Wong [3] permet d'échantillonner de cette distribution a posteriori de façon itérative en alternant deux étapes :

— **Étape I (Imputation) :** $Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta^{(t)})$.
— **Étape P (Postérieur) :** $\theta^{(t+1)} \sim p(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

Amelia II [4], de Honaker et al. (2011), est une implémentation pratique de cette approche sous hypothèse gaussienne multivariée. Elle suppose $Y \sim \mathcal{N}_k(\mu, \Sigma)$ et un mécanisme MAR, et produit des imputations multiples théoriquement optimales pour des données cliniques continues approximativement normales.

Les méthodes génératives modernes étendent ce cadre à des familles de distributions bien plus expressives, remplaçant la gaussienne par un réseau de neurones profond.

5.2. GAIN *Generative Adversarial Imputation Network*

GAIN [5], introduit par Yoon et al. à ICML 2018, adapte le cadre des GAN au problème de l'imputation. Deux réseaux de neurones sont entraînés de façon adversariale.

Le **Générateur** G prend l'entrée bruitée $\bar{X}$ (où les valeurs manquantes sont mises à zéro) et le masque M, et produit une version complétée $\hat{X} = G(\bar{X}, M)$. Le **Discriminateur** D prend les données complétées $\hat{X}$ et tente de deviner quelles entrées étaient originellement observées et lesquelles ont été imputées, variable par variable.

Algorithm 7 GAIN

1: Construire $\bar{X}$ en remplaçant les valeurs manquantes selon le masque M.
2: Générateur : $\hat{X} = G(\bar{X}, M)$.
3: Discriminateur : $\hat{M} = D(\hat{X})$.
4: Optimiser D par entropie croisée binaire sur M.
5: Optimiser G :

$$\min_{\theta_G} \sum_{i \in B} L_G(M_i, \hat{M}_i) + \alpha \, L_{\text{rec}}(\bar{X}_i, \hat{X}_i, M_i)$$

6: Itérer les étapes 2–5 jusqu'à convergence.

Un imputeur entraîné de façon adversariale apprend à produire des valeurs issues de la distribution marginale réaliste de chaque variable, pas seulement la moyenne conditionnelle. Les valeurs imputées ressemblent à de vraies données, pas à des moyennes lissées.

5.3. VAE — *Variational Autoencoder* : imputation itérative

Le VAE adopte une approche générative différente : il utilise l'inférence variationnelle pour apprendre une représentation latente compressée des données. L'imputation s'effectue en raffinant itérativement les valeurs imputées à l'aide du décodeur.

Algorithm 8 VAE Imputation Itérative

Entrées : $\bar{X}_{\text{train}}, M_{\text{train}}, \bar{X}_{\text{test}}, M_{\text{test}}, i_{\max}, e_{\min}$.
Entraîner : $\text{VAE} = \text{train}(\bar{X}_{\text{train}}, M_{\text{train}})$.
Initialiser $\epsilon \sim \mathcal{N}(0,1)$ et :

$$\hat{X}_{\text{test}} = \bar{X}_{\text{test}} \odot M_{\text{test}} + \epsilon \odot (1 - M_{\text{test}}).$$

Initialiser $i = 0$.
repeat

$$\hat{X}_{\text{test}} = \text{VAE}(\hat{X}_{\text{test}}), \quad e = L_{\text{rec}}(\bar{X}_{\text{test}}, \hat{X}_{\text{test}}, M_{\text{test}})$$
$$\hat{X}_{\text{test}} = \bar{X}_{\text{test}} \odot M_{\text{test}} + \hat{X}_{\text{test}} \odot (1 - M_{\text{test}}), \quad i \leftarrow i + 1$$

until $e \leq e_{\min}$ ou $i \geq i_{\max}$.

On peut voir ce processus comme une forme de descente de gradient projetée dans l'espace des données : on projette répétitivement l'observation incomplète sur la variété apprise des données, en comblant les dimensions manquantes avec ce que la variété prédit. Le terme de divergence KL dans la ELBO du VAE standard prévient le surapprentissage.

Limitation. Le prior gaussien du VAE standard peut ne pas correspondre à la vraie distribution des données cliniques, qui présentent souvent des distributions multimodales ou à queues lourdes.

5.4. Universal Marginalizer imputation en un seul passage

L'innovation de l'Universal Marginalizer est la suivante : au lieu d'itérer au moment du test, on entraîne le réseau à gérer directement des schémas de manque arbitraires.

Algorithm 9 Universal Marginalizer Entraînement

Entrées : jeu complet $\mathcal{D}$, réseau f_θ, nombre d'itérations N_{iter}.
for iter $= 1, \ldots, N_{\text{iter}}$ **do**
    Échantillonner un batch $X \sim \mathcal{D}$.
    Échantillonner $p \sim \mathcal{U}(0,1)$ et un masque $M \sim \text{Bernoulli}(p)$.
    Créer $X_{\text{in}} = X \odot M$, concaténer $\tilde{X} = [X_{\text{in}}, M]$.
    Prédire $\hat{X} = f_\theta(\tilde{X})$.
    Calculer :

$$\mathcal{L} = \frac{1}{\sum(1 - M)} \sum (1 - M) \cdot \text{Loss}(X, \hat{X})$$

    Mettre à jour θ par descente de gradient sur $\mathcal{L}$.
end for

En s'entraînant sur tous les schémas de masquage possibles, le réseau apprend une distribution conditionnelle universelle $p(X_{\text{mis}} \mid X_{\text{obs}}, M)$. La concaténation du masque M est cruciale : elle indique au réseau exactement quelles entrées manquent. L'inférence en un seul passage est extrêmement rapide au moment du déploiement.

5.5. VAEAC *Variational Autoencoder with Arbitrary Conditioning*

VAEAC [6] combine la puissance des autoencodeurs variationnels avec l'idée de conditionnement arbitraire. L'innovation architecturale clé réside dans la présence de **deux encodeurs**.

Encodeur Complet Traite toutes les variables (X_{obs} et X_{mis}) et produit une distribution a posteriori sur la variable latente z :

$$p_\phi(z \mid x, \bar{S}) = \mathcal{N}(z \mid \mu_\phi, \text{diag}[\sigma_\phi^2]).$$

Cet encodeur représente « ce que l'on sait du patient connaissant toutes les informations ».

Encodeur Masqué Traite uniquement les variables observées X_S, conditionnellement au masque :

$$p_\psi(z \mid x_S, \bar{S}) = \mathcal{N}(z \mid \mu_\psi, \text{diag}[\sigma_\psi^2]).$$

Cet encodeur représente « ce qu'on peut inférer à partir des seules données observées ».

L'objectif d'entraînement minimise l'ELBO, comportant deux termes : (i) la divergence KL entre p_ϕ et p_ψ (forçant l'Encodeur Masqué à approcher l'Encodeur Complet), et (ii) la log-vraisemblance espérée (forçant le Décodeur à reconstruire les valeurs manquantes).

À l'inférence, seuls l'Encodeur Masqué et le Décodeur sont utilisés. L'imputation multiple est triviale : il suffit d'échantillonner $z^{(k)}$ depuis p_ψ à k reprises. Les skip connections entre l'Encodeur Masqué et le Décodeur sont cruciales pour la stabilité de l'entraînement, en permettant au gradient de circuler directement depuis la perte de reconstruction.

5.6. Comparaison synthétique VAE vs VAEAC

Table 1 – Comparaison VAE itératif et VAEAC.

Dimension	**VAE itératif**	**VAEAC**
Flexibilité du masque	Schéma fixe ou itération	Masques arbitraires natifs
$p(X_{\text{mis}} \mid X_{\text{obs}})$	Approx. itérative	Distribution paramétrique directe
Scalabilité	Bonne	Limitée (malédiction dimensionnelle)
Inférence	Plusieurs passes	Un seul passage
Imputation multiple	Par échantillonnage	Par échantillonnage

6. Protocole expérimental

6.1. Données

Deux datasets médicaux réels ont été utilisés : **Breast Cancer** (classification) et **Diabetes** (régression), de taille petite à moyenne. Un troisième dataset simulé de haute dimension ($n = 800$, $p = 120$) a été construit pour mettre à l'épreuve la scalabilité des méthodes.

6.2. Simulation des données manquantes

Des schémas MCAR, MAR et MNAR ont été injectés artificiellement à des taux allant de 5 % à 70 % des valeurs. Cette large plage est intentionnelle : elle permet de comprendre comment chaque méthode se dégrade à mesure que le manque devient plus sévère.

6.3. Méthodes comparées et métriques

Neuf méthodes ont été comparées : Moyenne, Médiane, LOCF, KNN, MICE, SVD, GAIN, VAE, VAEAC. Les métriques d'évaluation sont les suivantes.

— **RMSE** (qualité d'imputation brute, avec masquage synthétique).
— **AUC** (performance prédictive en aval, sur une tâche de classification).
— **Temps de calcul** (scalabilité en grande dimension).
— **Divergence KL des valeurs de Shapley** par rapport à MICE comme référence.

MICE est utilisé comme référence pour la comparaison Shapley car il est largement considéré comme la méthode de référence sous MAR. La divergence KL entre la distribution des valeurs Shapley d'une méthode candidate et celle de MICE indique à quel point la méthode candidate distord l'importance des variables par rapport au benchmark de bonnes pratiques.

7. Analyse des résultats

7.1. Qualité d'imputation RMSE

Sous MCAR. MICE délivre le RMSE le plus faible sur les trois mécanismes : 0,7128 sous MCAR. KNN est systématiquement deuxième (0,87–0,90). Les méthodes simples (Moyenne, Médiane) se situent à 1,0–1,2, acceptables uniquement à faible taux de manque. Le résultat le plus frappant concerne LOCF, catastrophiquement mauvais sous MNAR avec un RMSE de 1,49 presque le double du RMSE de MICE. En termes pratiques : utiliser LOCF quand les données sont MNAR revient à produire des valeurs imputées presque aussi erronées qu'un bruit aléatoire en valeur absolue.

Méthodes génératives vs classiques. GAIN (1,21–1,24) et VAEAC (1,07–1,28) sous-performent leurs homologues classiques. Deux raisons à cela : d'une part, ces modèles sont conçus pour capturer la distribution marginale des données, pas nécessairement pour minimiser l'erreur de reconstruction point par point ; d'autre part, avec des tailles d'échantillon relativement petites, les réseaux de neurones manquent de données pour apprendre un bon modèle génératif. Le VAE est une agréable surprise à 0,87–0,96, compétitif avec KNN.

7.2. Performance prédictive AUC

Sous MCAR, MICE atteint le meilleur AUC (0,8324), suivi de près par KNN (0,8339) et Moyenne (0,8244). La leçon est que sous MCAR—le scénario le plus bénin—même les méthodes simples performent près du meilleur niveau. Sous MNAR, MICE maintient son avantage à 0,8452 ; KNN performe à 0,8345. LOCF s'effondre à 0,8083 et VAEAC à 0,8179 s'avère décevant compte tenu de sa sophistication architecturale. Le tableau d'ensemble confirme que MICE délivre systématiquement la meilleure performance prédictive sur les trois mécanismes. Les méthodes d'apprentissage profond ne surpassent pas systématiquement MICE sur des datasets de taille modérée, un résultat récurrent dans la littérature : les modèles complexes ne l'emportent que quand les données sont suffisantes pour justifier leur complexité.

7.3. Scalabilité temps de calcul

Les résultats sur le dataset haute dimension ($n = 800$, $p = 120$) révèlent des écarts considérables.

— **Moyenne :** quelques millisecondes, quel que soit le taux de manque. Référence de vitesse.

— **SVD :** environ 0,3–0,5 secondes, remarquablement stable. Le calcul SVD scale avec la taille de la matrice, pas avec le nombre d'entrées manquantes.

— **VAEAC :** environ 1–2 secondes, stable à l'inférence. Le coût d'entraînement est fixe ; seule l'étape d'inférence scale avec les observations.

— **MICE :** l'outlier spectaculaire, démarrant à $\sim$30 secondes même à faible taux de manque, et atteignant $\sim$50–60 secondes à 70 % de manque. MICE doit ajuster un modèle de régression séparé pour chaque variable avec des valeurs manquantes et itérer ; avec $p = 120$ variables et un fort manque, cela devient prohibitif.

7.4. Compromis global qualité/temps

L'agrégation sur tous les taux de manque et tous les datasets confirme la hiérarchie suivante. **MICE** est le gagnant clair sur les deux dimensions (AUC 0,915, RMSE $\approx$ 0,59). **KNN**

est un deuxième très convaincant (AUC 0,912, RMSE 0,79), avec un très bon rapport qualité/coût. Le **VAE** est comparable à KNN (AUC 0,910, RMSE 0,80) tout en apportant les avantages structurels d'une représentation apprise. GAIN et SVD se regroupent au milieu adéquats mais pas exceptionnels. **VAEAC** est une déception à cette échelle (AUC 0,905, RMSE 1,08) : la sophistication théorique ne se traduit pas par une supériorité pratique sur ces datasets cliniques relativement petits. **LOCF** est le grand perdant (AUC 0,888), à éviter dans tout contexte avec un manque structuré.

8. Analyse des valeurs de Shapley

8.1. Rappel valeurs de Shapley

La valeur de Shapley de la variable i pour la prédiction f_x est la contribution marginale moyenne de i sur tous les sous-ensembles possibles des autres variables :

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \Big[f_x(S \cup \{i\}) - f_x(S) \Big].$$

Les valeurs de Shapley constituent l'allocation équitable de la valeur de prédiction sur toutes les variables d'entrée, empruntée à la théorie des jeux coopératifs. Elles nécessitent de marginaliser sur les variables d'intégrer certaines variables en fixant d'autres. La façon de gérer les variables manquantes lors de cette marginalisation est profondément liée à la façon dont les données manquantes ont été traitées pendant l'entraînement. Une méthode d'imputation qui distord la distribution jointe des variables produira des marginalisations incorrectes, conduisant à des valeurs de Shapley biaisées. TreeSHAP, utilisé dans nos expériences, calcule des valeurs de Shapley exactes pour les modèles à base d'arbres en temps polynomial.

8.2. Résultats sous MCAR

Sous MCAR, les rankings de variables sont remarquablement cohérents entre les neuf méthodes. La variable `s5` est le prédicteur dominant (plage 0,25–0,31), suivie de `bmi` (0,13–0,17), `bp` (0,11–0,14) et `s3` (0,09–0,13). La cohérence est théoriquement attendue et empiriquement confirmée : quand les données sont MCAR, toute méthode d'imputation raisonnable récupère le vrai signal, car le schéma de manque ne contient aucune information sur le résultat. SVD se distingue légèrement en attribuant une importance relativement plus élevée à `bmi` reflet de la tendance de SVD à amplifier la structure de faible rang corrélée aux facteurs liés à l'IMC. LOCF présente une légère divergence sur la variable `sex`, ce qui a du sens : LOCF n'est pas pertinent pour les variables invariantes dans le temps.

Conclusion sous MCAR. Toute méthode d'imputation raisonnable produit des explica-

tions SHAP stables et cohérentes.

8.3. Résultats sous MNAR

Sous MNAR, le tableau change substantiellement. `bmi` devient le prédicteur principal (0,21–0,26), tandis que `s5` dominant sous MCAR passe en deuxième position (0,18–0,24). Ce n'est pas un simple artefact statistique : MNAR distord la distribution des covariables observées, rendant certaines variables plus ou moins prédictives qu'elles ne le sont réellement.

La divergence entre méthodes est bien plus large. Pour la variable `s4`, la plage passe de 5–10 % d'importance sous MCAR à 0,074 (MICE)–0,156 (VAEAC) sous MNAR—plus du double. C'est la différence entre ignorer une variable et la signaler comme cliniquement importante. GAIN présente un profil distinctif : il alloue une importance élevée à `s3` (0,150) par rapport aux autres méthodes (0,103–0,143), reflet de la sensibilité de GAIN aux décalages de distribution sous MNAR.

Implication clinique directe. Si l'on construit un modèle destiné à soutenir les décisions cliniques—par exemple, « sur quelle valeur biologique le clinicien doit-il se concentrer ? » le choix de la méthode d'imputation peut littéralement changer la recommandation. Ce n'est pas une abstraction statistique.

9. Conséquences d'une mauvaise imputation

Les conséquences d'une imputation incorrecte peuvent être appréhendées à plusieurs niveaux.

Baisse de précision prédictive. Une imputation inadéquate dégrade les variables d'entrée. Le modèle apprend à partir de données corrompues et généralise mal. Dans un modèle de prédiction clinique, cela signifie une moins bonne stratification du risque : des patients incorrectement classifiés comme à faible risque pourraient ne pas recevoir les interventions nécessaires.

Introduction de biais. Des hypothèses d'imputation incorrectes distordent la distribution empirique des variables. L'imputation par la moyenne contracte les distributions vers le centre, gonflant artificiellement la précision. LOCF sous MNAR suppose de façon optimiste une santé stable dans une population de patients qui se détériore. Les deux produisent des estimations de coefficients biaisées et des tailles d'effet trompeuses.

Dégradation de l'explicabilité. Comme l'a montré l'analyse Shapley ci-dessus, une mauvaise imputation rend les explications du modèle instables et potentiellement trompeuses. Une variable apparaissant importante peut simplement refléter des artefacts

d'imputation plutôt qu'un signal biologique réel.

Le cas spécifique du VAEAC à prior gaussien. Si VAEAC utilise un prior gaussien dans son espace latent, et que les données cliniques ont des distributions multimodales ou à queues lourdes, l'hypothèse gaussienne biaise les valeurs imputées vers le mode du prior. Lorsque ces imputations biaisées sont utilisées pour calculer les valeurs de Shapley, la marginalisation sur la distribution conditionnelle « imputée » est incorrecte, produisant des attributions systématiquement biaisées pour toute variable corrélée avec la variable imputée.

Conclusion. L'imputation des données manquantes n'est pas un détail de prétraitement que l'on peut régler une fois pour toutes. C'est un choix de modélisation avec des conséquences statistiques, cliniques et éthiques. Il doit être documenté, évalué empiriquement et analysé en sensibilité.

10. Guide décisionnel pratique

Nous traduisons ici l'ensemble des résultats en un guide opérationnel.

10.1. Quelle méthode choisir ?

Moins de 5 % de manque ET MCAR. La suppression par liste est acceptable. Simple, rapide, sans biais dans ces conditions.

Manque modéré (5–30 %) ET MAR, données tabulaires. Premier choix : MICE — robuste, imputation multiple appropriée, gestion des types mixtes. **Deuxième choix : KNN** — si la rapidité et l'interprétabilité sont prioritaires, ou si MICE est trop coûteux.

Manque élevé ($> 30\,\%$) ET structure complexe. VAEAC si les variables sont très interdépendantes et que $n > 500$; GAIN pour les types mixtes (continu + binaire + catégoriel) ; SVD comme alternative rapide et fiable.

MNAR suspecté. Aucune méthode d'imputation ne résout MNAR. Documenter le mécanisme explicitement. Conduire des analyses de sensibilité. Rapporter l'étendue des résultats.

Besoin d'explicabilité (SHAP, importance des variables). Comparer toujours les valeurs de Shapley sur au moins deux méthodes d'imputation. Préférer l'imputation multiple. Signaler les variables dont l'importance est sensible au choix d'imputation.

10.2. Bonnes pratiques incontournables

1. **Analyser le mécanisme en premier.** Visualiser la matrice de manque, tester MCAR avec le test de Little, chercher des schémas par sous-groupe. Documenter la conclusion

sur le mécanisme probable.

2. **Ne jamais utiliser une seule méthode isolément.** Tester au moins deux approches. Si les conclusions divergent significativement, un problème de sensibilité doit être signalé.
3. **Évaluer empiriquement.** Si possible, construire une expérience de masquage synthétique sur des portions complètes des données. RMSE et MAE sur les valeurs masquées donnent une validation avec vérité terrain.
4. **Utiliser l'imputation multiple.** Produire $m \geq 5$ datasets imputés. Combiner estimations et erreurs standard via les règles de Rubin. L'imputation simple surestime la précision.
5. **Rapporter la sensibilité à l'imputation.** La section Résultats doit inclure un tableau ou une figure montrant comment les estimations clés changent selon les méthodes d'imputation.
6. **Vérifier la stabilité de l'explicabilité.** Si des valeurs SHAP ou des rankings d'importance sont rapportés, vérifier leur cohérence entre méthodes d'imputation. Des rankings instables signalent un artefact d'imputation plutôt qu'un comportement genuinement appris.
7. **Documenter dans la section Méthodes.** Préciser : quel mécanisme a été supposé, quelle méthode a été utilisée, combien d'imputations, quelles variables ont été incluses dans le modèle d'imputation.

11. Conclusion

Plusieurs messages essentiels ressortent de cette revue.

Les données manquantes ne sont jamais un détail. Elles constituent une menace pour la validité, une source de biais et—via leur interaction avec les méthodes d'explicabilité—une source potentielle de recommandations cliniques trompeuses. Elles méritent d'être traitées avec la rigueur qu'impose leur impact réel.

Le mécanisme détermine le remède. MCAR est bénin, MAR est gérable, MNAR est dangereux. Diagnostiquer toujours avant de traiter.

Les méthodes classiques restent compétitives. MICE, KNN et SVD ne sont pas obsolètes. Sur des datasets cliniques réalistes, ils égalent ou surpassent régulièrement les approches d'apprentissage profond. Il ne faut pas utiliser un GAN quand MICE suffit.

Les méthodes génératives trouvent leur place dans les scénarios complexes. Quand les variables sont très interdépendantes, quand le manque est très élevé, ou quand les données d'entraînement sont suffisamment nombreuses, VAEAC et VAE offrent de véritables avantages.

Toujours évaluer RMSE, AUC et stabilité Shapley ensemble. La qualité d'imputation et la qualité de la tâche en aval peuvent diverger. Les deux dimensions doivent être évaluées.

L'imputation multiple n'est pas optionnelle pour les statistiques inférentielles. Un seul dataset imputé produit des résultats trop confiants. Les règles de Rubin existent pour une raison.

L'insight fondamental de Rubin en 1976 reste inchangé : la validité d'une analyse dépend de façon critique du diagnostic correct et du traitement rigoureux du mécanisme de manque. Aucun algorithme, aussi sophistiqué soit-il, ne peut compenser un mécanisme mal diagnostiqué.

Références

[1] Rubin, D.B. (1976). *Inference and missing data.* Biometrika, 63(3), 581–592.

[2] Little, R.J.A., & Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

[3] Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

[4] Honaker, J., King, G., & Blackwell, M. (2011). Amelia II : A program for missing data. *Journal of Statistical Software*, 45(7), 1–47.

[5] Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN : Missing data imputation using generative adversarial nets. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, PMLR 80, 5689–5698.

[6] Olsen, P.A., Langseth, H., Lindström, O., & Sjøberg, O. (2023). VAEAC : Variational Autoencoder with Arbitrary Conditioning for imputation and explainability. *arXiv :2302.00887.*

[7] Lundberg, S., & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

ANNEXE

Annexe

Résultats complémentaires

TABLE 2 – Comparaison de la RMSE de reconstruction des valeurs de Shapley (Moyenne ± Std Dev) sur différents datasets.

Dataset	Mean	Zero	KNN	MissForest	VAEAC
ABALONE	$\mathbf{0.7024 \pm 0.0618}$	0.7103 ± 0.0333	1.0832 ± 0.1390	1.1265 ± 0.1168	0.7463 ± 0.0397
CALIFORNIA	0.9695 ± 0.0297	$\mathbf{0.8882 \pm 0.0448}$	0.9491 ± 0.0152	0.9113 ± 0.0497	0.9407 ± 0.0035
DIABETES	0.8751 ± 0.0519	0.8833 ± 0.0444	0.8594 ± 0.0310	$\mathbf{0.8445 \pm 0.0159}$	0.8902 ± 0.0131
IID	1.0035 ± 0.0082	1.0093 ± 0.0051	1.0087 ± 0.0196	1.0036 ± 0.0034	$\mathbf{0.9948 \pm 0.0037}$
GMM	0.8810 ± 0.0322	0.8471 ± 0.0138	0.8493 ± 0.0342	$\mathbf{0.8312 \pm 0.0055}$	0.8502 ± 0.0120
COPULA	$\mathbf{0.6230 \pm 0.0477}$	0.7388 ± 0.0231	0.6482 ± 0.0320	0.6581 ± 0.0308	0.6922 ± 0.0751

TABLE 3 – Détail des performances (RMSE) selon la dimension du masque (k).

Dataset	k	Mean	MissForest	VAEAC
COPULA	1	**0.525**	0.748	0.636
	2	0.706	0.796	**0.699**
	4	**0.623**	0.684	0.625
	8	**0.652**	0.680	0.711
GMM	1	**0.653**	0.677	0.666
	2	**0.649**	0.697	0.662
	4	**0.646**	0.689	0.680
	8	0.721	**0.675**	0.727
IID	1	1.154	1.120	**1.106**
	2	1.119	1.148	**1.117**
	4	1.118	**1.097**	1.114
	8	1.091	**1.082**	1.086

Schémas complémentaires

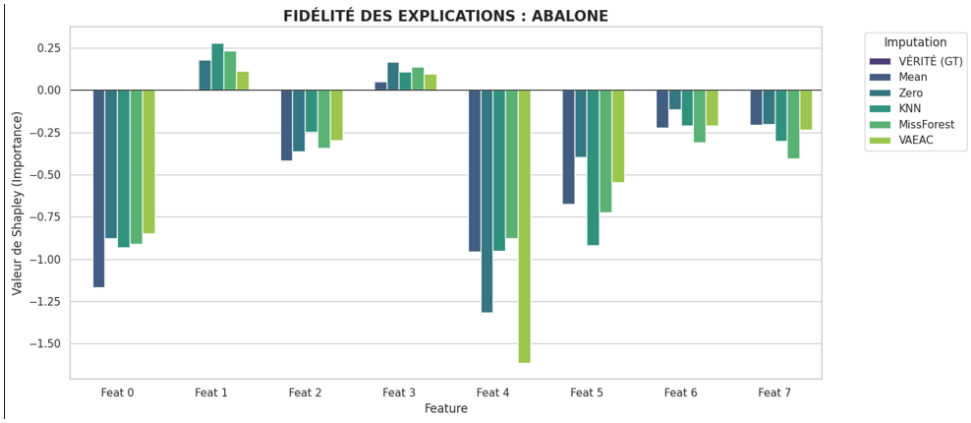


FIGURE 1 – Évaluation de la fidélité des valeurs de Shapley (Moyenne ± Écart-type) selon la méthode d'imputation sur le dataset *Abalone*. Les résultats sont agrégés sur 3 itérations (*runs*) indépendantes.

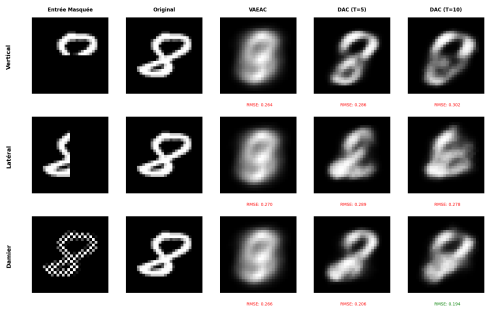


FIGURE 2 – **Comparaison visuelle et métrique de l'imputation (VAEAC vs DAC).** Reconstruction du chiffre "8" sous trois scénarios de masquage (Vertical, Latéral, Damier). Les résultats comparent le VAEAC classique et le modèle DAC avec $T = 5$ et $T = 10$ pas de diffusion. *Détails expérimentaux :* Entraînement sur **10 époques** avec une taille de lot (*batch size*) de **64**. Ces résultats sont issus d'une **architecture générale légèrement modifiée** et utilisent une **fonction de perte adaptée**. Les scores RMSE (en rouge/vert) indiquent la qualité de reconstruction pixel à pixel.

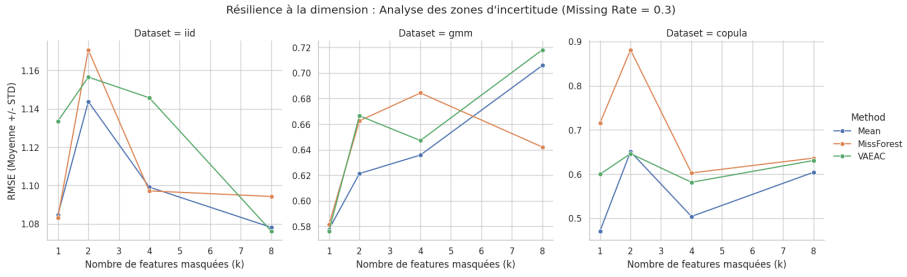


FIGURE 3 – **Analyse des zones d'incertitude :** Évolution de l'erreur de reconstruction (RMSE) en fonction du nombre de caractéristiques masquées ($k \in \{1, 2, 4, 8\}$) pour les datasets IID, GMM et Copula. Le taux de données manquantes global est fixé à 30%.