MisShapForest : Imputation Fiable des Données Manquantes par MissForest Amélioré avec SHAP

Sifeddine Sellami[1], Louenas Bounia[1], Juba Agoun[2], Sebastien Destercke[3], Céline Rouveirol[1]

[1] Université Sorbonne Paris Nord, LIPN-UMR CNRS 7030, Villetaneuse, France
[2] Université Lumière Lyon 2, Laboratoire ERIC, Lyon, France
[3] Université de technologie de Compiègne, UMR CNRS 7253 Heudiasyc, France

{sellami, bounia, rouveirol}@lipn.univ-paris13.fr, juba.agoun1@univ-lyon2.fr, sebastien.destercke@hds.utc.fr

Résumé

Les valeurs manquantes représentent un problème omniprésent dans les jeux de données réels, affectant sévèrement la fiabilité des analyses de données et des modèles prédictifs. Bien que de nombreuses techniques d'imputation existent — allant de méthodes statistiques simples à des approches avancées d'apprentissage automatique telles que missForest — la plupart imputent systématiquement toutes les valeurs manquantes sans évaluer la fiabilité de l'imputation. Cela entraîne des estimations biaisées et une propagation des erreurs, en particulier lorsque les imputations reposent sur des informations insuffisantes ou déjà imputées. Nous présentons MisShapForest, un nouveau cadre combinant missForest avec l'explicabilité basée sur SHAP afin de quantifier la fiabilité de chaque valeur imputée. Notre méthode n'impute les valeurs manquantes que lorsque les prédictions sont soutenues par des caractéristiques non manquantes suffisamment informatives ; dans le cas contraire, les lignes sont écartées. En exploitant les valeurs SHAP, nous identifions quelles caractéristiques influencent significativement chaque imputation et vérifions leur fiabilité. Des expériences menées sur plusieurs jeux de données de référence (Adult, Credit, Ecoli, Heart Disease et Parkinson), avec différents taux de valeurs manquantes, montrent que notre approche surpasse systématiquement missForest standard ainsi que les méthodes d'imputation classiques sur les lignes imputées de manière fiable, en obtenant des erreurs d'imputation plus faibles et une robustesse améliorée lorsque le taux de valeurs manquantes augmente. Ce travail souligne l'importance d'intégrer l'explicabilité dans les chaînes de prétraitement de données pour une imputation fiable des valeurs manquantes.

Mots-clés

Imputation des données manquantes, Valeurs SHAP, Forêts aléatoires, Évaluation de la qualité des données

Abstract

Missing values are a pervasive issue in real-world datasets, severely affecting the reliability of data analysis and predictive modeling. While numerous imputation techniques exist—from simple statistical methods to advanced machine learning approaches like missForest—most systematically impute all missing values without assessing imputation reliability. This leads to biased estimates and error propagation, particularly when imputations rely on insufficient or previously imputed information. We introduce MisShapForest, a novel framework that combines missForest with SHAP-based explainability to quantify the reliability of each imputed value. Our method imputes missing values only when predictions are supported by sufficiently informative non-missing features ; otherwise, rows are discarded. By leveraging SHAP values, we identify which features significantly influence each imputation and verify their reliability. Experiments on multiple benchmark datasets (Adult, Credit, Ecoli, Heart Disease, and Parkinson) under varying missingness rates demonstrate that our approach consistently outperforms standard missForest and classical imputation methods on reliably imputed rows, achieving lower imputation error and improved robustness as missingness increases. This work highlights the importance of integrating explainability into data preprocessing pipelines for trustworthy missing data imputation.

Keywords

Missing Data Imputation, SHAP Values, Random Forests, Data Quality Assessment

1 Introduction

Dans le domaine de la science des données, les données revêtent une importance majeure, car la construction d'analyses solides et l'obtention de conclusions précises nécessitent des données de haute qualité. Cependant, le processus de collecte de données est généralement désorganisé, ce qui conduit à des jeux de données contenant de nombreuses valeurs manquantes [8]. Ces valeurs manquantes posent un problème particulièrement significatif, car elles peuvent amener les analystes à tirer des conclusions inexactes[13]. La solution la plus simple consiste à réduire le jeu de données en supprimant les instances comportant des valeurs manquantes. Bien que directe, cette stratégie peut entraîner une perte substantielle d'information. Une autre solu-

tion possible est l'imputation des valeurs manquantes. Cette imputation doit être effectuée avec précaution afin d'éviter d'introduire des biais dans le jeu de données [2]. Les méthodes couramment utilisées pour imputer les valeurs manquantes incluent des techniques simples telles que l'imputation par la moyenne, ainsi que des approches plus sophistiquées comme l'Imputation Multiple par Équations Chaînées (MICE), les k Plus Proches Voisins (KNN), et la méthode missForest, qui repose sur les forêts aléatoires.

Le problème est que même avec ces méthodes d'imputation, une imputation médiocre peut introduire des biais dans les données, entraîner une perte de puissance statistique et, en fin de compte, conduire à des conclusions erronées, selon les mécanismes sous-jacents à l'origine des données manquantes [16]. Dans l'étude de [11], les auteurs montrent que les méthodes d'imputation surpassent généralement la simple suppression dans la plupart des scenarios. Cependant, lorsqu'il s'agit de jeux de données volumineux comportant plus de 50% de valeurs manquantes, la suppression peut produire de meilleures performances prédictives, bien qu'au prix d'une perte substantielle d'information. Sur la base de ces observations, les auteurs émettent l'hypothèse qu'une stratégie hybride combinant suppression et imputation pourrait être bénéfique, par laquelle les lignes ou colonnes présentant une forte proportion de valeurs manquantes sont supprimées, tandis que les données manquantes restantes sont imputées. Néanmoins, ils soulignent également que peu d'attention a été accordée dans la littérature pour combiner systématiquement les stratégies de suppression et d'imputation, mettant en évidence une lacune de recherche ouverte.

Dans notre travail, nous combinons à la fois l'imputation et la suppression sélective en employant une méthode d'imputation basée sur missForest [14] conjointement avec SHAP [10] pour l'explicabilité afin d'évaluer la fiabilité de l'imputation des valeurs manquantes. Lorsque suffisamment d'informations sont disponibles, missForest est utilisé pour imputer les valeurs manquantes. Inversement, lorsque les informations disponibles sont insuffisantes pour garantir une imputation fiable, les instances correspondantes sont supprimées. L'utilisation de l'explicabilité nous permet de déterminer si l'imputation d'une valeur manquante est appropriée, aboutissant à un jeu de données complété où seules les instances disposant d'informations support adéquates sont conservées.

Nous commençons par un aperçu des types de données manquantes, suivi d'une revue des méthodes d'imputation existantes et de leurs limites, avec un accent particulier sur la méthode missForest. Nous présentons ensuite notre approche, menons plusieurs expériences, et concluons finalement notre travail.

2 Techniques d'Imputation des Valeurs Manquantes

Plusieurs méthodes d'imputation classiques sont couramment utilisées dans la littérature pour traiter les données manquantes. Nous passons brièvement en revue les principales approches avant de nous concentrer sur missForest, le fondement de notre méthode.

2.1 Méthodes d'Imputation Classiques

— **Statistiques Récapitulatives (Moyenne/Médiane/Mode) :** Ces méthodes remplacent les valeurs manquantes par la moyenne, la médiane ou le mode des valeurs disponibles. Bien que simples à mettre en œuvre et préservant la taille de l'échantillon, elles peuvent introduire des biais et ne tiennent pas compte des relations entre les variables [4].

— **K Plus Proches Voisins (KNN) :** L'imputation KNN [15] estime les valeurs manquantes en utilisant les k voisins les plus proches basés sur les caractéristiques disponibles, typiquement en utilisant la distance euclidienne. La valeur manquante est imputée en utilisant une statistique récapitulative (moyenne, médiane ou moyenne pondérée) à partir des instances voisines. Cette méthode préserve les motifs locaux mais peut être coûteuse en calcul et sensible au choix de k.

— **Imputation Multiple par Équations Chaînées (MICE) :** MICE [17] génère plusieurs jeux de données complets en imputant de manière itérative les valeurs manquantes pour chaque variable en utilisant des modèles conditionnels avec d'autres variables comme prédicteurs. Cette approche permet l'estimation de l'incertitude mais nécessite de combiner les résultats à travers plusieurs jeux de données.

2.2 La Méthode MissForest

La méthode *missForest* [14] utilise les Forêts Aléatoires pour imputer les valeurs manquantes dans une matrice de données $X = (X_1, X_2, \ldots, X_p)$ de dimensions $n \times p$. Après une imputation initiale (par exemple, en utilisant la moyenne), les variables sont triées par ordre croissant du nombre de valeurs manquantes. Pour chaque variable X_s, les valeurs manquantes sont imputées en ajustant un modèle de Forêt Aléatoire sur les données observées en utilisant les autres variables comme prédicteurs. Cette procédure itère jusqu'à convergence, définie par un critère de stabilité entre les itérations successives (Algorithme 2.2).

2.3 Métriques d'Évaluation

La performance de l'imputation est évaluée en utilisant les métriques de [14]. Pour les **variables continues**, l'Erreur Quadratique Moyenne Normalisée (NRMSE) :

$$\text{NRMSE} = \frac{\sqrt{\text{mean}\left((X_{\text{true}} - X_{\text{imp}})^2\right)}}{\text{std}(X_{\text{true}})} \tag{1}$$

Pour les **variables catégorielles**, la Proportion de Valeurs Mal Classées (PFC) :

$$\text{PFC} = \frac{1}{|F|} \sum_{i \in F} \mathbb{I}(X_{\text{true},i} \neq X_{\text{imp},i}) \tag{2}$$

Algorithm 1 : Missing Data Imputation with Random Forest (missForest)

Require: A matrix $X \in \mathbb{R}^{n \times p}$, a stopping criterion γ
Ensure: The imputed matrix X_{imp}

```
Perform an initial estimation of missing values (e.g., mean)
Sort the columns of X according to the increasing amount of missing values, store the indices in k
while stopping criterion γ is not met do
    Save the previous imputed matrix X_imp^old
    for each variable s ∈ k do
        Fit a random forest : y_obs^(s) ~ x_obs^(s)
        Predict the missing values : y_mis^(s) ← x_mis^(s)
        Update the imputed matrix X_imp^new
    end for
    Update the criterion γ
end while
return X_imp
```

où X_{true} représente les valeurs vraies, X_{imp} les valeurs imputées, F est l'ensemble des indices de valeurs manquantes, et $\mathbb{I}(\cdot)$ est la fonction indicatrice. Des valeurs proches de 0 indiquent une bonne performance ; des valeurs proches de 1 indiquent une qualité médiocre.

2.4 Limites des Méthodes Précédentes

La principale limite de ces méthodes d'imputation est qu'elles remplissent systématiquement toutes les valeurs manquantes sans évaluer la fiabilité. Par exemple, avec missForest, si une ligne ne contient qu'une seule variable observée, l'imputation repose uniquement sur cette unique information, qui peut être insuffisante pour une estimation fiable. De telles lignes—malgré un taux de valeurs manquantes élevé—sont toujours complétées, et leurs valeurs imputées peuvent ensuite être utilisées pour imputer d'autres données, propageant potentiellement des erreurs et de l'incertitude.

Des études montrent que les méthodes d'imputation ignorant la fiabilité peuvent affecter négativement l'équité des modèles, en particulier pour les jeux de données présentant des motifs de valeurs manquantes variés et des taux élevés de valeurs manquantes [11, 6]. De plus, ces méthodes peuvent déformer les estimations de l'importance des caractéristiques, affectant l'interprétabilité du modèle [18]. Par exemple, [5] a démontré que l'application d'une imputation standard sur le jeu de données Wine [1] sous-estimait drastiquement l'importance de la caractéristique "Alcohol", qui est connue pour être critique dans la prédiction de la qualité du vin.

Pour résoudre ces problèmes, nous proposons une méthode qui non seulement impute les valeurs manquantes mais identifie et supprime également les lignes non fiables—celles pour lesquelles l'imputation est susceptible d'être inexacte ou non robuste. Ceci est réalisé en utilisant l'explicabilité basée sur SHAP pour déterminer si les caractéristiques observées soutiennent suffisamment l'imputation. Notre approche réduit les biais du modèle et préserve les estimations de l'importance des caractéristiques, garantissant des modèles prédictifs plus fiables et robustes.

3 Notre Méthode d'Imputation "MisShapForest"

Nous proposons une méthode, MisSHAPForest, qui combine missForest avec les explications SHAP pour imputer uniquement les lignes imputées de manière fiable tout en supprimant celles qui ne le sont pas. Pour chaque imputation de valeur manquante, notre méthode vérifie si elle est basée sur des valeurs observées (non manquantes) ou sur des données précédemment imputées qui peuvent elles-mêmes être non fiables.

3.1 Méthode de Détection de Fiabilité

Pour imputer une valeur manquante x_{ij} de l'instance i et de la colonne j, missForest entraîne un modèle de Forêt Aléatoire f_j sur les lignes où la colonne j est observée. Nous vérifions la fiabilité de cette prédiction comme suit (Algorithme 3.1) :

1. **Calculer les valeurs SHAP :** Nous utilisons les valeurs SHAP (SHapley Additive exPlanations) [9] calculées via TreeSHAP pour quantifier la contribution de chaque caractéristique à la prédiction. SHAP fournit une décomposition additive : $f_j(x_i) = \phi_0 + \sum_{c=1}^{p} \phi_c$, où ϕ_0 est la valeur de base (prédiction moyenne sur les données d'entraînement) et ϕ_c est la contribution de la caractéristique c à la prédiction pour l'instance i.

2. **Définir l'importance des caractéristiques :** Nous définissons l'effet en pourcentage de la caractéristique c comme :

   $$\text{Effet en pourcentage de la caractéristique } c = \frac{|\phi_c|}{|\phi_0|}$$

   Nous utilisons ϕ_0 comme référence pour mesurer l'effet absolu d'une variable sur la valeur imputée—combien elle modifie la prédiction par rapport à la prédiction moyenne du modèle—plutôt que son importance relative par rapport aux autres variables.

3. **Appliquer le seuil :** Un seuil τ défini par l'utilisateur détermine si une caractéristique influence significativement la prédiction. Si l'effet en pourcentage dépasse τ, la caractéristique est considérée comme importante. Nous fixons $\tau = 10\%$ comme un compromis raisonnable entre sensibilité et robustesse, bien que cela puisse être ajusté selon le contexte.

4. **Évaluer la fiabilité :** La valeur imputée x_{ij} est considérée comme non fiable si au moins une caractéristique importante est elle-même manquante ou a été précédemment imputée de manière non fiable (par exemple, via l'imputation par la moyenne). Sinon, l'imputation est jugée fiable.

Algorithm 2 : Function isReliable(Rf, x, j, X) – Checking the reliability of imputation

Input : Rf trained Random Forest model, x incomplete instance, j index of the column to impute, X complete dataset
Output : true if the prediction is reliable, otherwise false
$shap\_values \leftarrow$ compute SHAP values for instance x with respect to the prediction of column j using model Rf
$base\_value \leftarrow$ mean prediction of model Rf on dataset X for column j
$important\_features \leftarrow$ empty set
for each column c such that $c \neq j$ **do**
    $effect\_c \leftarrow \frac{|shap\_values[c]|}{|base\_value|}$
    if $effect\_c > 0.1$ **then**
        add c to $important\_features$
    end if
end for
for each column c in $important\_features$ **do**
    if $x[c]$ is missing in X_{reliable} **then**
        return false
    end if
end for
return true

3.2 Algorithme d'Imputation Fiable Basé sur missForest et SHAP

En s'appuyant sur la méthode précédente et l'algorithme missForest, nous proposons l'algorithme d'imputation suivant, qui effectue une imputation fiable tout en supprimant les instances qui ne peuvent pas être imputées de manière fiable (voir Algorithme 3 et figure 1). L'algorithme procède comme suit :

1. Imputer initialement les valeurs manquantes, par exemple par la moyenne.
2. Itérer sur les colonnes en commençant par celles contenant le plus de valeurs manquantes.
3. Pour chaque colonne c, entraîner un modèle de Forêt Aléatoire (régression si la colonne cible est numérique, classification si elle est catégorielle) sur les instances où la colonne c est observée (non manquante).
4. Prédire les valeurs manquantes pour la colonne c et, pour chaque prédiction, vérifier sa fiabilité en utilisant la méthode décrite précédemment. Ensuite, marquer chaque prédiction comme fiable ou non fiable.
5. Répéter les étapes 2 à 4 jusqu'à ce que le nombre de prédictions fiables se stabilise.
6. Répéter les étapes 2 à 5 jusqu'à ce que la convergence soit atteinte, définie par un critère de stabilité entre deux itérations successives (critère utilisé dans `missForest`).
7. Supprimer les lignes contenant des colonnes imputées de manière non fiable.
8. Retourner le jeu de données contenant uniquement les lignes avec des valeurs imputées de manière fiable.

Objectif de l'Étape 5. L'étape 5 est conçue pour assurer la gestion des situations où l'ordre d'imputation affecte la fiabilité. Par exemple, considérons une ligne avec deux valeurs manquantes : les colonnes C_1 et C_2. Supposons que C_2 soit importante pour prédire C_1, mais que C_1 ne soit pas importante pour prédire C_2. Si nous commençons par imputer C_1, son imputation sera non fiable car elle dépend de la valeur manquante C_2. Cependant, si nous imputons d'abord C_2, qui ne dépend pas de C_1, nous pouvons ensuite revenir à l'imputation de C_1 de manière fiable, en utilisant la valeur maintenant prédite de C_2. L'étape 5 itère sur les colonnes dans différents ordres pour résoudre de telles dépendances asymétriques et maximiser le nombre d'imputations fiables. Les seules lignes qui restent non fiables sont celles pour lesquelles aucun ordre d'imputation ne peut satisfaire toutes les dépendances ; par exemple, dans le cas de dépendances circulaires : C_i est importante pour prédire C_j, et C_j est importante pour prédire C_i, alors que les deux sont manquantes.

4 Expériences

Pour démontrer l'efficacité de notre méthode, *MisShapForest*, nous avons mené des expériences sur cinq jeux de données différents provenant de domaines variés : Adult, Credit, Ecoli, Heart Disease et Parkinson. Les valeurs manquantes ont été générées aléatoirement à différents taux de valeurs manquantes : 15%, 25%, 50% et 75%. Pour chaque taux, nous avons effectué cinq expériences indépendantes, en appliquant diverses méthodes d'imputation : Moyenne, Médiane, MICE, KNN, MissForest et notre méthode. Les valeurs manquantes ont été introduites en sélectionnant aléatoirement un ensemble de cellules dans la matrice de données selon le taux de valeurs manquantes spécifié. Ici, une cellule fait référence à une entrée unique dans la matrice, correspondant à l'intersection d'une ligne (une observation) et d'une colonne (une caractéristique). Le taux de valeurs manquantes représente le pourcentage de toutes les cellules dans le jeu de données, c'est-à-dire le nombre total de lignes multiplié par le nombre total de colonnes. Par exemple, un taux de valeurs manquantes de 5% indique que 5% de toutes les entrées dans la matrice sont définies comme manquantes. Les valeurs supprimées ont été enregistrées pour permettre la comparaison des résultats d'imputation avec les valeurs réelles. Dans chaque expérience, la performance d'imputation a été évaluée en utilisant les métriques suivantes :

— **NRMSE (Erreur Quadratique Moyenne Normalisée)** (1) pour les colonnes numériques
— **PFC (Proportion de Valeurs Mal Classées)** (2) pour les colonnes catégorielles

Nous avons ensuite calculé la performance moyenne pour chaque taux de valeurs manquantes afin de comparer l'effi-

Algorithm 3 : Imputation based on Random Forest and SHAP - MisShapForest

$X \leftarrow$ dataset of size (n, m)
$X_{\text{reliable}} \leftarrow X$
Initially estimate missing values in X (e.g., mean imputation)
while stopping criterion γ is not satisfied **do**
  $missing\_count \leftarrow -1$
  $previous\_missing\_count \leftarrow -2$
  while $missing\_count \neq previous\_missing\_count$ **and** $missing\_count \neq 0$ **do**
    $previous\_missing\_count \leftarrow missing\_count$
    for each column col in X **do**
      $missing\_indices \leftarrow$ indices of rows where $X[col]$ is missing
      $X_{\text{train}} \leftarrow X$ without rows $missing\_indices$ and without column col
      $y_{\text{train}} \leftarrow X[col]$ without rows $missing\_indices$
      if col is numerical **then**
        $Rf \leftarrow$ train Random Forest Regressor on $(X_{\text{train}}, y_{\text{train}})$
      else
        $Rf \leftarrow$ train Random Forest Classifier on $(X_{\text{train}}, y_{\text{train}})$
      end if
      $X_{\text{pred}} \leftarrow X$ with rows $missing\_indices$ and without column col
      for each instance x in X_{pred} **do**
        $i \leftarrow$ index of x in X
        if isReliable(Rf, x, col, X) **then**
          $X_{\text{reliable}}[i, col] \leftarrow Rf.predict(x)$
        end if
        $X[i, col] \leftarrow Rf.predict(x)$
      end for
    end for
    $previous\_missing\_count \leftarrow missing\_count$
    $missing\_count \leftarrow$ current number of missing cells in X_{reliable}
  end while
  $X_{\text{out}} \leftarrow$ remove rows with missing values from X_{reliable}
  return X_{out}
end while

cacité des différentes méthodes d'imputation. Les résultats sont présentés dans les Tableaux 1 et 2. Certains jeux de données, tels que **Adult, Credit et Ecoli**, contiennent un mélange de colonnes numériques et catégorielles. Puisque notre méthode vise à identifier les instances qui sont mal imputées et à les supprimer, il est nécessaire de définir une métrique qui combine **NRMSE** et **PFC**. Ceci est dû au fait qu'une instance unique peut contenir des valeurs manquantes à la fois dans des colonnes numériques et catégorielles, et la décision de supprimer l'instance devrait dépendre de la performance sur les deux types de caractéristiques. Nous proposons donc une métrique d'erreur combinée :

$$\text{Erreur} = \alpha \cdot \text{NRMSE} + (1 - \alpha) \cdot \text{PFC}, \tag{3}$$

où α représente la proportion de colonnes numériques par rapport au nombre total de colonnes. Cette métrique combinée nous permet d'évaluer la fiabilité globale de chaque instance de manière plus précise que de considérer NRMSE et PFC séparément.

Nous utilisons cette métrique pour comparer la performance à travers trois scénarios :

— Sur les **instances conservées** (c'est-à-dire, celles considérées comme imputées de manière fiable par notre méthode),
— Sur les **instances supprimées** (c'est-à-dire, celles considérées comme imputées de manière non fiable),
— Sur **toutes les instances** en utilisant l'imputation MissForest standard.

Ces résultats sont illustrés dans le Tableau 3 et la Figure 2

Enfin, nous analysons l'évolution de l'erreur d'imputation en fonction du taux de valeurs manquantes, en comparant MissForest et notre méthode. Les résultats sont présentés dans la Figure 3

5 Discussion

Les résultats expérimentaux présentés dans les Tableaux 1 et 2 démontrent que notre méthode proposée atteint une meilleure performance globale comparée aux approches d'imputation de référence. Pour les colonnes numériques, le Tableau 1 montre que notre méthode surpasse systématiquement les méthodes concurrentes sur les cinq jeux de données, atteignant le NRMSE le plus faible dans 18 confi-

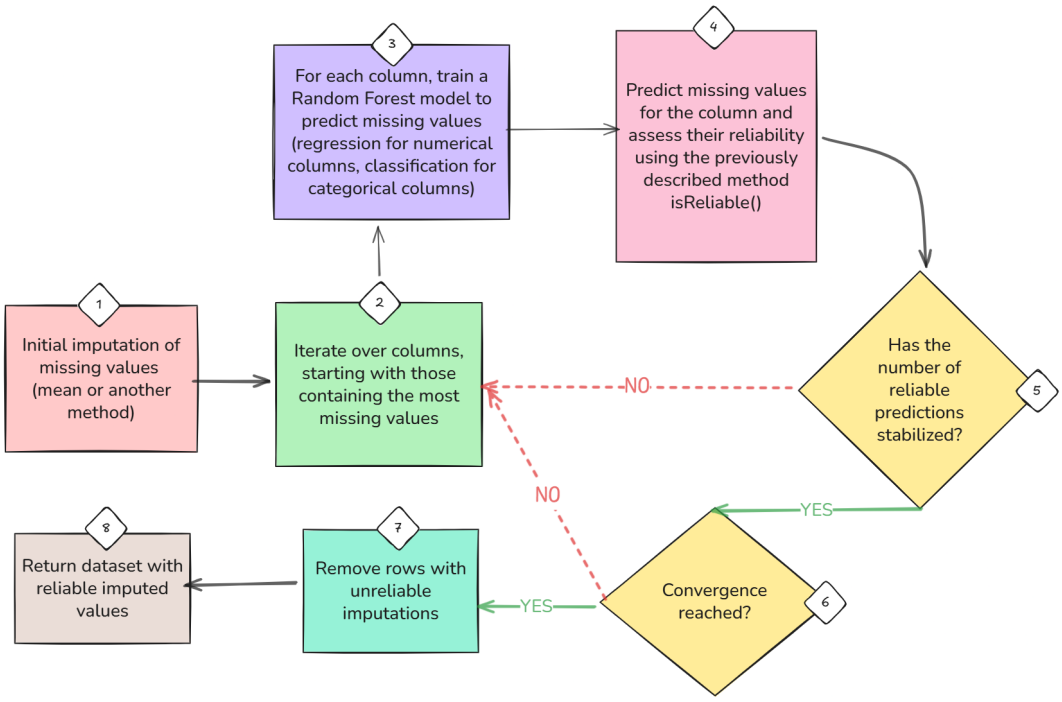

FIGURE 1 – Imputation itérative par Forêt Aléatoire avec vérification de fiabilité

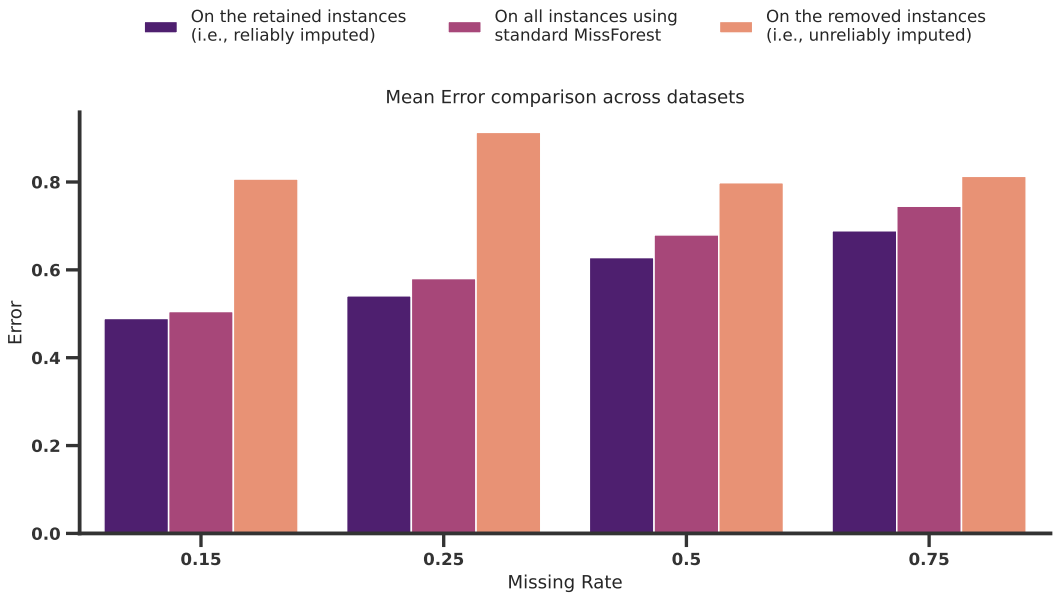

FIGURE 2 – Comparaison de la métrique d'erreur combinée (Équation 3) à travers trois scénarios : instances conservées, instances supprimées, et toutes les instances utilisant MissForest standard.

gurations expérimentales sur 20.

Pour les colonnes catégorielles, notre méthode fournit la meilleure performance aux taux de valeurs manquantes élevés (50% et 75%). Aux taux de valeurs manquantes plus faibles (15% et 25%), MissForest surpasse légèrement notre approche. Ce comportement peut s'expliquer par le mécanisme de sélection au niveau des instances de notre méthode. Spécifiquement, lorsqu'une instance contient des valeurs manquantes à la fois dans des colonnes numériques et catégorielles, elle peut être supprimée si l'imputation de ses caractéristiques numériques est jugée non fiable, même si l'imputation catégorielle est précise. En conséquence, certaines valeurs catégorielles bien imputées sont écartées, ce qui peut conduire à une augmentation marginale de la métrique PFC. Cette observation explique pourquoi MissForest obtient occasionnellement de meilleurs scores PFC aux niveaux de valeurs manquantes plus faibles.

Pour résoudre cette limitation et fournir une évaluation

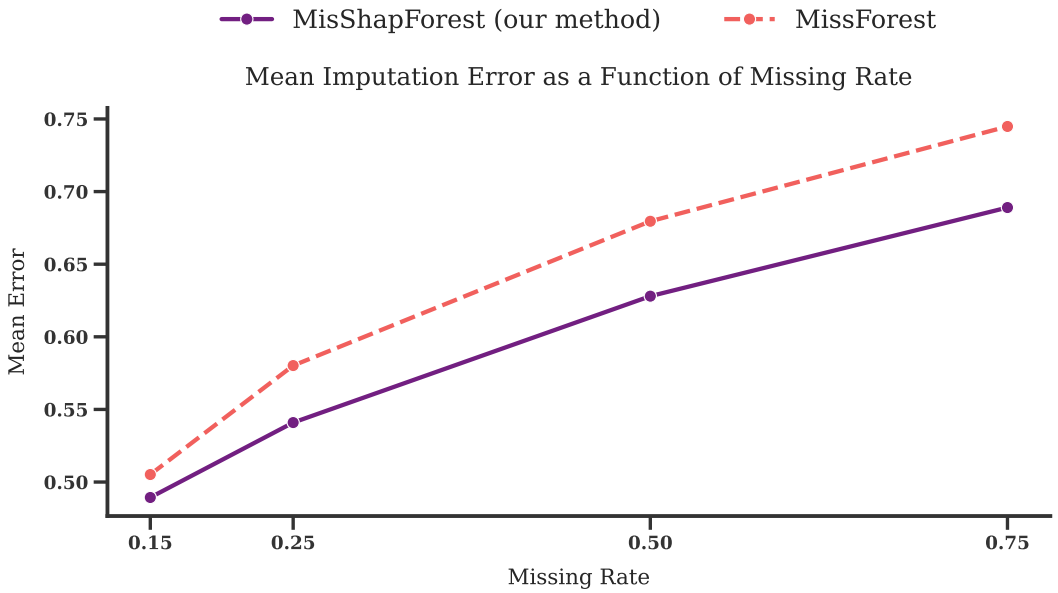

FIGURE 3 – Évolution de l'erreur d'imputation en fonction du taux de valeurs manquantes, en comparant MissForest et notre méthode proposée.

équitable pour les jeux de données contenant à la fois des caractéristiques numériques et catégorielles, nous avons introduit la métrique d'erreur combinée définie dans l'Équation 3. Comme le montre le Tableau 3, lorsque cette métrique unifiée est utilisée, notre méthode surpasse systématiquement MissForest sur tous les jeux de données et tous les taux de valeurs manquantes. Ceci confirme que notre approche capture plus efficacement la qualité globale de l'imputation au niveau des instances. De plus, les résultats rapportés dans le Tableau 3 et illustrés dans la Figure 2 démontrent la capacité de notre méthode à identifier et supprimer avec précision les instances imputées de manière non fiable. En particulier, l'erreur mesurée sur les instances supprimées est significativement plus élevée que celle observée sur les instances conservées, indiquant que les lignes écartées correspondent à des imputations véritablement non fiables. Cette suppression sélective permet à notre méthode de produire un jeu de données complété plus fiable et d'atteindre une erreur globale plus faible comparée à MissForest.

Enfin, les Figures 3 mettent en évidence la robustesse de notre approche à mesure que le taux de valeurs manquantes augmente. Alors que l'erreur d'imputation de MissForest croît régulièrement avec des niveaux plus élevés de données manquantes, notre méthode présente un comportement plus stable en excluant les instances dont les imputations sont basées sur des informations insuffisantes ou non fiables.

5.1 Limitations de notre méthode

La capacité de notre méthode à identifier de manière fiable les lignes imputées en utilisant SHAP, une approche basée sur l'explicabilité, et à écarter celles imputées de manière non fiable constitue un avantage clair par rapport à MissForest et aux autres méthodes d'imputation. Cependant, comme le montre le Tableau 3, lorsque le taux de valeurs manquantes est élevé, une proportion plus importante de lignes est supprimée. Par exemple, dans le jeu de données Credit avec un taux de valeurs manquantes de 75%, jusqu'à 80,26% des instances sont écartées. Ce comportement est attendu, car une forte proportion de valeurs manquantes parmi les caractéristiques augmente la probabilité qu'une instance soit imputée de manière non fiable, déclenchant ainsi sa suppression. Néanmoins, cela peut être considéré comme une limitation de notre approche, puisque dans des conditions de valeurs manquantes extrêmes, seul un sous-ensemble relativement restreint d'instances reste disponible après l'imputation.

Une autre limitation de notre approche est sa complexité computationnelle plus élevée comparée à MissForest. Alors que MissForest prédit uniquement les valeurs manquantes en utilisant des modèles de Forêts Aléatoires, notre méthode nécessite des calculs supplémentaires après chaque prédiction, à savoir le calcul des valeurs SHAP et la vérification de la fiabilité de l'imputation, comme décrit dans la Section 3.1. Par conséquent, en termes de temps de calcul, notre approche est plus coûteuse que MissForest et les autres méthodes d'imputation standard.

Enfin, notre méthode nécessite de définir un seuil pour déterminer si une caractéristique est considérée comme importante pour une prédiction donnée, comme discuté dans la Section 3.1. Dans nos expériences, nous avons utilisé un seuil de 10% ; cependant, cette valeur est un paramètre défini par l'utilisateur. Il devrait être sélectionné en fonction du niveau d'influence souhaité qu'une caractéristique doit avoir, soit en augmentant soit en diminuant la prédiction—pour être considérée comme importante dans le processus.

Dataset	Method	15%	25%	50%	75%
Adult	Mean	0.877	0.949	1.001	0.993
	Median	0.881	0.961	1.010	1.001
	MICE	0.873	0.937	1.022	1.012
	KNN	0.915	0.959	1.006	0.993
	MissForest	0.854	0.920	1.086	1.125
	Our method	**0.799**	**0.845**	**0.993**	**0.978**
Credit	Mean	0.886	0.875	0.971	0.939
	Median	0.952	0.941	1.031	1.005
	MICE	0.820	0.807	1.057	1.031
	KNN	0.859	0.855	0.950	**0.928**
	MissForest	0.842	0.885	1.045	1.099
	Our method	**0.798**	**0.777**	**0.937**	1.026
Ecoli	Mean	0.909	0.994	1.065	1.038
	Median	0.902	1.004	1.074	1.053
	MICE	0.741	0.867	1.123	1.002
	KNN	0.757	0.838	0.958	0.973
	MissForest	0.710	0.889	0.919	0.992
	Our method	**0.684**	**0.789**	**0.871**	**0.952**
Heart	Mean	0.995	1.001	0.996	0.999
	Median	1.130	1.137	1.150	1.126
	MICE	0.882	0.892	1.053	0.985
	KNN	0.880	0.896	0.902	**0.941**
	MissForest	0.498	0.638	0.885	1.001
	Our method	**0.480**	**0.619**	**0.839**	0.945
Parkinson	Mean	0.960	0.972	0.991	0.993
	Median	0.993	0.998	1.021	1.026
	MICE	0.388	0.425	0.509	0.600
	KNN	0.870	0.881	0.922	0.939
	MissForest	0.360	0.397	0.472	0.540
	Our method	**0.334**	**0.339**	**0.362**	**0.441**

TABLE 1 – NRMSE comparison for 4 datasets at different missing data rates.

Dataset	Method	15%	25%	50%	75%
Adult	Mean	0.6168	0.6202	0.6310	0.6251
	Median	0.4361	0.4477	0.4501	0.4452
	MICE	0.5822	0.5866	0.6107	0.6056
	KNN	0.5857	0.5668	0.6080	0.5866
	MissForest	**0.2742**	**0.3005**	0.3359	0.3725
	Our method	0.2873	0.3200	**0.3343**	**0.3725**
	% deleted	4.00	6.11	15.83	26.51
Credit	Mean	0.5486	0.5624	0.5628	0.6008
	Median	0.4961	0.4960	0.4987	0.4996
	MICE	0.4563	0.4592	0.4929	0.5018
	KNN	0.5785	0.5672	0.5553	0.5455
	MissForest	**0.3541**	0.3566	0.3865	0.4160
	Our method	0.3573	**0.3528**	**0.3860**	**0.4037**
	% deleted	10.23	23.56	49.10	80.26
Ecoli	Mean	0.9935	0.9945	0.9976	0.9965
	Median	0.8792	0.8779	0.8863	0.8787
	MICE	0.8673	0.9070	0.8978	0.9211
	KNN	0.8967	0.8599	0.9419	0.9597
	MissForest	**0.6005**	**0.6196**	**0.6244**	0.6816
	Our method	0.6369	0.6482	0.6311	**0.6497**
	% deleted	2.56	6.96	16.55	28.15

TABLE 2 – PFC comparison for 3 datasets at different missing data rates.

6 Travaux Connexes

Les recherches existantes sur l'imputation des valeurs manquantes peuvent être organisées en trois thèmes principaux : la quantification de l'incertitude, l'évaluation de l'impact et les stratégies hybrides.

Approches de Quantification de l'Incertitude. Plusieurs travaux ont abordé l'incertitude inhérente à l'imputation des valeurs manquantes. Les méthodes d'imputation multiple [12] génèrent plusieurs jeux de données complets et analysent la variabilité entre les imputations pour fournir des mesures d'incertitude rigoureuses. Les approches basées sur des ensembles estiment l'incertitude à travers le désaccord entre modèles ou la variance des prédictions. Par exemple, [7] utilise plusieurs modèles d'apprentissage automatique pour l'imputation, le désaccord inter-modèles servant d'indicateur d'incertitude. Bien que ces méthodes reconnaissent que toutes les imputations ne sont pas également fiables, elles ne fournissent pas de mécanismes explicites pour identifier les imputations incorrectes ou de stratégies systématiques pour améliorer la qualité de l'imputation basée sur la fiabilité.

Impact sur les Tâches en Aval. D'autres études ont démontré les conséquences négatives des imputations non fiables. [3] a montré que des imputations inexactes peuvent dégrader significativement la performance des modèles prédictifs, soulignant que traiter toutes les imputations comme équivalentes est problématique. Plus récemment, [5] a démontré que des imputations incorrectes peuvent déformer les estimations de l'importance des caractéristiques, altérant ainsi l'interprétabilité du modèle et conduisant potentiellement à des conclusions trompeuses. Ces résultats soulignent la nécessité d'une imputation fiable plutôt que d'une complétion systématique de toutes les valeurs manquantes.

Stratégies Hybrides de Suppression-Imputation. [11] a montré que lorsque le taux de valeurs manquantes dépasse 50%, la suppression peut surpasser l'imputation malgré la perte d'information. Ils ont proposé une stratégie hybride combinant suppression et imputation : supprimer les lignes ou colonnes avec un taux élevé de valeurs manquantes tout en imputant les valeurs restantes. Cela met en évidence la nécessité de distinguer entre imputations fiables et non fiables, bien qu'ils n'aient pas fourni de méthode systématique pour cette distinction.

Notre Contribution. Contrairement à ces approches, nous proposons un mécanisme explicite d'évaluation de la fiabilité utilisant les valeurs SHAP. En identifiant quelles variables observées influencent significativement chaque imputation et en vérifiant leur fiabilité, nous conservons uniquement les lignes avec des imputations basées sur des informations suffisantes et fiables tout en écartant celles qui ne le sont pas. Cela fournit un cadre principiel et interprétable pour combiner imputation et suppression sélective.

Dataset	Scenario	15%	25%	50%	75%
Adult	Retained	**0.4921**	**0.5298**	**0.5979**	**0.6147**
	Removed	0.8414	0.8307	0.8370	0.8363
	MissForest	0.5061	0.5482	0.6358	0.6735
	% deleted	4.00%	6.11%	15.83%	26.51%
Credit	Retained	**0.4674**	**0.4588**	**0.5238**	**0.5593**
	Removed	0.5530	0.5862	0.5795	0.5935
	MissForest	0.4761	0.4888	0.5512	0.5868
	% deleted	10.23%	23.57%	49.10%	80.26%
Ecoli	Retained	**0.6738**	**0.7581**	**0.8175**	**0.8849**
	Removed	1.1420	1.7791	1.0365	1.0193
	MissForest	0.6858	0.8292	0.8538	0.9227
	% deleted	2.56%	6.96%	16.55%	28.15%
Heart	Retained	**0.4798**	**0.6188**	**0.8385**	**0.9450**
	Removed	0.8715	0.7583	0.9598	1.0324
	MissForest	0.4980	0.6379	0.8851	1.0014
	% deleted	4.66%	13.66%	38.46%	64.55%
Parkinson	Retained	**0.3338**	**0.3393**	**0.3622**	**0.4412**
	Removed	0.6258	0.6110	0.5793	0.5833
	MissForest	0.3599	0.3970	0.4719	0.5399
	% deleted	8.92%	21.23%	50.56%	69.44%

TABLE 3 – Combined error metric (Equation 3) for retained (reliably imputed), removed (unreliably imputed), and MissForest instances.

7 Conclusion

Dans cet article, nous introduisons un nouveau cadre pour l'imputation des valeurs manquantes qui intègre l'algorithme missForest avec l'explicabilité basée sur SHAP pour évaluer explicitement la fiabilité de chaque valeur imputée. Contrairement aux méthodes d'imputation standard qui remplissent systématiquement toutes les entrées manquantes, notre approche distingue entre imputations fiables et non fiables en analysant si les prédictions sont soutenues par des caractéristiques suffisamment informatives et non manquantes.

À travers des expériences approfondies sur cinq jeux de données de référence (Adult, Credit, Ecoli, Heart Disease et Parkinson) sous différents taux de valeurs manquantes (15%, 25%, 50% et 75%), nous avons démontré que notre méthode atteint systématiquement une erreur d'imputation plus faible sur les lignes imputées de manière fiable comparée aux techniques classiques, incluant missForest standard, MICE, KNN et les imputations statistiques simples. Notamment, notre méthode a obtenu les meilleurs scores NRMSE dans 18 configurations expérimentales sur 20 pour les variables numériques. Les résultats montrent que les lignes identifiées comme non fiables par notre critère basé sur SHAP présentent effectivement une erreur d'imputation significativement plus élevée, validant l'efficacité de notre mécanisme de détection de fiabilité.

De plus, notre approche présente une robustesse améliorée à mesure que la proportion de valeurs manquantes augmente en prévenant la propagation d'erreurs qui survient couramment lorsque des imputations non fiables sont réutilisées pour prédire d'autres entrées manquantes. En combinant sélectivement imputation et suppression, notre méthode aborde une lacune importante mise en évidence dans la littérature et contribue à des pipelines de prétraitement de données plus fiables. Au-delà de l'amélioration de la précision de l'imputation, notre travail démontre la valeur plus large de l'intégration de l'explicabilité dans le prétraitement des données. L'évaluation de fiabilité basée sur SHAP améliore non seulement la qualité mais fournit également des justifications interprétables pour déterminer quelles instances sont conservées ou écartées, améliorant la transparence dans le processus de préparation des données. Ce travail met finalement en évidence l'importance d'incorporer l'explicabilité dans la gestion des données manquantes pour assurer à la fois la précision et l'interprétabilité dans les analyses en aval.

8 Travaux Futurs

Une limitation de l'approche proposée réside dans le choix partiellement arbitraire de certains paramètres, en particulier le seuil de fiabilité utilisé pour distinguer les imputations fiables des non fiables. Dans les travaux futurs, cette décision binaire pourrait être remplacée par un score de fiabilité continu, permettant de classer les lignes de la plus fiable à la moins fiable. Une telle formulation offrirait une plus grande flexibilité, permettant aux utilisateurs soit d'écarter une proportion contrôlée d'observations à faible fiabilité, soit d'identifier un seuil optimal basé sur l'évolution de l'erreur de généralisation. Cela faciliterait également l'analyse du compromis entre performance prédictive et rejet de données à travers des courbes erreur-rejet, offrant une évaluation plus nuancée de l'impact de la fiabilité de l'imputation.

De plus, nous avons l'intention d'étudier l'impact de notre approche sur des problèmes du monde réel et de comparer son efficacité par rapport à d'autres techniques d'imputation existantes. Cela fournira des perspectives pratiques sur la robustesse et l'applicabilité de notre méthode dans divers contextes de données.

Références

[1] Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. DOI : https://doi.org/10.24432/C5PC7J.

[2] Geeta Chhabra, Vasudha Vashisht, and Jayanthi Ranjan. Missing value imputation using hybrid k-means and association rules. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2018.

[3] Alireza Farhangfar, Lukasz Kurgan, and Jennifer Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12) :3692–3705, 2008.

[4] Menna Ibrahim Gabr, Yehia Mostafa Helmy, and Doaa Saad Elzanfaly. Effect of missing data types and imputation methods on supervised classifiers : An evaluation study. *Big Data and Cognitive Computing*, 7(1) :55, 2023.

[5] Pegah Golchian and Marvin N Wright. Imputation uncertainty in interpretable machine learning methods. *arXiv preprint arXiv :2512.17689*, 2025.

[6] Falaah Arif Khan, Denys Herasymuk, Nazar Protsiv, and Julia Stoyanovich. Still more shades of null : An evaluation suite for responsible missing value imputation. *arXiv preprint arXiv :2409.07510*, 2024.

[7] Kamakshi Lakshminarayan, Steven A Harp, Robert P Goldman, and Tariq Samad. Imputation of missing data using machine learning techniques. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[8] Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation : a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53 :1487–1509, 2020.

[9] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv :1802.03888*, 2018.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] Lijuan Ren, Tao Wang, Aicha Sekhari Seklouli, Haiqing Zhang, and Abdelaziz Bouras. A review on missing values for main challenges and methods. *Information Systems*, 119 :102268, 2023.

[12] Donald B Rubin. *Multiple imputation for survey nonresponse*. Wiley, New York, 1987.

[13] Ismail Setiawan, Rahmat Gernowo, and Budi Warsito. A systematic literature review on missing values : research trends, datasets, methods and frameworks. In *E3S Web of Conferences*, volume 448, page 02020. EDP Sciences, 2023.

[14] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112–118, 2012.

[15] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6) :520–525, 2001.

[16] Stef Van Buuren. *Flexible imputation of missing data*. CRC Press, Boca Raton, FL, 2nd edition, 2012.

[17] Stef Van Buuren and Karin Oudshoorn. *Flexible multivariate imputation by MICE*. TNO, Leiden, 1999.

[18] Tuan L Vo, Thu Nguyen, Luis M Lopez-Ramos, Hugo L Hammer, Michael A Riegler, and Pål Halvorsen. Explainability of machine learning models under missing data. *arXiv preprint arXiv :2407.00411*, 2024.