The Scientific Temper (2025) Vol. 16 (2): 3727-3735

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.05

https://scientifictemper.com/

RESEARCH ARTICLE

Enhancing data imputation in complex datasets using lagrange polynomial interpolation and hot-deck fusion

Amala Deepa V.* and T. Lucia Agnes Beena

Abstract

Data imputation is vital in preserving the quality of datasets in machine learning, where missing data leads to decreased model accuracy. This research proposes a new imputation method called Lagrange Polynomial Interpolation with Hot-Deck Fusion (LPIHD) to enhance the quality and reliability of imputed datasets, mainly when the data is multifaceted and comprises multiple types. LPIHD combines Lagrange Polynomial Interpolation and Hot-Deck Fusion. Lagrange Polynomial Interpolation estimates missing values using known data points. Hot-Deck Fusion refines these estimates by borrowing similar values from a donor population. This hybrid approach, applied to two distinct datasets about wine quality and heart diseases, enhances precision by achieving lower MAE and RMSE values than those previously recorded. LPIHD achieved better accuracy for the wine quality and heart disease datasets, respectively, at varying rates of missing data. MAE and RMSE were also notably reduced across both datasets, affirming the method's efficacy. These findings suggest that LPIHD can produce better and more accurate data imputations, making it a helpful technique for the field that needs a strong analytical platform.

Keywords: Data Imputation, Hot-Deck Fusion, Hybrid Methods, Lagrange Polynomial Interpolation, Machine Learning.

Introduction

The imputation of missing data is playing an important role in data science. and, more specifically, in machine learning because data plays a central role in learning the algorithm (Ahmad et al., 2024). Data can be missing due to one of the reasons, such as human mistakes, machine errors, or lack of data collection, and these types of inadequacies significantly threaten the study's credibility (Albahri et al., 2023). If these gaps are not eliminated, the models derived from such research may contain inherent bias or errors. It potentially leads to incorrect conclusions and subsequent erroneous decisions. Therefore, imputation is not just a fix but an enrichment asset that strengthens statistics and artificial intelligence forecasts by creating approximations of missing values while maintaining the structure and connection of the data.

Also, data imputation methods have implications for the quality of the models used in predictive analytics (Shadbahr et al., 2023). Incomplete data can positively impact machine training since it creates overfitting or underfitting models when applied in the real world. By doing the imputation correctly, data scientists can preserve the data's quality and guarantee that the models derived therefrom are precise and usable on other related tasks. It aids in optimizing the usage of current information within a business to enhance understanding made in giving out conclusions for pertinent analytics within the organization. The imputation techniques, therefore, stand as the foundation of high-quality data analysis and enhancement of the reliable machine learning model that is crucial in various industries to date.

Problem Statement

Current data imputation methods could be more decisive in handling large datasets with non-linear and heterogeneous values. Other techniques like mean imputation or regression analysis might ensure that the complexity of data details is manageable while hindering the complexity of the linked

Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620002, Tamil Nadu, India.

***Corresponding Author:** Amala Deepa V., Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620002, Tamil Nadu, India., E-Mail: lindseyamala@gmail.com

How to cite this article: Deepa, A. V., Beena, L.A. (2025). Enhancing data imputation in complex datasets using lagrange polynomial interpolation and hot-deck fusion. The Scientific Temper, **16**(2):3727-3735.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.2.05

Source of support: Nil

Conflict of interest: None.

Received: 15/01/2025 **Accepted:** 27/02/2025 **Published** : 20/03/2025

variables. This issue is even more so when analyzing datasets with a high level of heterogeneity, and the data types include nominal, ordinal, and interval levels of measurements. Such methods may bring a specific bias or mask variability that is initially present in data, which can result in reduced accuracy of analytics and even untruthful results in the field of predictive analytics. Therefore, it is growing important to develop sophisticated imputation mechanisms due to the increased complexity and heterogeneity of contemporary data sources while at the same time ensuring that the data analysis results are reliable and accurate.

Research Objectives

The primary objective of this research is to develop a hybrid imputation method that significantly enhances the accuracy and robustness of data handling, particularly in datasets plagued by missing values. This novel approach seeks to amalgamate the strengths of various imputation techniques to create a more reliable and efficient method for dealing with incomplete data across diverse scenarios. Key aims include:

- To integrate advanced mathematical models to improve the precision of imputed values, thereby increasing the overall accuracy of subsequent data analyses and machine learning models.
- To design the method to be resilient across various types of data, including categorical and ordinal, ensuring consistent performance regardless of the dataset's complexity.
- To enable the imputation method to dynamically adjust to the specific characteristics of the dataset, such as the distribution of missing data and the presence of non-linear relationships among features.

Related Works

Data imputation techniques in healthcare were examined, emphasizing challenges and ethical concerns related to complex health data. Both traditional and AI-driven methods were explored, demonstrating their effectiveness through real-world examples (Nayak & Khilar, 2024). Lagrange interpolation for volatile datasets was analyzed, with findings indicating that cubic interpolation was most effective, especially in IoT systems (Oktaviani, Abdurohman, & Erfianto, 2023). The Cyclical Tree-Based Hot Deck (CTBHD) method was introduced for complex survey data, enhancing stability and reducing bias through extensive customization and a cyclical approach (Sukasih & Scott, 2023). A new robust imputation algorithm, imputeRobust, was developed to improve the precision and reliability of large-scale data analyses by effectively managing outliers and missing data (Templ, 2023).

Various imputation methods within S&P 500 financial datasets were compared, with MissForest identified as superior in enhancing predictive accuracy (Zamri et al., 2024). A probabilistic model for imputing data in employee datasets was presented, demonstrating high accuracy in diverse applications, from Kaggle competitions to real-world settings (Arefin & Masum, 2024). A new approach to handling missing data in accelerometer-based studies was developed, using hot deck multiple imputation to achieve less bias and better confidence interval coverage (Butera et al., 2019). Imputation for small and structured datasets was enhanced with a neural network-based architecture that uses adversarial learning to estimate uncertainty, improving traditional imputation techniques (Hameed & Ali, 2022).

A survey of imputation techniques offered insights into their effectiveness and limitations for managing large datasets, guiding future studies in quantitative research (Hameed & Ali, 2023). Challenges of multivariate polynomial interpolation were tackled with the novel Random Lagrange Multivariate Polynomial Interpolation Algorithm (RLMVPIA), enhancing computational efficiency (Essanhaji & Errachid, 2022). A multi-feature generation network for industrial time-series data was introduced, significantly improving data imputation accuracy (Zheng et al., 2023). Missing data imputation was improved by integrating high-order polynomial equations with CNNs, achieving superior accuracy on UCI datasets, and maintaining data integrity (Khan et al., 2024).

Methodology

Figure 1 illustrates the structured workflow of the Lagrange Polynomial Interpolation with Hot-Deck Fusion (LPIHD) methodology. The process begins with the raw input data, where missing values are identified across each feature. In Phase 1, Lagrange Polynomial Interpolation is applied to these identified missing points, leveraging observed neighboring values to estimate the missing data. Once these initial estimates are generated, the workflow transitions to Phase 2, where Hot-Deck Imputation is used for further refinement. This phase involves selecting similar donor records from the complete cases in the dataset to adjust or replace the estimates from Phase 1. The integration of results from both phases leads to the final imputed dataset, which is then analyzed using machine learning models such as Naive Bayes (NB) and Multi-Layer Perceptron (MLP) to validate the imputation's effectiveness. This cohesive process ensures that the final dataset is robust and ready for advanced analytical applications.

This work introduced artificial missingness into the datasets to evaluate the performance of the imputation methods. This approach allowed for controlled experimental conditions to assess how well each imputation method recovers lost information. Missingness was induced at three different rates: 10%, 20%, and 30%, representing varying levels of data sparsity that might be encountered in real-world scenarios.

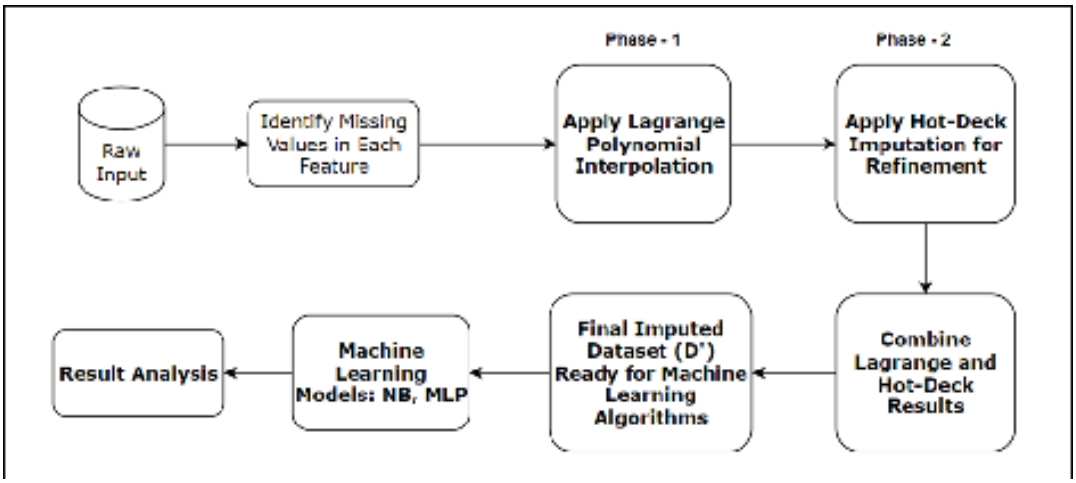

Figure 1: LPIHD Workflow

The process of creating artificial missingness involves randomly selecting a specified percentage of data points within the dataset and systematically removing their values. This is mathematically represented by the equation (1):

$$M(x) = \begin{cases} NaN & with\ probability\ p \\ x\ with\ probability\ (1-p) \end{cases} \quad (1)$$

where M(x) denotes the potentially missing data point, x is the original data value, NaN represents a missing value and p is the probability of a data point being missing, set to $0.1, 0.2, or\ 0.3$ depending on the desired missingness level.

Example

Consider a small dataset for demonstration: $[4, 8, 15, 16, 23, 42]$. Applying a 20% artificial missingness rate, each data point has a 20% chance of being replaced with NaN. A possible outcome might be $[4, NaN, 15, 16, 23, 42]$, indicating that the second position in the dataset was selected to be missing under the induced conditions.

Lagrange Polynomial Interpolation

Lagrange Polynomial Interpolation (LPI) is a classic mathematical method that estimates the values of a function at specific points by leveraging its known values at other points. This technique is particularly relevant in the realm of data imputation, where it facilitates the estimation of missing values through the utilization of known data points. The general interpolation formula is given in equation (2):

$$P(x) = \sum_{j=0}^{k} y_j L_j(x) \quad (2)$$

Where:

- $P(x)$ is the polynomial that estimates missing values.
- y_j are the known data values.
- $L_j(x)$ are the Lagrange basis polynomials.

Each basis polynomial $L_j(x)$ is defined by the product as in the equation (3):

$$L_j(x) = \prod 0 \le m \le k \ \ m \ne j \ \ \frac{x - x_m}{x_j - x_m} \quad (3)$$

This method is exceptionally effective in datasets where relationships between variables are non-linear, as the polynomial can flexibly fit a wide range of data patterns. The derivation of the Lagrange polynomial involves creating a series of basis polynomials, each corresponding to one of the known data points. These polynomials are designed such that each one equals 1 at its corresponding data point and 0 at all other data points included in the interpolation.

Consider a simple dataset with three known data points: $(x_0, y_0) = (1,2), (x_1, y_1) = (3,6), (x_2, y_2) = (4,8)$. To estimate the value of the function at x = 2, the Lagrange interpolation formula would be applied as follows:

The basis polynomials would be computed using the equation (3):

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x-3)(x-4)}{(1-3)(1-4)} = \frac{(x-3)(x-4)}{6}$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x-1)(x-4)}{(3-1)(3-4)} = \frac{(x-1)(x-4)}{-2}$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x-1)(x-3)}{(4-1)(4-3)} = \frac{(x-1)(x-3)}{3}$$

Then, the Lagrange polynomial P(x) would be assembled by combining these basis polynomials with the known y values using the equation (2):

$$P(x) = 2 \cdot \frac{(x-3)(x-4)}{6} + 6 \cdot \frac{(x-1)(x-4)}{-2} + 8 \cdot \frac{(x-1)(x-3)}{3}$$

Calculating P(2) results in:

$$P(2) = 2 \cdot \frac{(2-3)(2-4)}{6} + 6 \cdot \frac{(2-1)(2-4)}{-2} + 8 \cdot \frac{(2-1)(2-3)}{3}$$

$$P(2) = 2 \cdot \frac{(-1)(-2)}{6} + 6 \cdot \frac{(1)(-2)}{-2} + 8 \cdot \frac{(1)(-1)}{3}$$

$$P(2) = 2 \cdot \frac{2}{6} - 6 + 8 \cdot \frac{-1}{3}$$

$$P(2) = \frac{2}{3} - 6 - \frac{8}{3} = \frac{2-18-8}{3} = \frac{-24}{3} = -8$$

Lagrange Polynomial Interpolation's strength lies in its ability to precisely model the intricate relationships inherent in real-world data. This capability makes it a valuable tool in fields such as climatology, economics and any other area where prediction models based on historical data are used. Its adaptability ensures that the interpolations are both accurate and practical for data-driven decision-making processes, thereby maintaining the integrity and reliability of statistical analyses and predictive modeling.

Hot-Deck Fusion

Hot-Deck Fusion (HDF) is a sophisticated method used in data imputation to handle missing values by drawing upon a pool of donors—records within the dataset that have complete data. Unlike other imputation techniques that rely on statistical or model-based assumptions, HDF utilizes actual data to ensure the imputed values are realistic and consistent with observed data patterns. This method is particularly effective in maintaining the integrity of categorical and ordinal data, as well as in datasets where preserving the distribution of the data is crucial.

Hot-Deck Fusion operates by identifying 'donor' records that are like records with missing data. For each record or data point with a missing value, a donor is selected from the pool of complete records based on specific criteria, such as proximity in statistical or demographic characteristics. The missing value is then replaced with the value from the selected donor. This process can be formally represented as in the equation (4):

$$I(m) = D(i) \quad (4)$$

Where:

- $I(\ )$ t is the imputed value for the missing data point·
- $D(i)$ is the donor value selected based on the closeness to the characteristics of the missing data point.

The selection of donors is a critical step in HDF. It typically involves calculating a similarity index or distance measure between the incomplete record and each potential donor record. The record with the smallest distance or highest similarity score is selected as the donor. The distance can be computed using the equation (5):

$$S_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \quad (5)$$

where:

- S_{ij} is the similarity or distance between the incomplete record i and donor record j.
- x_{ik}, x_{jk} are the values of the k-th attribute for records i and j, respectively.
- n is the number of attributes considered for determining similarity.

For categorical data, a common approach involves using a matching algorithm that counts the number of attributes identical between two records, given in the equation (6).

$$M_{ij} = \sum_{k=1}^{n} \ddot{a}\left(x_{ik}, x_{jk}\right) \quad (6)$$

where δ is an indicator function that returns 1 if $x_{ik} = x_{jk}$ and 0 otherwise.

Once a donor is selected, the missing value is replaced directly with the corresponding value from the donor record. This approach can be extended to multivariate missing data by conducting a donor selection for each missing component individually or by finding a single donor for all missing components. The process of replacing the missing value is expressed in the equation (7):

$$X_{miss} = X_{donor} \quad (7)$$

where:

- X_{miss} represents the vector of missing values.
- X_{donor} represents the vector of values from the selected donor that corresponds to the missing components.

Consider a dataset where a record is missing values for attributes A and B. Assume the dataset has three complete records as potential donors. The HDF process would involve calculating the similarity or distance from the missing record to each of the donor records using a chosen metric, selecting the donor with the highest similarity or lowest distance and then replacing the missing values in attributes A and B with those from the selected donor.

This process ensures that the imputed values are realistic and maintain the original data distribution, thus minimizing the introduction of bias that can often occur with model-based imputation techniques.

Hot-Deck Fusion is particularly useful in settings where the accuracy of categorical and ordinal data imputation is

crucial. It is widely applied in survey data analysis, clinical data management, and any field where data integrity and accuracy are paramount. By relying on actual data points rather than estimated or modeled values, HDF provides a practical and reliable method for data imputation, enhancing the quality of data analyses and the reliability of subsequent conclusions drawn from the data.

Integration of Techniques

The integration of LPIHD starts by applying the Lagrange Polynomial to each missing entry in the dataset, providing an initial estimate. Following this, the HDF method examines these initial estimates: if an estimate closely matches a donor value from the pool, it is retained; otherwise, it is adjusted based on the most similar donor to ensure the imputed values are realistic and consistent with the dataset's overall characteristics.

This dual approach ensures a robust imputation process. LPI ensures mathematical precision in estimating missing values based on observable data trends, while HDF adjusts these estimates to reflect the dataset's real-world complexity and diversity. The combination offers a comprehensive solution to data imputation challenges, particularly in complex scenarios where single-method approaches may fall short.

Algorithm-1: Artificial Missingness Induction

Input: Original Dataset D, Missingness Rates p = [0.1, 0.2, 0.3]
Output: Dataset D′ with induced missing values

1. for each rate p_i in p
2. Copy original dataset $D\,to\,D'$
3. for each element $d \in D'$
4. Generate random number r from uniform distribution U(0,1)
5. if $r < p_i$:
6. Set $d = \mathrm{NaN}$ (mark as missing)
7. Return dataset D′ with missing values induced as per rate p_i

The algorithmic framework for the LPIHD is structured to efficiently handle missing data in datasets, particularly where the data exhibits non-linear relationships and involves various data types. The following detailed description outlines each step of the process, from initialization to evaluation.

Algorithm 2: LPIHD

Input: Dataset D with missing values
Output: Imputed Dataset D′

1. Initialize: $D' \leftarrow D$
2. for each $X[i] \in D'$ with missing:
3. $\mathrm{Obs} \leftarrow observeddata \in X[i]$
4. $\mathrm{Miss} \leftarrow missingpositions \in X[i]$
5. for $m \in \mathrm{Miss}$:
6. $N_m \leftarrow nearest\ neighbors\ of\ min\,\mathrm{Obs}$
7. $$P(m) \leftarrow \sum_{j=0}^{k} y_j \prod_{0 \le m \le km \ne j} \frac{m - x_m}{x_j - x_m}$$
8. Replace m in X[i] with P(m)
9. for $m \in \mathrm{Miss}$:
10. $\mathrm{DonorPool} \leftarrow complete\,cases\ of\ D$
11. $d \leftarrow select\ donor\ from\,\mathrm{DonorPool}$
12. $if\ |d - P(m)|\ is\ small$:
13. $X[i][m] \leftarrow P(m)$
14. else:
15. $X[i][m] \leftarrow d$
16. Return D′

Notations and Symbols Description:

- D: Original dataset
- D′: Imputed dataset
- X[i]: i-th column of dataset D
- Obs : Set of observed (non-missing) data points in X[i]
- Miss : Indices of missing data points in X[i]
- m: Index of a missing data point in X[i]
- N_m : Nearest observed data points to m
- P(m): Imputed value at index m calculated using Lagrange Polynomial Interpolation
- y_j : Observed value corresponding to neighbor x_j
- x_j, x_m : Indices of the observed and missing data points used in the interpolation formula
- $\mathrm{DonorPool}$: Set of complete cases from D used for selecting donor values
- d: Selected donor value from $\mathrm{DonorPool}$
- $|d - P(m)|$: Absolute difference between donor value and interpolated value

Results

Dataset Description

The datasets utilized in this study were selected for their distinct characteristics: the Wine Quality dataset represents a complex, non-linear dataset, while the heart disease dataset exemplifies large-scale, highly imbalanced data. Table 1 summarizes the key details of these datasets.

Table 1: Datasets Overview

Dataset	*Source*	*Attributes*	*Instances*	*Characteristics*
Wine Quality	UCI Repository	Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality (score between 0 and 10).	4,898	Complex and non-linear relationships among features.
Heart Disease	Kaggle	HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer.	319,400	Large-scale dataset with significant class imbalance, requiring robust handling of minority classes.

Wine Quality Results

The results for the wine quality dataset using the LPIHD method show significant improvements across various metrics when compared to the HPCNN (Khan et al., 2024). As detailed in Table 2 and Figure 2 the LPIHD method outperformed HPCNN in accuracy measurements across all percentages of missing data. Specifically, using the Multi-Layer Perceptron (MLP) model, LPIHD achieved a peak accuracy of 75.4% at 10% missing data a substantial increase from the 51.9% observed with HPCNN. This improvement underscores the capability of LPIHD to enhance predictive accuracy effectively, even under conditions of significant data missingness.

Furthermore, Table 3 and Figure 3 illustrates the comparative analysis of Mean Absolute Error (MAE) metric. LPIHD consistently showed lower MAE values compared to HPCNN, indicating that the imputed values deviate less from the actual values, and thus are more accurate. For instance, at 10% missing data, the MAE with LPIHD using the MLP model were reduced to 0.246, demonstrating a significant decrease from the values recorded by HPCNN.

Table 2: Comparative results of accuracy for wine quality dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.427	0.698	0.431	0.69	0.429	0.673
MLP	0.519	0.754	0.514	0.753	0.504	0.713

Table 3: Comparative results of MAE for wine quality dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.695	0.302	0.679	0.31	0.696	0.327
MLP	0.534	0.246	0.545	0.247	0.557	0.287

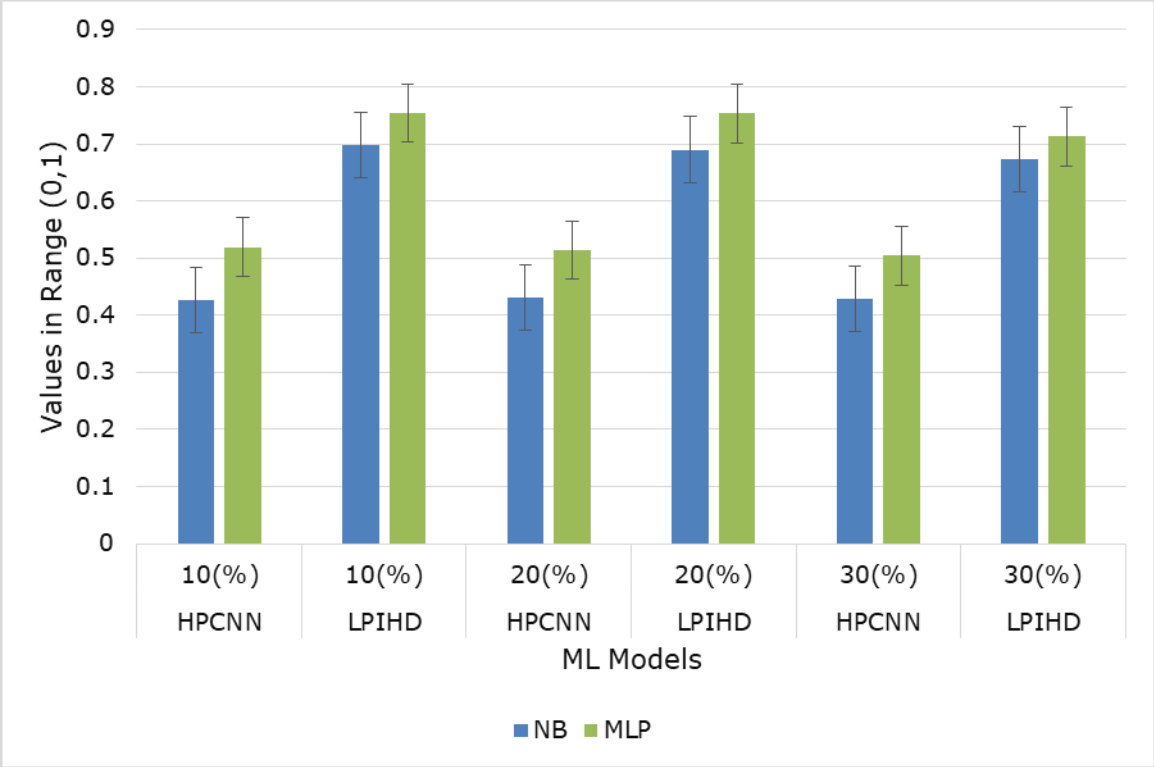

Figure 2: Comparative results of accuracy for wine quality dataset

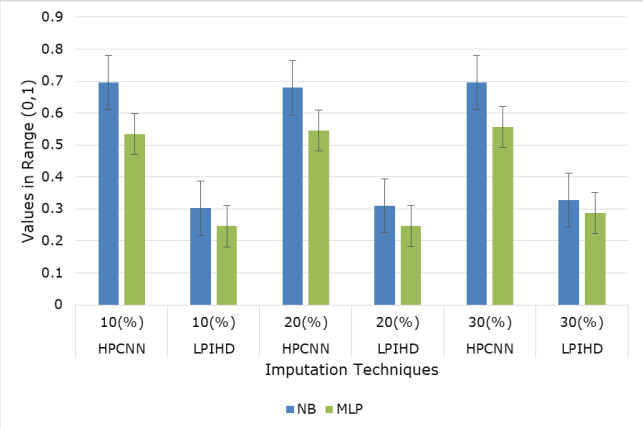

Figure 3: Comparative results of MAE for wine quality dataset

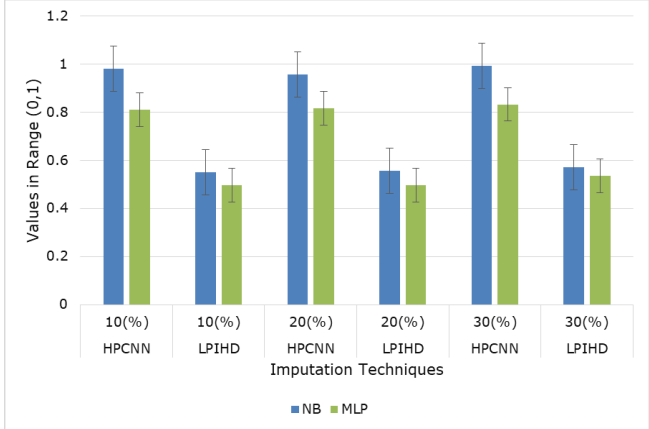

Figure 4: Comparative results of RMSE for wine quality dataset

Table 4 and Figure 4 illustrates the comparative analysis of Root Mean Square Error (RMSE) for wine quality dataset. LPIHD steadily showed lower RMSE values compared to HPCNN, indicating that the imputed values deviate less from the actual values, and thus are more accurate. For instance, at 10% missing data, the RMSE with LPIHD using the MLP model were reduced to 0.496, representing a substantial decline from the values recorded by HPCNN.

Heart Disease Results

As shown in Table 5 and Figure 5 LPIHD enhanced accuracy notably, especially for the MLP model, achieving 90.4% accuracy at 20% missing data level for heart disease dataset. This represents a significant increase compared to the HPCNN method, which peaked at 87.4% under similar conditions. The Naive Bayes (NB) model also saw improved accuracy with LPIHD, reaching up to 78.9% at 30% missing data.

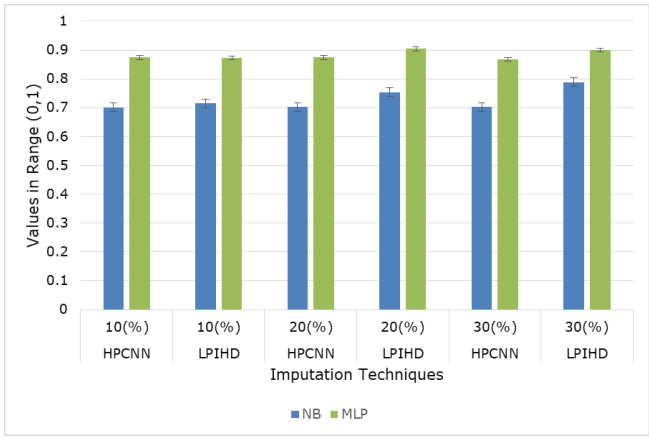

Figure 5: Accuracy Results for Heart Disease Dataset

Table 4: Comparative results of RMSE for wine quality dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.982	0.55	0.958	0.557	0.994	0.571
MLP	0.81	0.496	0.817	0.497	0.833	0.535

Table 5: Accuracy Results for Heart Disease Dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.701	0.715	0.703	0.754	0.702	0.789
MLP	0.874	0.873	0.874	0.904	0.867	0.901

Table 6: MAE Results for Heart Disease Dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.299	0.285	0.297	0.246	0.298	0.211
MLP	0.126	0.127	0.126	0.096	0.133	0.099

Table 6 and Figure 6 details the MAE for the heart disease dataset, illustrating that LPIHD method significantly reduced MAE across all levels of missing data. For instance, with the MLP model under LPIHD, the MAE decreased to 0.096 at 20% missing data from 0.126 recorded by HPCNN, indicating a more precise imputation of missing values. Even the NB model saw a reduction in MAE from 0.299 with HPCNN to 0.285 with LPIHD at 10% missing data.

In terms of RMSE, Table 7 and Figure 7 reflects similar improvements brought about by the LPIHD method. The RMSE for the MLP model decreased notably from 0.355 with HPCNN to 0.310 with LPIHD at 20% missing data. These results further validate the effectiveness of the LPIHD method in reducing error rates and enhancing the reliability of data imputation, particularly in complex medical datasets.

Discussion

The results obtained from the implementation of the LPIHD method have been promising across various metrics

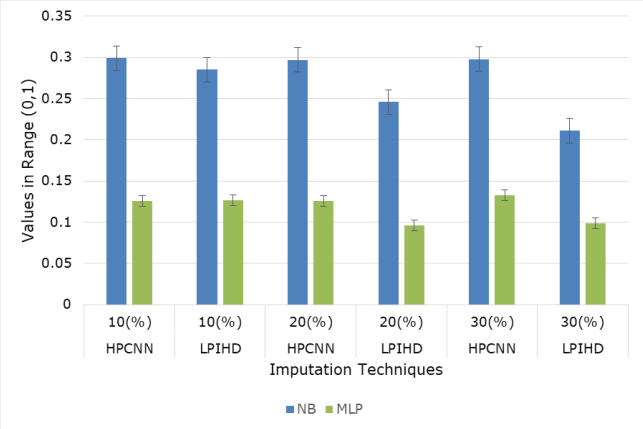

Figure 6: MAE Results for Heart Disease Dataset

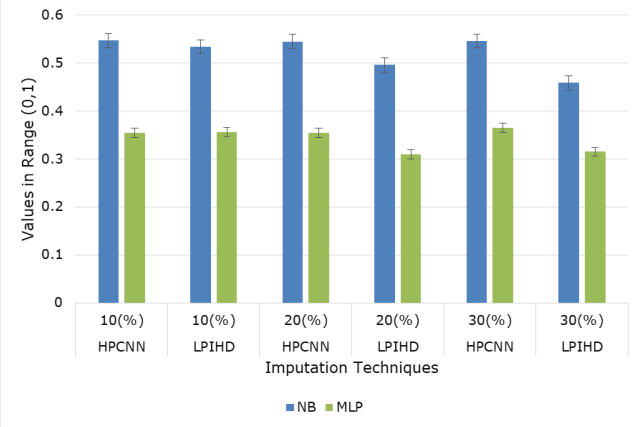

Figure 7: RMSE Results for Heart Disease Dataset

Table 7: RMSE Results for Heart Disease Dataset

ML Models	*HPCNN 10(%)*	*LPIHD*	*HPCNN 20(%)*	*LPIHD*	*HPCNN 30(%)*	*LPIHD*
NB	0.547	0.534	0.545	0.496	0.546	0.459
MLP	0.355	0.356	0.355	0.31	0.365	0.315

including accuracy, MAE and RMSE as shown in Tables 2 to 7. The LPIHD method consistently outperformed previous HPCNN method in accuracy metrics. For instance, in the wine quality dataset, accuracy improvements were evident with LPIHD achieving a peak accuracy of 75.4% at 10% missing data rate using an MLP model, a substantial increase from the 51.9% observed with HPCNN. Such enhancements are attributable to the method's capability to integrate polynomial interpolation and hot-deck imputation, thereby tailoring imputation more closely to the underlying data structures and patterns.

Furthermore, LPIHD demonstrated significant reductions in MAE and RMSE across both datasets. These metrics are critical as they indicate a closer match between the imputed and actual values, essential for maintaining the integrity and utility of the dataset in subsequent analyses. Notably, the heart disease dataset saw MAE improvements, with LPIHD reducing the MAE to 0.096 at 20% missing data, compared to 0.126 with HPCNN. This precision is particularly beneficial in healthcare datasets where accurate data representation is crucial for patient diagnosis and treatment planning.

Despite its advantages, LPIHD's implementation is not devoid of challenges. The complexity of integrating two distinct imputation methods demands careful tuning and validation to ensure optimal performance across different types of datasets. The method's dependency on the quality and arrangement of available data for polynomial interpolation and the selection of appropriate donors for hot-deck imputation could limit its applicability in extremely sparse or irregular datasets. Additionally, the computational overhead involved in executing two sequential imputation phases may impact its scalability and efficiency in larger datasets.

Conclusion

Data imputation plays a critical role in ensuring the accuracy and integrity of datasets used in machine learning, where missing data can significantly impair the performance and reliability of predictive models. This study introduced a novel hybrid imputation method, LPIHD, aimed at enhancing the accuracy and robustness of imputed data, particularly in complex datasets with multiple data types. The method combines Lagrange Polynomial Interpolation, which utilizes known data points to estimate missing values, with Hot-Deck Fusion, where these estimates are refined using similar values from a donor pool. Applied to two distinct datasets—wine quality and heart disease—LPIHD demonstrated significant improvements. Specifically, it achieved accuracy increases up to 75.4% and 90.1%, while reducing MAE to 0.246 and RMSE to 0.310 at varying rates of missing data for the respective datasets. Despite its effectiveness, LPIHD's computational demands and reliance on the availability of appropriate donor data present limitations, particularly in sparsely populated or highly irregular datasets. Future work will focus on enhancing the computational efficiency of LPIHD and expanding its application to real-time data streaming environments, aiming to broaden its utility across more dynamic and diverse data scenarios. These efforts seek to establish LPIHD as a foundational tool for reliable data imputation in critical analytical applications.

Acknowledgements

The author extends gratitude to the academic mentors and colleagues at the department whose insights and expertise greatly contributed to the research, as well as the financial support from the university which facilitated this project.

References

Ahmad, A. F., Alshammari, K., Ahmed, I., & Sayed, M. D. (2024). Machine Learning for Missing Value Imputation. arXiv preprint arXiv:2410.08308. https://doi.org/10.48550/

arXiv.2410.08308

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion, 96, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Arefin, M. N., & Masum, A. K. M. (2024). A Probabilistic Approach for Missing Data Imputation. *Complexity*, *2024*(1), 4737963. https://doi.org/10.1155/2024/4737963

Butera, N. M., Li, S., Evenson, K. R., Di, C., Buchner, D. M., LaMonte, M. J., ... & Herring, A. (2019). Hot deck multiple imputation for handling missing accelerometer data. *Statistics in biosciences*, *11*, 422-448. https://doi.org/10.1007/s12561-018-9225-4

Essanhaji, A., & Errachid, M. (2022). Lagrange multivariate polynomial interpolation: a random algorithmic approach. *Journal of Applied Mathematics*, *2022*(1), 8227086. https://doi.org/10.1155/2022/8227086

Hameed, W. M., & Ali, N. A. (2022). Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning. *Periodicals of Engineering and Natural Sciences (PEN)*, *10*(3), 350-367.

Hameed, W. M., & Ali, N. A. (2023). Missing value imputation techniques: a survey. *UHD Journal of Science and Technology*, *7*(1), 72-81. https://doi.org/10.21928/uhdjst.v7n1y2023.pp72-81

Khan, H., Rasheed, M. T., Liu, H., & Zhang, S. (2024). High-order polynomial interpolation with CNN: A robust approach for missing data imputation. *Computers and Electrical Engineering*, *119*, 109524. https://doi.org/10.1016/j.compeleceng.2024.109524

Lv, Z., Chen, K., Zhang, T., Zhao, J., & Wang, W. (2023). Multi-feature generation network-based imputation method for industrial data with high missing rate. *Expert Systems with Applications*, *227*, 120229. https://doi.org/10.1016/j.eswa.2023.120229

Nayak, S., & Khilar, P. M. (2024). Data Imputation in Healthcare Applications. In *AI Healthcare Applications and Security, Ethical and Legal Considerations* (pp. 49-67). IGI Global. DOI: 10.4018/979-8-3693-7452-8.ch004

Oktaviani, I. D., Abdurohman, M., & Erfianto, B. (2023). Fluctuating Small Data Imputation with Lagrange Interpolation Based. In *Information Systems for Intelligent Systems: Proceedings of ISBM 2022* (pp. 211-217). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7447-2_19

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., ... & Schönlieb, C. B. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. Communications Medicine, 3(1), 139. https://doi.org/10.1038/s43856-023-00356-z

Sukasih, A. S., & Scott, V. (2023). *Cyclical Tree-Based Hot Deck Imputation*. RTI Press.

Templ, M. (2023). Enhancing precision in large-scale data analysis: an innovative robust imputation algorithm for managing outliers and missing values. *Mathematics*, *11*(12), 2729. https://doi.org/10.3390/math11122729

Templ, M. (2023). Enhancing precision in large-scale data analysis: an innovative robust imputation algorithm for managing outliers and missing values. *Mathematics*, *11*(12), 2729. https://doi.org/10.3390/math11122729

Zamri, N. A., Jaya, M. I., Irawati, I. D., Rassem, T. H., & Kasim, S. (2024). Comparative Analysis of Imputation Methods for Enhancing Predictive Accuracy in Data Models. *JOIV: International Journal on Informatics Visualization*, *8*(3), 1271-1276. http://dx.doi.org/10.62527/joiv.8.3.1666