INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

A Hybrid Approach for Missing Data Imputation using Polynomial Interpolation and Backfill

V. Amala Deepa[1]*, **T. Lucia Agnes Beena**[2]

1 Research Scholar, Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, TamilNadu, India
2 Research Supervisor, Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, TamilNadu, India

OPEN ACCESS

Received: 06/04/2025

Accepted: 05/05/2025

Published: 21/05/2025

Citation: Amala Deepa V, Lucia Agnes Beena T (2025) A Hybrid Approach for Missing Data Imputation using Polynomial Interpolation and Backfill. Indian Journal of Science and Technology 18(19): 1478-1488. https://doi.org/10.17485/IJST/v18i19.636

* **Corresponding author**.

lindseyamala@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose a method for imputing missing values within non-linear datasets with class imbalance issues. **Methods:** The proposed methodology includes two stages of imputation through the Polynomial with Backfill (PoBa) process. This imputation technique applies quadratic polynomial interpolation for individual class type missing value estimation as well as implementing a backfill strategy to complete remaining gaps. The evaluation process utilized six ML models to assess dataset imputations under different non-linear and class imbalance conditions within two publicly available datasets. **Findings:** The PoBa method showed superior performance than HPCNN baseline in all the metric evaluations for different degrees of missingness. The Wine Quality dataset received an accuracy score of 0.807 and MAE of 0.193 and RMSE of 0.439 when feeding PoBa to Multi-Layer Perceptron (MLP) under 30% missing data conditions. The PoBa method reached 0.950 accuracy along with 0.050 MAE and 0.224 RMSE in the heart disease dataset despite its imbalanced characteristics. The experiments demonstrate that PoBa effectively preserves data organization and improves predictive performance while handling large real-world datasets. **Novelty:** The method proposes a class-specific combination of polynomial interpolation with backfill to handle missing data in an efficient and interpretable way.

Keywords: Missing Data; Polynomial Interpolation; Backfill Imputation; Machine Learning; Class Imbalance

1 Introduction

The application of machine learning techniques has revolutionized data-based decision processes throughout different industrial sectors[(1)]. Accurate prediction depends heavily on data quality at its source. Missing values creates persistent problems in data acquired from real-world environments that span healthcare along with finance and environmental sciences[(2)]-[(3)]. Introducing missing data results in two negative impacts on model performance and statistical power reduction while producing biased results.

The preprocessing of missing data requires strong techniques that preserve the original data structure.

Mean substitution along with median and mode substitution approach data imputation easily yet they do not understand specific data contexts[4]. Such techniques lack ability to recognize variable relationships which results in wrong estimation results. Monthly imputation methods require computationally complex systems together with domain-specific parameter optimization for effective execution. The requirement exists for developing strategies to merge both accuracy and simple execution techniques.

Multiple studies during recent times have examined new approaches to handling missing data. The medical datasets have been studied for deep learning-based imputation strategies by Xu *et al.*[5] and kose *et al.*[6]. The authors at Pastorini et al. developed hybrid solutions which integrate statistical approaches with machine learning tools to manage environmental data incompleteness[7]. These publications show both the rising awareness of handling missing data problems and the remaining shortcomings in current available solutions.

Research has recently concentrated on understanding the problems related to missing data. Healthcare predictions experience harmful consequences because of missing data according to Nijman et al.[8]. These researchers pointed out that data imputation methods must preserve existing relationships between data variables. The research conducted by Wang et al.[9] demonstrated successful implementation of clustering and regression framework which delivered excellent outcomes in financial data processing.

Oktaviani et al.[10] presented their findings about temporal polynomial interpolation as an accurate improvement strategy for tiny datasets including live data. The proposed quadratic interpolation produced superior outcomes than both simple regression and other approaches. The proposed method produced Mean Squared Error (MSE) results of 0.051 for 10% data loss along with 0.033 and 0.035 for 30% and 50% data loss while simple regression obtained MSE scores of 0.71, 0.59 and 0.44 for the respective levels of missing data. The demonstrated strength of quadratic interpolation shows its ability to precisely manage small datasets encountering different levels of missing data.

Las-Heras et al.[11] present MARS (Multivariate Adaptive Regression Splines) as their new method for handling missing data. The evaluation demonstrated that this method succeeded in completing RMSE and MAPE and MAE tests while processing environmental data in Madrid. Tests were conducted on the methodology with three distinct implementations that included all data first and then omitted outliers and finally used previous-month data only. Researchers need to investigate how the approach works with varied datasets because it demonstrates high success in specific applications. Khumukcham et al.[12] developed a framework which effectively handles missing data in mixed-type datasets using linear regression analysis for numbers and decision trees for categories. The methodology shows switch-typing resistance across diverse datasets and classification systems through its single approach to interpretable mixed-type imputation which surpasses conventional predictive approaches for accuracy and model extension abilities.

In a recent study, HPCNN which combines high-order polynomial equations alongside convolutional neural networks (CNNs) for resolving missing data imputation problems[13]. HPCNN represents the method which applies trained kernels to spatially arranged data matrices to optimize polynomial coefficients. Experimental data revealed outstanding results because the system accomplished above 95% data accuracy on wine quality and heart disease sets. This method proved successful in working with both extensively correlated information and data with minimal correlations. Computational requirements of using CNNs for optimization reduces the availability of HPCNN in resource-limited applications. The method lacks exploration of its capabilities to work with datasets that change at speed.

Deepa and Beena[14] established LPIHD through the unification of Lagrange Polynomial Interpolation and Hot-Deck Fusion. Testing on wine quality with heart disease datasets showed this method decreased MAE and RMSE values while increasing accuracy because of its ability to manage heterogeneities and while preserving missing value integrity.

Almeida et al. (2025)[15] developed a new framework which integrated neural networks with meta-learning and LSTM models to find replacements for missing univariate time series data. The model adopted meta-learning capabilities which enabled it to adjust for different temporal patterns to improve its generalization across multiple datasets. The integration of LSTM allowed the model to detect temporal patterns which resulted in improved accuracy of imputation predictions.

Research Gap

The development of existing methods in imputation has progressed rapidly but still fails to solve fundamental issues. Since traditional approaches lack background understanding they generate prediction estimates that become biased. Advanced deep learning applications for imputation need powerful computational systems that limit their use across different applications. Many imputation methods fail to recognize how unequal class distribution affects the prediction accuracy. The present research shows how numeric precision competitors with human understanding in solution models. High-accuracy methods prove difficult to use in healthcare applications since they generally don't provide transparent implementations. The scalability of data methods becomes complicated since they demand substantial parameter adjustments along with domain-specific modifications.

The field remains short on research about implementing methods which handle unbalanced classes during data preparation. The method of imputation applies to uniform approaches to all data points despite neglecting specific characteristics of main and minority groups. A lack of attention towards data imbalance poses risks because it produces faulty prediction results specifically in classification models. Most data imputation approaches focus solely on achieving numeric precision but fail to emphasize interpretability and scalability of their methods.

The proposed Polynomial with Backfill (PoBa) method fills the current gaps in the field. The combination of polynomial interpolation with backfill strategies creates a strong and interpretable solution for dealing with missing data through PoBa framework. The strategy utilizes quadratic functions to replace missing values by performing backfill completion at the same time. A dual-stage approach between the techniques provides both precise data restoration and preserves original data organization structures which benefit various types of data.

Research Contributions

Development of the PoBa Technique: The PoBa Technique uses degree 2 polynomial interpolation together with backfill algorithms for handling missing values. The dual-stage method delivers precise and entire data which maintains the connections among values during imputation procedures.

Class-Specific Imputation: Unlike traditional methods, PoBa considers class imbalance during imputation. This feature provides improved fairness in addition to minimizing classification bias through its operation.

2 Methodology

2.1 Dataset Description

The investigation used the Wine Quality dataset as a complex non-linear system alongside the heart disease dataset as its example of large-scale imbalanced data. The essential specifications of these datasets are explained within Table 1.

Table 1. Datasets Overview

Dataset	Source	Attributes	Instances	Characteristics
Wine Quality	UCI Repository [16]	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 [17]	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.

The research datasets demonstrate effective evaluation conditions for the PoBa technique. The Wine Quality dataset presents multiple complex feature relations which makes it difficult for imputation methods to keep non-linear patterns. A method with scalability and the ability to retain minority class information needs to be developed because the heart disease dataset presents a heavily unbalanced dataset. The unique data problems require a combination of approaches like PoBa to establish both maximal data integrity and reliable model performance.

2.2 PoBa Technique

The PoBa technique works as a powerful dataset completion tool which addresses problems with both non-linear data patterns and uneven class distribution. The approach contains two phases that function as depicted in Figure 1.

Phase 1: Class-Specific Polynomial Interpolation

At the first stage the data source separates into distinct groups of majority and minority. Each distinct class maintains its distinctive properties after separation to prevent introduction of biases during the imputation process. The algorithm operates independently on both classes to complete imputation by implementing polynomial interpolation methods. The utilization of quadratic polynomials offers efficient non-linear relationship modeling, so they become the preferred choice for this method.

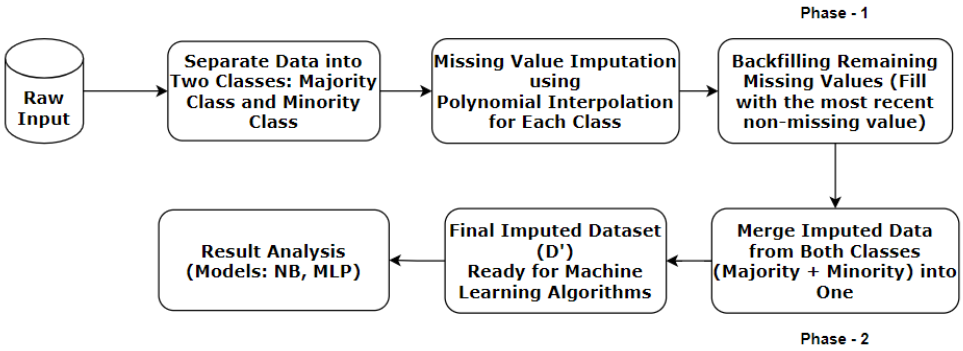

Fig 1. Workflow of the PoBa Technique

The interpolation function $f(x)$ represents the approach to fill missing values according to the following definition:

$$f(x) = \sum_{i=0}^{n} a_i x^i + \lambda \sum_{i=1}^{m} \left[\frac{\partial^2}{\partial x^2} (a_i x^i) \right]^2 \tag{1}$$

where a_i are polynomial coefficients are optimized for feature relationships, λ is a regularization parameter controlling the curve smoothness, n is the polynomial degree, and m represents the total number of features with missing values.

Through regularization we prevent overfitting which produces imputed data values that match the observed data patterns throughout the whole dataset. The Wine Quality non-linear dataset achieves accurate estimation through this approach.

Phase 2: Backfilling and Data Merging

Residual unobservable values need to be filled using backfill techniques after implementing polynomial interpolation. The backfill method propagates the last observed non-missing value forward. The merged imputed dataset D' maintains consistency between all data points through the combination of completed majority and minority class subsets. The combined data forms the following statement:

$$D' = \cup_{c \in \{\text{majority}, \text{minority}\}} (D_c - \{m_i\} \cup \hat{m_i}) \tag{2}$$

where D_c represents class-specific datasets, m_i are original missing values and $\hat{m_i}$ are their imputed counterparts.

2.3 PoBa Algorithm

The PoBa algorithm is a hybrid imputation technique that combines polynomial interpolation and backfill strategies to address missing data. It begins by dividing the dataset into two subsets: the majority and minority classes, to preserve class-specific characteristics and reduce imputation bias. For each subset, missing values are initially estimated using a quadratic polynomial interpolation model, which captures non-linear relationships between features. Residual missing values, if any, are handled using a backfill process, which propagates the last observed non-missing value forward. Finally, the majority and minority subsets are merged to create a complete dataset. This structured and class-specific approach ensures data integrity, improves imputation accuracy and enhances model performance for complex and imbalanced datasets.

Notations and Symbols

Here it demonstrates through detailed guidelines the step-by-step calculation process of the proposed PoBa technique on a simplified dataset that contains missing values. The calculation demonstrates how second-degree polynomial interpolation works together with backfilling when implemented through the PoBa framework. Any continuous attribute can be represented by a feature vector named x which includes the following collection of values.

$$x = [2.1, 2.5, \text{NaN}, 3.8, \text{NaN}, 4.5]$$

Here, 'NaN' denotes missing values. The goal is to impute these missing values using PoBa.

Step 1: Identify Observed Values and Their Indices

Observed data points:

- Indices: $i = [0, 1, 3, 5]$

Algorithm: Polynomial with Backfill (PoBa)

Input: Raw data

Output: Imputed data

1: $D = \{DF_{\text{majority}}, DF_{\text{minority}}\}$, separate D into:
1.1: $DF_{\text{majority}} \leftarrow$ Majority Class Instances in D
1.2: $DF_{\text{minority}} \leftarrow$ Minority Class Instances in D
2: $For\ DF \in \{DF_{\text{majority}}, DF_{\text{minority}}\}$:
2.1: $For\ X_i \in DF(features\ with\ missing\ values)$:
2.1.1: $X_{\text{obs}} \leftarrow$ Observed indices in $DF(X_i)$
2.1.2: Fit: $X_i = \beta_0 + \sum_{k=1}^{d} \beta_k X_k +$
2.1.3: $For\ j \in X_{\text{missing}} (missing\ indices \in DF(X_i))$:
2.1.3.1: Predict: $X_{i,j} = \beta_0' + \sum_{k=1}^{d} \beta_k' X_{k,j}$
2.1.3.2: Replace $X_{i,j} \in DF(X_i)$
3: If $X_{\text{missing}} \neq \emptyset$ (remaining missing values after interpolation):
3.1: $For\ X_i \in DF$ (columns with missing values):
3.1.1: Apply Backfill: $X_{i,k} = X_{i,k-1}$, where $X_{i,k} \in X_{\text{missing}}$
4: Merge:
4.1: $D' \leftarrow DF_{\text{majority}} \cup DF_{\text{minority}}$
5: Output D'

Symbol	*Definition*
D	Raw dataset with missing values
DF_{majority}	Data frame containing majority class instances
DF_{minority}	Data frame containing minority class instances
X_i	Feature i
X_{obs}	Observed indices of feature X_i
X_{missing}	Missing indices of feature X_i
β_k	Coefficients of polynomial interpolation
$\in$	Error term in polynomial interpolation
$X_{i,j}$	Value of feature X_i at instance j
$X_{i,k}$	Backfilled value of X_i at index k
D'	Final imputed dataset after merging

- Values: $y = [2.1, 2.5, 3.8, 4.5]$

Step 2: Fit a 2nd-Degree Polynomial

Using the least squares method, fit a quadratic polynomial of the form:

$$f(i) = ai^2 + bi + c$$

Using the observed points $((0, 2.1), (1, 2.5), (3, 3.8), (5, 4.5))$, solve the normal equations to determine coefficients (a), (b), and (c).

Let us denote the design matrix A and target vector Y:

$$A = \begin{bmatrix} 0^2 & 0 & 1 \\ 1^2 & 1 & 1 \\ 3^2 & 3 & 1 \\ 5^2 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 9 & 3 & 1 \\ 25 & 5 & 1 \end{bmatrix}, Y = \begin{bmatrix} 2.1 \\ 2.5 \\ 3.8 \\ 4.5 \end{bmatrix}$$

Solving the normal equation $A^\top A \cdot \beta = A^\top Y$, it gives:

$$\beta = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \approx \begin{bmatrix} 0.0512 \\ 0.3137 \\ 2.0739 \end{bmatrix}$$

Thus, the polynomial becomes:

$$f(i) = 0.0512i^2 + 0.3137i + 2.0739$$

Step 3: Estimate Missing Values Using Polynomial

Now compute the imputed values at the missing indices $i = 2, 4$:

- For i = 2:
  $f(2) = 0.0512 \cdot 4 + 0.3137 \cdot 2 + 2.0739 = 0.2048 + 0.6274 + 2.0739 \approx 2.9061$
- For i = 4:
  $f(4) = 0.0512 \cdot 16 + 0.3137 \cdot 4 + 2.0739 = 0.8192 + 1.2548 + 2.0739 \approx 4.1479$

Step 4: Backfill Fallback

When polynomial estimation does not work in PoBa (stemming from inadequate observed data points) the system uses the following observation to fill in missing values. Because polynomial fitting worked out properly this dataset will not require backfilling of data points.

However, if $x = [\text{NaN}, \text{NaN}, 3.2]$, the polynomial cannot be constructed. Then, backfilling would yield:

$$x = [3.2, 3.2, 3.2]$$

Step 5: Final Imputed Vector

After applying PoBa on the original vector:

$$x_{\text{imputed}} = [2.1, 2.5, \boxed{2.9061}, 3.8, \boxed{4.1479}, 4.5]$$

The manual computation shows how the proposed PoBa method implies an interpretable method to manage missing data through its structured framework. The observed data points are submitted to quadratic interpolation in PoBa to calculate accurate estimates of missing values by identifying the non-linear trends present in the data set. If the dataset lacks sufficient data points to develop a valid polynomial, then PoBa applies the backfill strategy to preserve a usable complete dataset. Through its two-stage mechanism PoBa makes sure local consistency stays consistent and there is smoothness across feature values that primarily benefit from conditions of partial missingness. PoBa possesses a transparent and easily interpretable methodology that explains every imputed value through a simple algorithm addressing the black-box limitations of current deep-learning imputation methods.

3 Results and Discussion

3.1 Results

Results from the proposed PoBa method revealed better imputation results in all experimental trials. Table 2 demonstrates PoBa delivers better accuracy than both HPCNN and LPIHD throughout different levels of data loss (10%, 20%, and 30%) on Wine Quality and Heart Disease data. Under Heart Disease dataset with 30% missing data the combination of PoBa with MLP achieved accuracy of 0.950 thereby taking an edge over HPCNN-MLP (0.867) and LPIHD-MLP (0.901). In the Wine Quality dataset PoBa-MLP achieved its highest accuracy of 0.807 when handling 30% missing data while performing better than both HPCNN which attained 0.504 accuracy and LPIHD which reached 0.713 accuracy in utilizing the same model.

Under 30% missingness conditions PoBa-MLP demonstrated the best accuracy performance measured by MAE of 0.193 in Wine Quality and 0.050 for Heart Disease shown in Figure 2. The RMSE score reached its minimum point at 0.439 when measuring Wine Quality and 0.224 when assessing Heart Disease using the same missing rate under PoBa's management as Figure 3 demonstrates.

The performance comparison between PoBa imputation and LPIHD was measured using paired t-tests and Wilcoxon signed-rank tests under three different missingness levels (10%, 20%, and 30%) for both Wine Quality and Heart Disease datasets. Table 3 shows that PoBa produced substantial enhancements of Wine Quality dataset performance. The significance threshold of 0.05 was surpassed by the p-values obtained from both paired t-tests and Wilcoxon tests which produced results of 0.0050, 0.0000, 0.0003 and 0.0312 across the analyses at different missingness levels. Analysis results demonstrate that using PoBa improves Wine Quality classification accuracy beyond what LPIHD achieves.

The Heart Disease dataset shows no identifiable significant difference regarding accuracy between imputation methods. Both paired t-test p-values at 0.8381, 0.7951, 0.9550 and Wilcoxon test p-values at 0.6875, 0.5879, 1.0000 show identical accuracy results between PoBa and LPIHD across all chosen levels of missing data. The structural numeric data in Wine Quality benefits from PoBa but its effectiveness against Heart Disease dataset results in neutral performance as shown in Table 3.

Table 2. Comparative results of accuracy across various imputation techniques

Imputation Technique	Model	Wine quality - Accuracy			Heart Disease - Accuracy		
		10(%)	20(%)	30(%)	10(%)	20(%)	30(%)
HPCNN	LR	0.505	0.5	0.495	0.918	0.91	0.911
	LDA	0.501	0.489	0.5	0.909	0.901	0.894
	KNN	0.519	0.523	0.494	0.905	0.903	0.907
	NB	0.427	0.431	0.429	0.701	0.703	0.702
	SVM	0.522	0.503	0.508	0.913	0.913	0.913
	MLP	0.519	0.514	0.504	0.874	0.874	0.867
LPIHD	LR	0.733	0.702	0.697	0.908	0.912	0.917
	LDA	0.738	0.702	0.702	0.908	0.908	0.913
	KNN	0.741	0.728	0.673	0.903	0.912	0.911
	NB	0.698	0.69	0.673	0.715	0.754	0.789
	SVM	0.747	0.741	0.711	0.913	0.913	0.913
	MLP	0.754	0.753	0.713	0.873	0.904	0.901
PoBa	LR	0.743	0.759	0.759	0.919	0.911	0.923
	LDA	0.744	0.755	0.772	0.906	0.908	0.911
	KNN	0.755	0.764	0.804	0.909	0.91	0.903
	NB	0.723	0.739	0.771	0.689	0.738	0.731
	SVM	0.776	0.789	0.812	0.914	0.914	0.921
	MLP	0.777	0.795	0.807	0.891	0.916	0.95

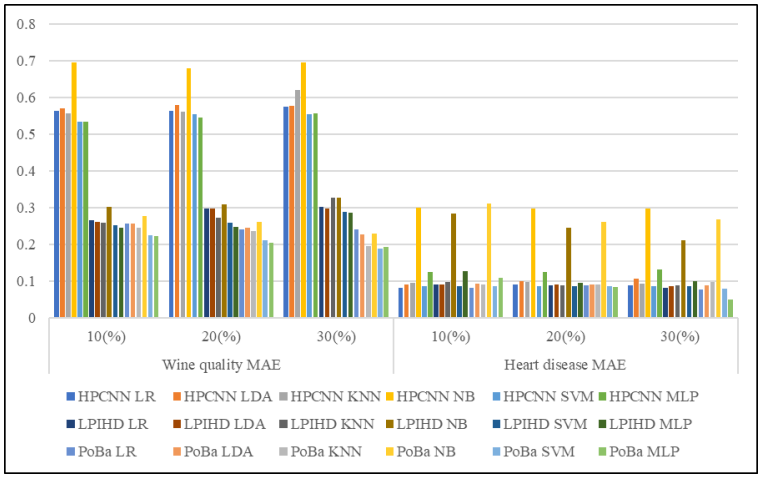

Fig 2. Comparative results of MAE across various imputation techniques

Table 3. Statistical Significance Test Results for Accuracy Comparison (PoBa vs LPIHD)

Dataset	Missing Rate	Paired t-test (p-value)	Wilcoxon Test (p-value)
Wine Quality	10%	0.0050	0.0312
	20%	0.0000	0.0312
	30%	0.0003	0.0312
Heart Disease	10%	0.8381	0.6875
	20%	0.7951	0.5879
	30%	0.9550	1.0000

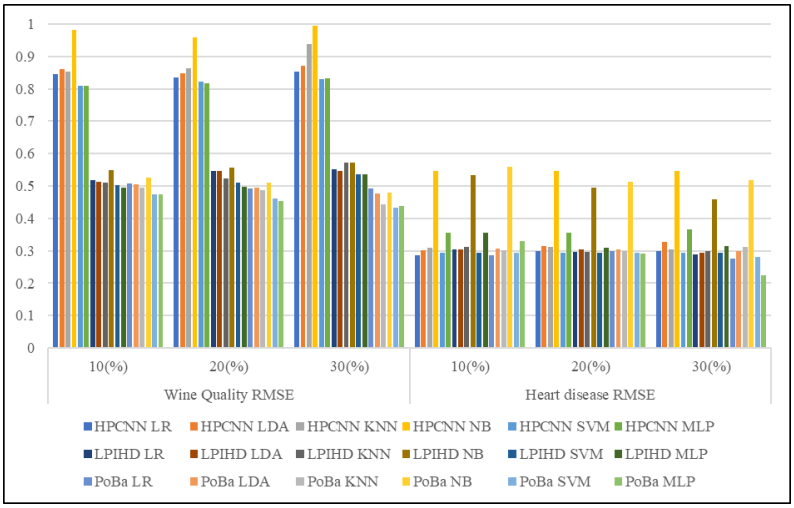

Fig 3. Comparative results of RMSE across various imputation techniques

The density distribution of higher accuracy rates for PoBa persists particularly when the missing data reaches 30% levels as shown in Figure 4. The precise and mirror-like peaks of PoBa signal improved data reliability as well as decreased outcome variability compared to LPIHD measurements.

The effectiveness of PoBa exceeds LPIHD accuracy levels as presented in Figure 5. The accuracy performance of KNN NB SVM and MLP increases substantially as missing data numbers rise to 30% where they gain an average of 13% accuracy improvement. The enhanced performance rates confirm the results that were statistically significant.

The accuracy distribution levels of both PoBa and LPIHD demonstrate an almost identical pattern as shown in Figure 6. Both median and standard deviation values show matching patterns while displaying close similarity of variation at the 10% and 20% data loss levels. The non-significant statistical results receive additional backing from these findings.

Figure 7 reveals that NB and similar classification models experience limited or adverse accuracy enhancement as missing values rise to 30%. Nevertheless, MLP and SVM models demonstrate slight advancements during this case scenario. The data indicates PoBa performs equally well as LPIHD for this dataset independent of improvement strength.

Visual analysis confirms that PoBa enhances Wine Quality dataset accuracy in all missingness conditions (**Figure 4 and 5**) but produces performance results for Heart Disease that are similar to LPIHD with at most no notable improvements (**Figure 6 and 7**).

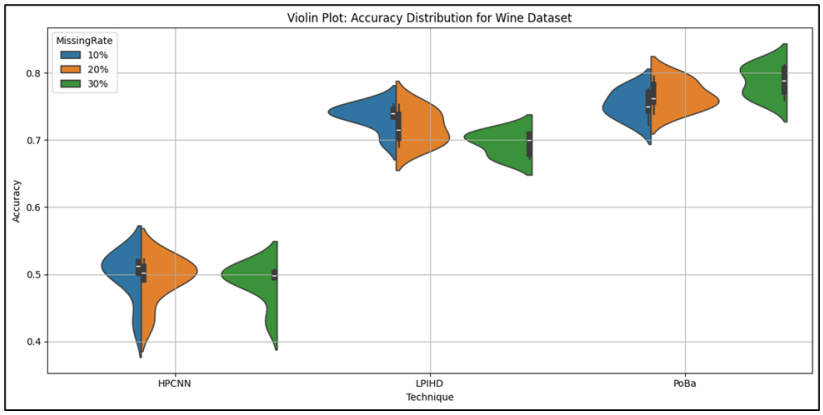

Fig 4. Violin Plot - Accuracy Distribution for Wine Dataset

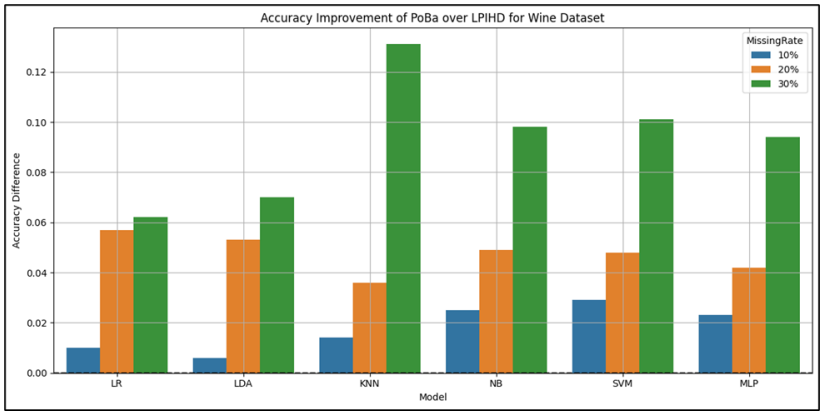

Fig 5. Accuracy Improvement of PoBa over LPIHD for Wine Dataset

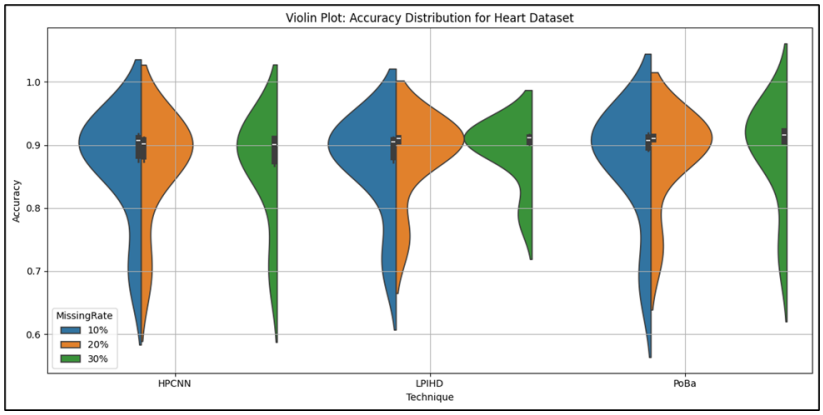

Fig 6. Violin Plot - Accuracy Distribution for Heart Dataset

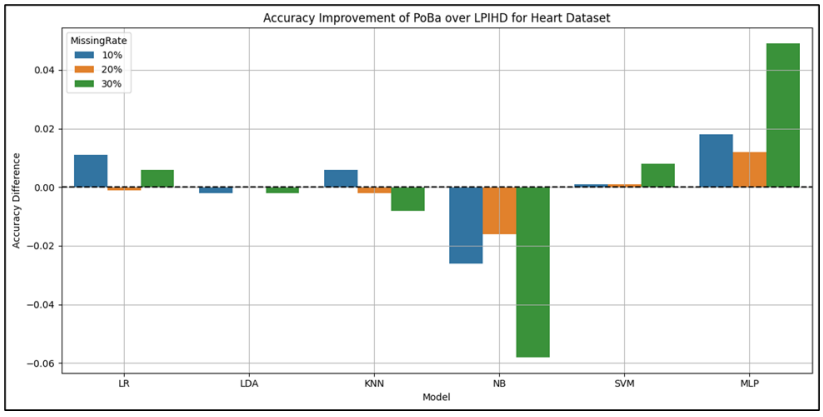

Fig 7. Accuracy Improvement of PoBa over LPIHD for Heart Dataset

3.2 Discussion

PoBa delivers superior results than existing approaches in the field as shown by experimental results when operating on datasets that demonstrate both non-linear patterns and imbalanced class distributions. The CNN-driven HPCNN model[13] requires extensive spatial data processing with optimized kernel operations, yet PoBa operates as a light-weight interpretable framework that supports efficient computation. The Heart Disease data showed acceptable performance from HPCNN, but it failed to deliver quality results on Wine Quality data because the model struggled with non-linear patterns since it lacked specific knowledge of domain-related spatial structures.

The PoBa algorithm exhibited better flexibility than the hybrid LPIHD model[14] at handling increasing missing data while detecting Wine Quality dataset non-linear relationships. The LPIHD approach achieves satisfactory generalization with moderate data imbalance yet its inability to address imputation at a granular class level where PoBa directly performs by utilizing its two-step approach. The designed class-specific approach delivered effective bias mitigation because it addressed prediction challenges that arise during high-missingness situations when standard imputation approaches introduce systematic biases to majority class results. Statistically significant Wine Quality improvements that PoBa achieved throughout different missingness stages in Table 3 validate the effectiveness of its specific interpolation and backfilling strategy toward discovering intricate intra-class feature patterns.

Results published by Oktaviani et al.[10] support PoBa system because they established the effectiveness of quadratic interpolation in temporal datasets even with small samples. The backfilling mechanism included in PoBa extends basic missing value interpolation by completing data through a specific approach that uses no ensemble modeling or iterative regressors. Las-Heras et al.[11] employed MARS to handle missing data imputation with success in particular environmental conditions yet showed limited universality for varied data distributions. PoBa demonstrates equal data distribution characteristics across Wine and Heart Disease datasets indicating its ability to perform consistently with dataset-independent results.

The outcomes in Table 2 establish that PoBa maintains feature structure integrity because they demonstrate minimal distributional differences between primary and interpolated data points throughout key characteristics.

The results from Figure 7 suggest PoBa delivers additional benefits to LPIHD in handling high levels of missing data despite statistical performance insignificance in Table 3. Analysis by Nijman et al.[8] confirmed that imputation techniques need to guarantee fairness between different classes while PoBa fulfills this requirement.

Different organizations and institutions can use the scalable and transparent design of PoBa as an effective tool for resource-limited infrastructure or healthcare applications that demand interpretability in their operations. HPCNN[13] demands specialized infrastructure together with model-specific tuning whereas PoBa operates effectively without infrastructure requirements or special tuning needs thus providing a practical choice for real-world deployment.

PoBa represents an advanced solution in imputation approaches by combining custom class interpolation methodology with unparameterized backfilling regulations. This combined approach creates a system which optimizes both interpretability and achieves high accuracy while maintaining scalability according to the previous studies[5],[7],[11]. PoBa fulfills both strict quantitative assessment requirements and preserves data distribution, so it becomes a dependable framework available across diverse and high-scale datasets.

Table 2 shows that the proposed PoBa technique produces enhanced classification outcomes when the missing rate rises even though this behavior seems unexpected at first. Several interconnected design elements of PoBa together with characteristics of datasets creates this behavioral pattern. The class-sensitive imputation system in PoBa maintains original distributions within classes because it divides majority instances from minority instances before performing imputation. The separation of sample classes by PoBa reduces both bias in reconstructed data and maintains consistent features in the imputation process. The size of missing data leads polynomial interpolation to work with progressively longer index spans which reduces local noise and produces cleaner model-friendly feature values. The approach works as a regularizing mechanism because it benefits SVM and MLP models that depend on high-quality features. The backfill mechanism serves as a protection measure that preserves both database integrity by maintaining complete structures and proximity between elements. The data imputation process naturally eliminates both noisy data points and inconsistent labels that would most likely negatively impact model generalization. The present improvements demonstrate both PoBa's tolerance towards data errors and its implicit mechanism which establishes a more standardized representation of features under uncertain conditions.

4 Conclusion

This study introduces PoBa a new method for missing data imputation that solves real-world datasets' two essential problems regarding feature trend non-linearity and distribution class imbalance. Values created through second-degree polynomial interpolation and backfilling ensure both conceptual integrity and legal accuracy. Reliable alternatives to HPCNN[12] and

LPIHD[13] exist in PoBa which represents an interpretable yet lightweight system that works within specific classes without deep architecture requirements.

The Heart Disease and Wine Quality datasets proved the superior capability of PoBa to maintain effective performance during various classifier tests. PoBa achieved exceptional results as a combination with MLP where it demonstrated 0.950 accuracy and 0.050 MAE and 0.224 RMSE while processing Heart Disease data with 30% missingness indicating its capability to work effectively with medical data imbalance. The reliable performance enhancements at higher missing rates stem from PoBa's feature noise reduction capability which preserves intra-class feature relationships as proved by statistical results and visual distributions.

The key features of PoBa consist of its combination between interpolation and backfill strategies alongside a tailored imputation process for each class category and general applicability across data domains at a small computational cost. The current performance strength of PoBa with numeric features extends to most situations however it does not scale well with high-dimensional sparse datasets or attributes comprised of many categorical values when polynomial assumptions prove invalid. The approach aims to add categorical encoding methods and parallelized polynomial fitting algorithms along with ensemble-based imputation models to enhance it. It will enable PoBa to expand its capabilities for real-time sensitive applications like healthcare diagnostics and financial forecasting and industrial process monitoring while resolving data incompleteness both rapidly and accurately.

5 Acknowledgement

The authors acknowledge the support of the Department of Science and Technology (DST), Government of India, for providing the FIST (Fund for Improvement of S&T Infrastructure) program funding. This assistance has significantly contributed to the research and development activities outlined in this work.

References

1) Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*. 2021;2(5):377. https://doi.org/10.1007/s42979-021-00765-8.
2) Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*. 2022;22(1):287. https://doi.org/10.1186/s12874-022-01768-6.
3) Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *Journal of Big Data*. 2021;8(1):1–37. https://doi.org/10.1186/s40537-021-00516-9.
4) Alam S, Ayub MS, Arora S, Khan MA. An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*. 2023;9:100341. https://doi.org/10.1016/j.dajour.2023.100341.
5) Xu D, Hu PJH, Huang TS, Fang X, Hsu CC. A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. *Journal of Biomedical Informatics*. 2020;111:103576. https://doi.org/10.1016/j.jbi.2020.103576.
6) Köse T, Özgür S, Coşgun E, Keskinoğlu A, Keskinoğlu P. Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. *BioMed Research International*. 2020;2020(1):1895076. https://doi.org/10.1155/2020/1895076.
7) Pastorini M, Rodríguez R, Etcheverry L, Castro A, Gorgoglione A. Enhancing environmental data imputation: A physically-constrained machine learning framework. *Science of The Total Environment*. 2024;926:171773. https://doi.org/10.1016/j.scitotenv.2024.171773.
8) Nijman SW, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJL, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*. 2022;142:218–229. https://doi.org/10.1016/j.jclinepi.2021.11.023.
9) Wang P, Hu T, Gao F, Wu R, Guo W, Zhu X. A hybrid data-driven framework for spatiotemporal traffic flow data imputation. *IEEE Internet of Things Journal*. 2022;9(17):16343–16352. https://doi.org/10.1109/JIOT.2022.3151238.
10) Oktaviani ID, Abdurohman M, Erfianto B. Increasing tiny data imputation accuracy using temporal polynomial interpolation. IEEE. 2022;p. 357–361. https://doi.org/10.1109/ICoICT55009.2022.9914838.
11) Lasheras FS, Nieto PJG, García-Gonzalo E, Gómez FA, Iglesias FJR, Sánchez AS. Missing data imputation for continuous variables based on multivariate adaptive regression splines. Cham. Springer. 2020;p. 73–85. https://doi.org/10.1007/978-3-030-61705-9_7.
12) Khumukcham RS, Mayanglambam D, Urikhimbam BC, Hoque N. MVI-DR: An Efficient Missing Value Imputation Method Using Decision Tree and Regression Analysis. *Indian Journal of Science and Technology*. 2023;16(43):3862–3874. https://doi.org/10.17485/IJST/v16i43.1864.
13) Khan H, Rasheed MT, Liu H, Zhang S. High-order polynomial interpolation with CNN: A robust approach for missing data imputation. *Elsevier Preprint*. 2023. https://doi.org/10.1016/j.compeleceng.2024.109524.
14) Deepa AV, Beena TLA. Enhancing data imputation in complex datasets using Lagrange polynomial interpolation and hot-deck fusion. *The Scientific Temper*. 2025;16(1):3727–3735. https://doi.org/10.58414/SCIENTIFICTEMPER.2025.16.1.19.
15) Almeida MM, Almeida JD, Quintanilha DB, Júnior GB, Silva AC. A meta-learning based neural network and LSTM for univariate time series missing data imputation. *Applied Soft Computing*. 2025;172:112845. https://doi.org/10.1016/j.asoc.2025.112845.
16) Cortez P, Cerdeira A, Almeida F, Matos T, Reis J. *Wine Quality*. 2025. https://doi.org/10.24432/C56S3T.
17) Internet source accessed on Feb-20-2025 from [Dataset]. Available from: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data.