Decision Analytics Journal 9 (2023) 100341

Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity

Shafiq Alam [a,1,*], Muhammad Sohaib Ayub [b,1], Sakshi Arora [c,1], Muhammad Asad Khan [d,1]

[a] School of Management, Massey University, Auckland, New Zealand
[b] Department of Computer Science, SBASSE, LUMS, Lahore, Pakistan
[c] School of Computing, Whitireia Community Polytechnic, Auckland, New Zealand
[d] Department of Telecommunication, Hazara University, Mansehra, Pakistan

ARTICLE INFO

Keywords:
Classification
Clustering
Imputation
Ordinal data
Partitioning Around Medoids
Multilayer Perceptron

ABSTRACT

Missing data can significantly impact dataset integrity and suitability, leading to unreliable statistical results, distortions, and poor decisions. The presence of missing values in data introduces inaccuracies in clustering and classification and compromises the reliability and validity of such analyses. This study investigates multiple imputation techniques specifically designed for handling missing values in ordinal data commonly encountered in surveys and questionnaires. Quantitative approaches are used to evaluate different imputation methods on various datasets with varying missing value percentages. By comparing the performance of imputation techniques using clustering metrics and algorithms (e.g., k-means, Partitioning Around Medoids), the study provides valuable insights for selecting appropriate imputation methods for accurate data analysis. Furthermore, the study examines the impact of imputed values on classification algorithms, including k-Nearest Neighbors (kNN), Naive Bayes (NB), and Multilayer Perceptron (MLP). Results demonstrate that the decision tree method is the most effective approach, closely aligning with the original data and achieving high accuracy. In contrast, random number imputation performs poorly, indicating limited reliability. This study advances the understanding of handling missing values and emphasizes the need to address this issue to enhance data analysis integrity and validity.

1. Introduction

Missing data can severely hinder effective decision-making [1]. Particularly within the realm of large, real-world datasets, the absence of data is a common and inherent challenge. Such gaps in data prevent comprehensive analysis and obstruct the generation of valuable inferences necessary for decision-making [2]. Existing data analysis methodologies may fall short or produce inaccurate results when faced with missing values. Consequently, it is critical to develop strategies to manage these missing values effectively, thereby enhancing decision-making systems [1]. In instances like surveys and questionnaires, multiple factors contribute to missing values, ranging from poorly designed surveys and deliberate non-responsiveness from participants to the omission of sensitive data for confidentiality purposes [3].

Recent advancements in data analysis techniques offer a myriad of methods to tackle the issue of missing data. One such simplistic approach is to discard data vectors containing missing values. However, this technique, despite its simplicity, can introduce biases in the resulting analysis and is impractical for datasets with substantial missing data [4]. In the real world, datasets often contain missing values that obstruct statistical inferences and decision-making [3]. It is thus critical to assign values – whether arbitrary or informed – to complete these datasets [5]. Imputation is a solution to missing values, wherein missing data is replaced by plausible approximations [6]. Uninformed patterns of absent data commonly appear in real-world datasets [7]. The chosen imputation technique should aim to improve data analysis and deliver unbiased results [8]. Simple techniques such as replacing missing values with the dataset's mean, median, or mode exist, but there are also more complex classification-based methods [8].

Ordinal data refers to observations that consist of variables ranked on a scale [5]. Examples include scales such as high, medium, and low. While ordinal values can be ordered, their distances cannot be measured. On the other hand, missing values refer to the absence of data due to various factors such as transmission errors, non-responses, data entry errors, or data capture errors [9]. These missing values can

* Corresponding author.
E-mail addresses: salam1@massey.ac.nz (S. Alam), 15030039@lums.edu.pk (M.S. Ayub), sakshi.arora@whitireianz.ac.nz (S. Arora), asadkhan@hu.edu.pk (M.A. Khan).
[1] Equal contribution from each Author.

https://doi.org/10.1016/j.dajour.2023.100341
Received 19 July 2023; Received in revised form 6 October 2023; Accepted 8 October 2023
Available online 12 October 2023
2772-6622/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

have a significant impact on the insights that can be derived from the data.

Missing data can be categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [10]. In the case of MCAR, the absence of values occurs without any identifiable pattern or reason. For example, participants may accidentally overlook a question or skip a page in surveys or questionnaires [10]. MAR refers to missing values that are related to other observed variables. For instance, missing values related to gender may occur if male participants are less likely to undergo a depression survey [11]. In the case of MNAR, the missing values are directly related to the values themselves. For example, participants with high alcohol consumption may choose not to answer a question about their drinking habits [10].

The proportion of missing values in a dataset plays a role in selecting appropriate treatment methods. Missing values can be considered trivial if they account for less than 1% of the complete dataset [9]. Percentages between 1% and 5% are still manageable, while percentages between 5% and 15% require more sophisticated approaches for handling missing data. Once the missing data exceeds 15%, it can significantly impact the analysis, necessitating careful treatment measures. Real-world data is often imperfect, unreliable, and noisy. To prepare such data for further computations and analysis, it undergoes preprocessing, which involves cleaning, reducing, transforming, and unifying the data [12]. Preprocessing ensures the quality and fitness of the data for subsequent analysis.

Imputation is the process of replacing missing instances of data with valid values [12]. While one simple approach is to delete vectors containing missing values, this method is often unsuitable [9]. Instead, imputation is a preferred method for treating missing values. In this study, various imputation methods are evaluated, including mean, median, mode, class mean, class median, class mode, random value imputation, k-Nearest Neighbor (k-NN) distance-based median, and several classifier-based imputations such as Support Vector Machine (SVM), Decision tree, and Neural Network (NN).

Classification is a machine learning technique used to predict a group for a given instance based on training data. It enables efficient categorization of new data [13]. Among various classification algorithms, decision trees and Naive Bayes (NB) are commonly used for imputation [14]. Clustering is an unsupervised data mining technique that groups similar data together [15]. In the process of clustering, vectors are grouped to form cohesive clusters containing similar elements. The quality of a cluster is determined by the closeness of elements within the cluster and the distance of cluster elements from non-cluster elements [15]. To assess the identified imputation methods, this study employs two clustering methods: k-means and PAM clustering.

The motivation for this research is primarily driven by the many drawbacks of missing values, including classification inaccuracy, mining inconsistencies, and certain algorithms' inability to process data. One of the imputation methods to address missing values involves imputation based on domain knowledge. However, evaluating such imputations is challenging due to the actual missing values' absence. This study explores different imputation methods to substitute near-real values for missing values and the effect of such imputations on various data mining algorithms. The research employs supervised mining methods to evaluate the proposed imputation methods' impacts. The research contributes by identifying different imputation methods for the missing data value problem, evaluating the performance of these methods, identifying the best technique for imputing missing values in ordinal data, and assessing whether the imputed values for ordinal data can improve data clustering. This research is driven by the following objectives:

- Evaluate and compare various imputation techniques for handling missing ordinal data.
- Assess the impact of imputation techniques on clustering performance and quality using clustering algorithms like k-means and Partitioning Around Medoids (PAM).
- Analyze the effect of imputed values on classification accuracy for various algorithms including k-Nearest Neighbors (kNN), Naive Bayes (NB), and Multilayer Perceptron (MLP).

This study uses a quantitative approach and leverages open-source tools like R-language [16] and WEKA [17] for data analysis. Several benchmark datasets with ordinal values from the UCI machine learning repository [18] are selected. These datasets are then manipulated to create subsets with varying degrees of missing values: 1%, 5%, 10%, 15%, 20%, and 25%. Various imputation methods are accessed to fill in the missing values. The effectiveness of these methods is evaluated by applying different clustering algorithms, such as k-means and Partitioning Around Medoids (PAM), to the completed datasets and comparing the clustering results with the original datasets. This analysis provided insights into the impact of imputation techniques on the clustering performance and allowed for a comprehensive evaluation of the imputation methods' effectiveness. In addition to clustering analysis, classification experiments are conducted to evaluate the effect of imputed values on the accuracy of various classification algorithms, including k-Nearest Neighbors (kNN), Naive Bayes (NB), and Multilayer Perceptron (MLP). By assessing the classification accuracy of the imputed datasets compared to the original datasets, the study provides insights into the impact of different imputation techniques on the performance of classification algorithms in the context of ordinal data. We have made the following contributions:

- We have conducted comprehensive experiments on five datasets, each comprising six subsets containing missing values, by the evaluation of various methods for handling missing values in ordinal datasets and found the most suitable technique for imputing missing instances, specifically in ordinal data.
- We have performed a comprehensive evaluation of clustering performance by measuring the impact of these eleven imputation techniques on clustering performance using four key metrics, including Sum of Squares Within clusters (SSW), cluster diameter, average intra-cluster distance, and average inter-cluster distance.
- We have performed the evaluation of classification performance using various datasets and diverse machine learning algorithms to analyze the impact of imputation methods on classification accuracy.

The rest of the paper is structured as follows: Section 2 delves into the existing research and establishes the context for the study. Section 3 outlines the proposed experimental methodology for data collection and performance assessment. Section 4 performs an in-depth analysis of the results and findings obtained through rigorous experimentation, providing valuable insights and interpretations. The last section summarizes the key findings and their implications, discusses the limitations of the study, and suggests potential avenues for future research.

2. Related work

This section provides a summary of existing research conducted in the field of data analysis related to topics such as ordinal data, missing data, imputation methods, and clustering algorithms. Data quality has become a crucial factor in Data Mining and Business Intelligence, as incomplete and noisy data pose challenges in analysis and can lead to erroneous statistical inferences and decision-making [11,19]. Therefore, it is essential to address missing values before performing any analysis. The summary of the commonly used techniques for treating missing values is presented in Table 1.

Various methods have been developed to handle missing values in numeric data, although their suitability for ordinal datasets may

Table 1
Summary of Related Approaches Missing Data Imputation.

Work	Description
[5]	Joreskog's classification of missing data types
[3]	Factors contributing to missing values in surveys
[8]	Imputation techniques for improving data analysis
[2]	Challenges posed by missing data in analysis
[1]	Impact of missing data on decision-making
[6]	Imputation as a solution for replacing missing data
[7]	Uninformed patterns of absent data in real-world datasets
[18]	UCI Machine Learning Repository for benchmark datasets
[4]	Critique of discarding data vectors with missing values
[16]	R-language as an open-source tool for data analysis
[20]	Artificial Neural Networks for model-based imputation
[21]	Evaluation of *k*-Nearest Neighbors imputation method
[22]	Missing data imputation using predictive models
[23]	Impact of removing samples with high missing values proportion
[24]	Determining the acceptable threshold for missing values
[25]	Limitations of simple statistical imputation approaches
[26]	Systematic review of data imputation methods in data mining
[27]	Data imputation for missing values and ensuring data integrity
[28]	Regression techniques for model-based imputation
[29]	Deep Learning-based imputation approaches
[30]	Comparison of dynamic imputation techniques

vary [11]. In the context of treating missing values, several commonly used techniques are worth mentioning.

Case Deletion (CD) involves discarding vectors that contain missing data, resulting in a new dataset for further processing. However, this method is not appropriate for datasets with a higher degree of missing values. Even for datasets with smaller quantities of missing data, caution must be exercised to evaluate potential bias introduced by the resultant dataset [9].

Random Value Imputation can be used to replace missing values, thereby creating a completed dataset. While this approach is straightforward, it does not utilize any information from the dataset and may introduce randomness that affects subsequent analysis [31].

Mean Imputation (MI) replaces missing values with the mean value of the corresponding feature or attribute in the complete dataset, as shown in Eq. (1). This type of replacement is also known as complete mean imputation [5]. An alternative approach involves calculating the mean for a given class, similar to the class of the vector containing the missing value [19]. It is important to note that mean imputation may not be suitable for datasets with numerous missing values, as it reduces variance and tends to overestimate the sample size. Furthermore, mean imputation is not appropriate for ordinal data, given that the concept of mean does not apply in an ordinal scale.

$$\hat{x}ij = \sum i : x_{ij} \in C_k \frac{x_{ij}}{n_k} \tag{1}$$

Most Common Imputation (MCI) method replaces missing values with the most frequently occurring value in the dataset [13,32]. This approach assumes that the most common value represents a plausible estimate for the missing data.

Median Imputation replaces missing values with the median value of the corresponding feature in the dataset. Eq. (2) represents complete median imputation [33]. Alternatively, class median imputation involves replacing missing values with the median of the feature within a specific class. The class should correspond to the class variable of the vector containing the missing value.

$$\hat{x}ij = \text{median}\left(xij : x_{ij} \in C_k\right) \tag{2}$$

The aforementioned imputation methods can be further customized to improve their effectiveness. Mean, median, and mode values can be calculated for each class, increasing the chances of approximating the original values. For example, instead of replacing missing values with the mean income of all families, the mean income of only married samples can be utilized. However, it is important to acknowledge that this approach may introduce bias, as all elements within a group will have the same mean value.

k-Nearest Neighbor (k-NN) distance-based median procedure replaces missing values with k instances that are most similar in terms of relevance [34]. The similarity between instances is determined using a distance metric. The procedure involves splitting the dataset into two fragments: DTm, which represents a subset containing at least one vector with a missing feature or instance, and DTc, which contains the remaining vectors without any missing data. For each row in DTm, the instance row is partitioned into observed and missing sections. The distance between the observed section and the rows of DTc is computed, followed by the identification of the k most similar instances using the k-nearest neighbors approach [34].

Several techniques have been proposed for handling missing data in ordinal datasets. Decision trees have demonstrated effectiveness in the classification of ordinal data, as they facilitate the categorization of data into branches or splits [35]. Traditional approaches for handling missing values can lead to biased estimates and may reduce or exaggerate statistical power [14,36]. Deleting missing instances is often chosen as the quickest method for treating missing data and is the default procedure in many statistical data analysis tools. However, this approach may result in the deletion of a significant portion of the data in real-world scenarios.

Neural networks offer a suitable technique for addressing the missing value problem in ordinal data. They represent a classifier-based imputation technique that emulates the physiological functioning of the brain. Neural networks typically consist of an input layer, a hidden layer, and an output layer. Through training algorithms, neural networks effectively solve complex mathematical equations, with the hidden layer adapting during the process [37,38]. Another classifier-based imputation technique for handling missing data is Support Vector Machines (SVM) [39,40].

Classification-based imputation techniques have emerged as viable approaches for estimating and imputing missing values in datasets [41,42]. These techniques leverage various classification methods, such as neural networks, decision trees, and similar procedures, to address the missing value problem. Notably, previous research has predominantly focused on handling numerical or nominal missing data values [43]. In contrast, this study aims to bridge the existing research gap by specifically addressing the treatment of missing values in ordinal data and investigating the impact of these treatment methods on unsupervised learning techniques, particularly clustering.

Applying classification-based imputation techniques entails constructing a projection model that facilitates the estimation of values closely resembling the actual data, ultimately filling in the missing instances [44]. In this process, the vectors with missing features serve as the reaction vector, while the remaining vectors contribute to the construction of the projection model. However, it is essential to recognize the limitations associated with this approach, as highlighted by Shah et al. [4]. Firstly, the estimated values obtained through classification-based imputation often exhibit superior performance compared to the actual values. Secondly, the accuracy of valuing missing values may be compromised when there are no inherent associations between complete vectors and vectors with missing values. Finally, the computational cost associated with building multiple models for predicting missing values is relatively high. These considerations underscore the need for a critical evaluation of the classification-based imputation techniques when addressing missing values in ordinal data. Additionally, works such as [45] provide essential insights into technologies and phases vital for addressing missing data challenges.

3. Experimental design

In this section, the design of our study for the performance of different imputation techniques in handling missing values across varying datasets and missing value scenarios is thoroughly examined. The implementation does not necessitate any special hardware requirements. The software employed in this study includes widely accessible open-source tools, namely R-language [16] and WEKA [17].

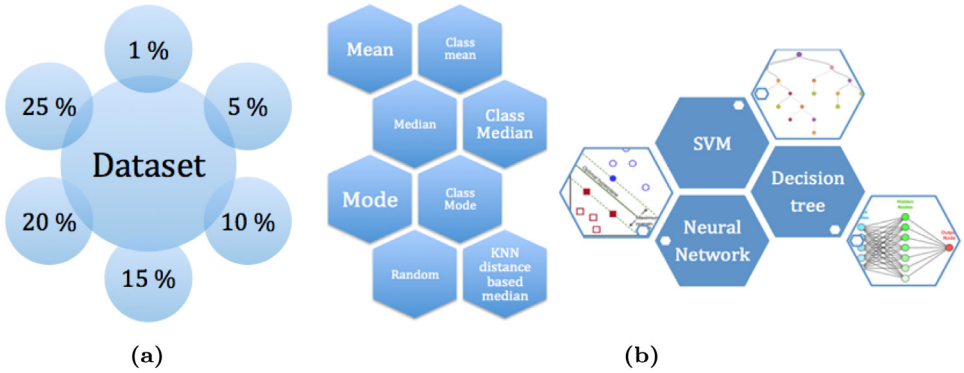

Fig. 1. (a) Data subsets with missing data percentages and (b) imputation methods used for experimentation.

Table 2
Datasets Details.

Dataset	#Instances	#Attributes	Class detail
CAR	1728	5	4 (unacc,acc,good,vgood)
ESL	488	5	9 (1,2,3,4,5,6,7,8,9)
LEV	1000	5	5 (0,1,2,3,4)
ERA	1000	5	9 (1,2,3,4,5,6,7,8,9)
SWD	1000	11	4 (2,3,4,5)

3.1. Dataset description

We have selected five benchmark datasets from the UCI machine learning repository [18] that exhibit the presence of ordinal values, as summarized in Table 2. These datasets serve as the foundation for evaluating diverse imputation methods. Subsequently, the selected datasets undergo processing to generate six additional datasets, each characterized by a distinct proportion of missing values. The experimental outcomes are acquired for subsets derived from the original dataset, comprising missing value proportions of 1%, 5%, 10%, 15%, 20%, and 25%, as visually represented in Fig. 1(a). This systematic approach enables us to thoroughly examine the performance of different imputation techniques in handling missing values across varying datasets and missing value scenarios.

3.2. Problem assessment

A comprehensive evaluation of various data imputation methods is conducted, with a specific focus on their applicability in data mining techniques, particularly clustering and classification. The assessment encompasses a range of imputation techniques, including random imputation, class mean imputation, global mean value imputation, global median value imputation, class mode value imputation, class median value imputation, most common value imputation, and classification-based imputation methods such as Support Vector Machines (SVM), Decision Tree, and Neural Networks (NN), as visually depicted in Fig. 1(b).

3.3. Choice of imputation algorithms

The selection of imputation algorithms was driven by the need to ensure a comprehensive evaluation of methods for handling missing ordinal data within diverse datasets. Each chosen technique serves a specific purpose, allowing us to cover a wide spectrum of imputation strategies.

- **Mean, Median, and Mode Imputation:** These basic techniques were included to provide a fundamental comparison. Mean imputation offers simplicity and efficiency, while median and mode imputation provide robustness against outliers and skewed data, respectively.
- **Class-based Imputation:** Considering class-specific statistics in imputation is crucial for datasets with inherent class structures. Class mean, class median, and class mode imputation ensure that imputed values align with the characteristics of their respective classes, preserving important class-related information.
- **Random Number Imputation:** The randomness introduced by this method allows for variation in the imputed dataset. This is essential in scenarios where uncertainty or variability in missing values is a significant consideration.
- **k-NN Distance-Based Median Imputation:** Utilizing the local neighborhood of a missing value for imputation is essential for datasets with localized patterns. k-NN imputation leverages the proximity of instances to ensure that imputed values reflect the local data structure accurately.
- **SVM and Decision Tree Imputation:** Classification-based imputation techniques are valuable when the relationship between features is complex. SVM and decision tree imputation leverage the power of classification algorithms to predict missing values, capturing intricate data patterns effectively.
- **Neural Network Imputation:** Neural networks, being versatile and powerful models, can approximate complex data patterns. Imputing missing values using a neural network ensures that the imputed dataset aligns well with the original data's underlying patterns.

This diverse selection of imputation techniques enables a thorough evaluation, considering both the inherent characteristics of the missing data and the complexities present within the ordinal datasets.

3.4. Performance assessments

We evaluate the efficacy of different imputation methods in enhancing the learning outcomes for datasets afflicted with missing values. The evaluation centers around various clustering and classification algorithms suitable for ordinal data and compares their performance with that of the original datasets. Specifically, the clustering algorithms employed are k-means and Partitioning Around Medoids (PAM). k-Nearest Neighbors (kNN), Naive Bayes (NB) and Multilayer Perceptron (MLP) are used for classification assessment.

4. Experimental evaluation

In this section, we present a detailed analysis of the results obtained from conducted experimentation. The experimentation encompasses five datasets, each consisting of six subsets with missing data ranging from 1% to 25%. For each subset, eleven imputation techniques are implemented to address the missing values. We discuss the accuracy achieved and the average variance of imputed values compared to the original values for each dataset. We highlight the performance of different imputation techniques, including mean, median, mode, class

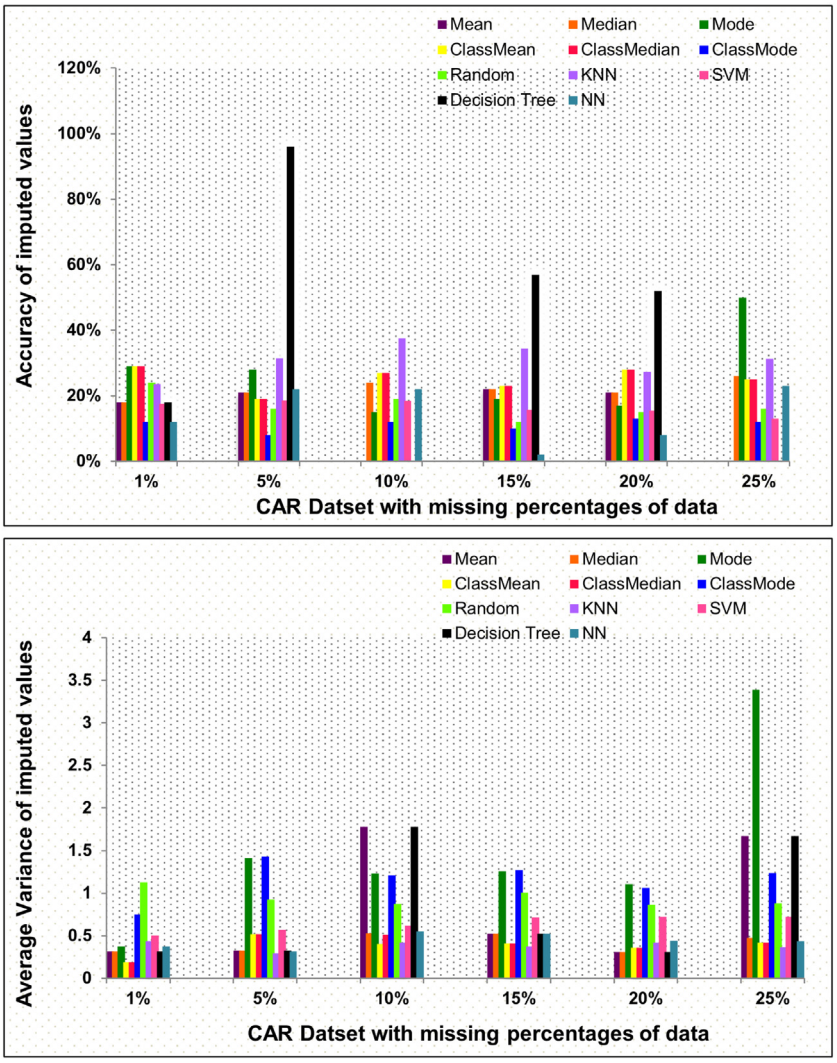

Fig. 2. Accuracy and average variance of various imputation techniques for the CAR dataset.

mean, class median, class mode, random number imputation, k-NN distance-based median imputation, SVM, Decision Tree, and Neural Network. Each of these techniques is applied to the datasets to fill in the missing values and obtain completed datasets. The subsequent analysis and evaluation of these imputed datasets allow for a comprehensive assessment of the performance and effectiveness of each imputation technique. Furthermore, we examine the clustering performance of the imputed data using the k-means and PAM algorithms, comparing the clusters formed with the original dataset. This analysis helps us identify the most suitable techniques for addressing the problem of missing values in ordinal data and provide valuable insights for future research and applications.

The choice of clustering and classification algorithms in this study was deliberate to ensure a robust evaluation in the context of ordinal data. For clustering, we utilized k-means and PAM algorithms. k-means offers efficiency, suitable for large datasets, while PAM handles outliers and noise effectively. These algorithms align well with ordinal data, grouping based on ordinal rankings to assess imputed data clusters against the original dataset. In classification evaluation, SVM, Decision Tree, and Neural Network were chosen for their ability to capture nonlinear relationships in data, essential for ordinal data exhibiting such patterns. This selection ensures a meaningful analysis of imputation effects on classification accuracy, aligning with ordinal data's characteristics.

4.1. Imputation performance for ordinal data

This section presents the accuracy achieved and the average variance of imputed values compared to the original values using eleven selected techniques for imputing missing values using each subset of the dataset containing missing values in varying proportions. The accuracy and variance of imputed values highlight the suitability of different imputation methods. The decision tree emerges as a promising technique for imputing missing values in ordinal datasets, demonstrating high accuracy and minimal deviation from the original dataset. The detailed results and findings for each dataset are presented in the subsequent sections.

4.1.1. Imputing missing values in the CAR dataset

Fig. 2 shows the accuracy and average variance across all subsets of the CAR dataset when different imputation methods are applied, represented by different colors. The decision tree method, a way of filling in missing values based on patterns in the data, achieves the best accuracy of 96% in one of the subsets. This high accuracy suggests that the decision tree method works effectively for this type of dataset. On the other hand, the neural network-based method does not perform well, indicating that it might not be the best choice for handling missing values in this dataset.

Two methods that fill in missing data with the most common value, called mode and class mode, show similar results for the CAR dataset,

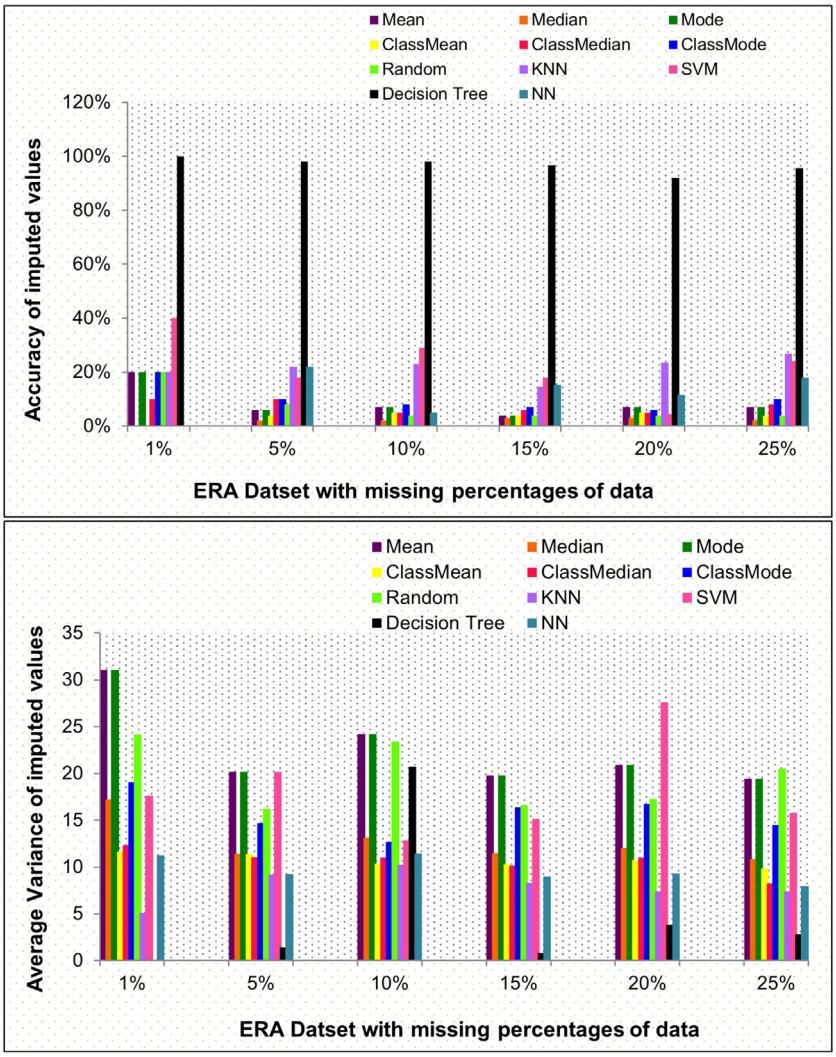

Fig. 3. Accuracy and Average variance of various imputation techniques for the ERA dataset.

which shows that these two methods behave similarly when used on this dataset. When looking at the average variance, which is a measure of how much the imputed dataset deviates from the original data, the highest variances are found for values replaced with the mode and class mode. Random number imputation also shows a high variance. These high variances suggest that these methods might distort the original data structure, which could affect the results of any analysis done on the imputed data.

4.1.2. Imputing missing values in the ERA dataset

Fig. 3 illustrates the accuracy and average variance of various imputation techniques applied to the subsets of the ERA dataset. The decision tree technique, a type of classification-based imputation, consistently displays excellent accuracy, over 90%, across all subsets with different proportions of missing data. In contrast, the class median technique struggles with this dataset. Its accuracy remains significantly low, not crossing the 11% threshold in any of the subsets. This suggests that the class median might not be an appropriate choice for dealing with missing data in the ERA dataset.

When considering the average variance, the decision tree method emerges as the most accurate, showing the least variation. This reinforces the superiority of this method, as it manages to preserve the original data structure while substituting missing values. In contrast, the mean imputation method results in the highest variation compared to the original dataset. This could potentially distort subsequent data analysis and affect the reliability of any conclusions drawn from it. Thus, while mean imputation may be a simple and convenient method, it seems to be less suitable for handling missing data in the ERA dataset.

4.1.3. Imputing missing values in the LEV dataset

Fig. 4 shows the percentage of correctly imputed values for the missing data compared to the original dataset, along with the averaged variance as a summary of the results for the LEV dataset, which is further divided into six sub-datasets containing missing values at varying percentages.

For the LEV dataset, we observe that the decision tree approach, a classifier-based imputation technique, outshines the other methods for replacing missing values and achieves up to 80% accuracy, highlighting its optimal fit for this context. Contrarily, the random number imputation method does not perform well, achieving less than 10% accuracy for most subsets of missing values in the LEV dataset. The class median imputation technique also shows a similar trend, with a low accuracy of about 13% in some subsets. These poor performances emphasize the superiority of the decision tree method for imputing missing values in this dataset.

4.1.4. Imputing missing values in the SWD dataset

Fig. 5 summarizes the accuracy of the imputed data and the variance of the new data compared to the original dataset using various imputation techniques for the SWD dataset divided into six sub-datasets, each containing missing instances at different percentages.

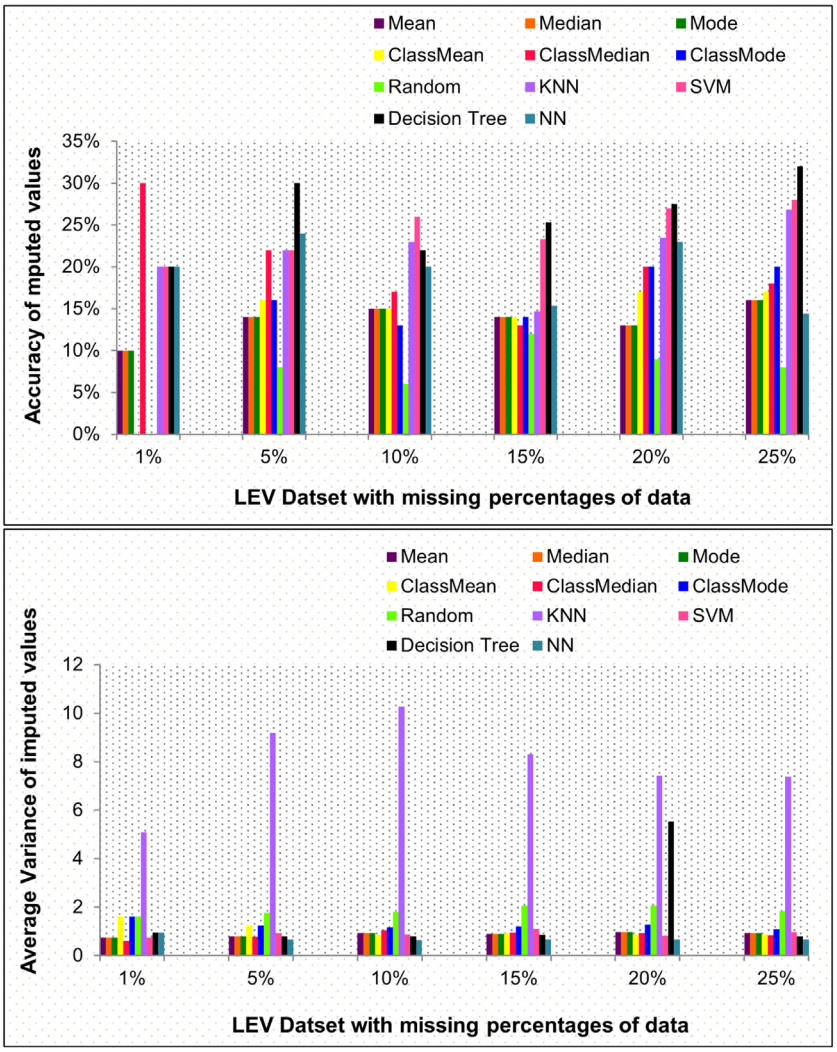

Fig. 4. Accuracy and Average variance of various imputation techniques for the LEV dataset.

For the SWD dataset, the decision tree method consistently excels in imputing missing values and maintains an average accuracy close to 80% across all subsets, a testament to its robust performance. Simultaneously, this technique displays the lowest average variance compared to the original data, reinforcing the accuracy of the imputed values. Conversely, the Support Vector Machine (SVM) classifier-based imputation method exhibits a significantly high variance from the original data, indicating its limitations in reproducing the characteristics of the original data.

Further, the method of substituting missing instances with random numbers reveals subpar results, underscoring its unsuitability for this task. It achieves the lowest accuracy among all eleven techniques, never surpassing roughly 15% accuracy in any of the SWD dataset subsets. This significant contrast emphasizes the superiority of the decision tree method for imputing missing values in the SWD dataset.

4.1.5. *Imputing missing values in the ESL dataset*

Fig. 6 shows the accuracy of the imputed values and depicts the average variance of the completed dataset compared to the original ESL dataset for various imputation methods. Regarding the ESL dataset, the decision tree method demonstrates a steady performance with approximately 40% accuracy across all missing value subsets. Despite its consistency, it may not represent the optimal choice for managing missing values in this specific dataset.

Interestingly, the imputation technique using the mode of a given feature offers superior results. This mode-based imputation reaches accuracy heights up to 80%, with an impressive average accuracy of nearly 35%. This performance suggests that in some datasets, simpler statistical measures like the mode could outperform more complex classifier-based methods. Conversely, the strategy of filling missing instances with random numbers performs poorly with the ESL dataset. It records the lowest accuracy among all eleven methods, underscoring its ineffectiveness for this dataset. These observations highlight the importance of selecting the most appropriate imputation method based on the unique characteristics of each dataset.

4.2. *Clustering performance using imputed data*

After appropriately imputing missing values into the dataset, accounting for varying degrees of missing data, this section analyzes the clustering performance of the imputed data in relation to the original dataset, applying k-means and Partitioning Around Medoids (PAM) clustering algorithms to assess the clustering performance.

The PAM clustering algorithm is considered to be an improved version of the k-means algorithm for analyzing unclassified data. The comparisons presented in this section are conducted on the original dataset as well as on the eleven imputation techniques mentioned earlier. The evaluation is performed based on the Sum of Squares Within (SSW) criterion, as proposed by Zhao et al. [46], which measures the dispersion of elements within a cluster. Additionally, four other metrics are employed to compare the various formed clusters,

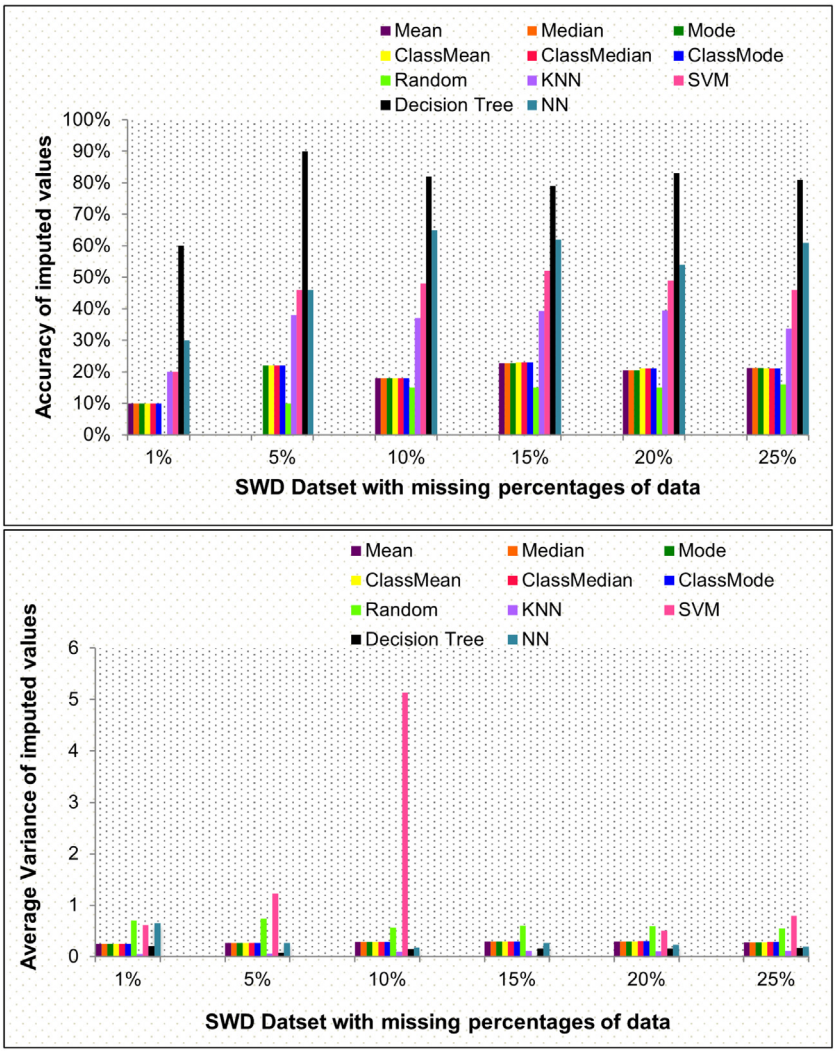

Fig. 5. Accuracy and Average variance of various imputation techniques for the SWD dataset.

including the diameter of the cluster, the average distance between elements within a cluster (referred to as intra-cluster distance), the average distance between elements of different clusters (referred to as inter-cluster distance), and the size of the clusters formed.

4.2.1. Clustering performance in the CAR dataset

Fig. 7 presents the outcomes for Partitioning Around Medoids (PAM) and k-means (KM) clustering algorithms as applied to the CAR dataset.

The most compatible imputation method, in terms of generating clusters closely matching the original data, is the decision tree technique, a classifier-based imputation approach. The imputation accomplished via the Support Vector Machines (SVM) results in the most optimal Sum of Squares Within clusters (SSW) values. The collective agreement of additional parameters, such as the diameter of the clusters formed, the average distance between elements within a cluster, and the average distance among elements from different clusters, reaffirms the SVM method's suitability for imputing missing values in this dataset. On the contrary, other techniques implemented for handling missing values failed to generate clusters that are sufficiently reliable or consistent to replace the original dataset's clusters. This signifies that different datasets may benefit more from specific imputation methods over others, underscoring the importance of carefully selecting the most suitable imputation technique for each dataset.

4.2.2. Clustering performance in the ERA dataset

Fig. 8 shows the comparison of the clustering performance of imputed and original data for the ERA dataset, utilizing both the Partitioning Around Medoids (PAM) and k-Means clustering algorithms.

The ERA dataset lends itself to the formation of highly refined clusters, distinguished by small intra-cluster distances and substantial inter-cluster separations. The decision tree technique, a classifier-based approach for managing missing values, accurately reproduces these attributes. The decision tree method aligns closely with the original data based on the SSW metric, comparing clusters formulated after imputing values with the decision tree technique to those generated utilizing the original data values. Conversely, the clusters created by implementing the class median technique for imputation deviate significantly from those based on the original dataset, particularly when evaluated on the SSW parameter. Clusters engendered by substituting missing data with class median values are found to be significantly smaller than the original clusters. Therefore, the decision tree method emerges as the most compatible imputation technique for the ERA dataset, effectively preserving the dataset's inherent clustering properties.

4.2.3. Clustering performance in the LEV dataset

Fig. 9 shows the analytical findings associated with the LEV dataset for the comparison across eleven imputation techniques using four distinct comparative metrics employing k-means (KM) and Partitioning Around Medoids (PAM) clustering methodologies.

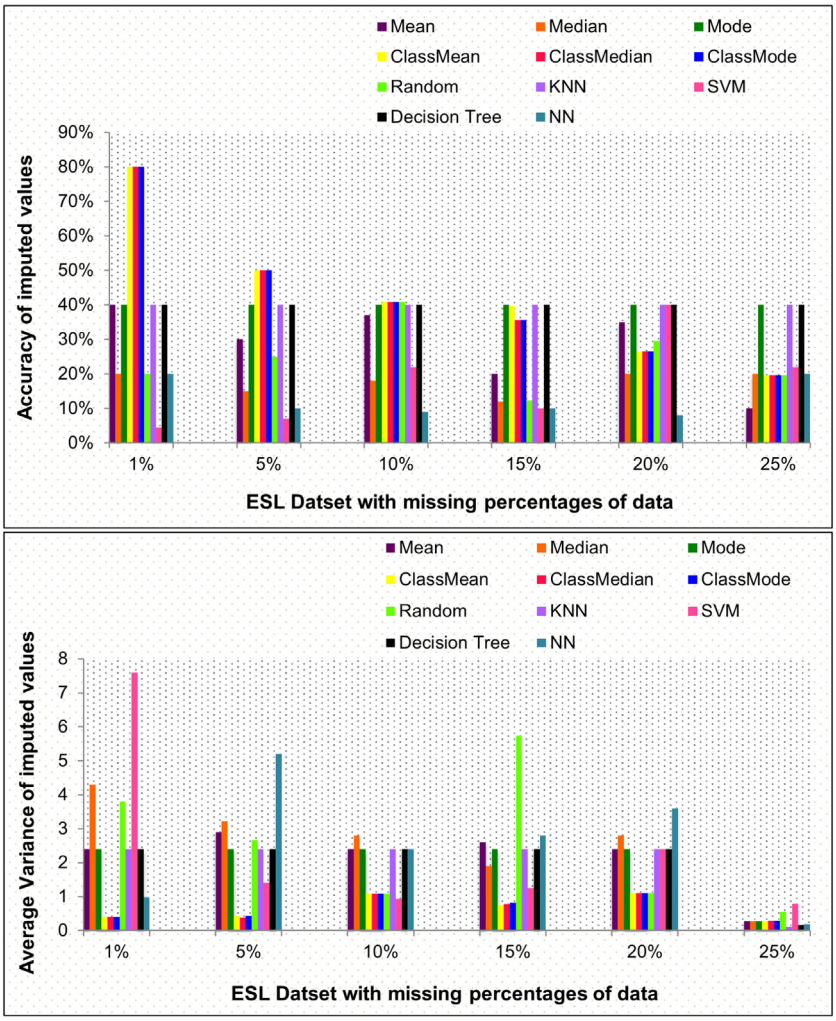

Fig. 6. Accuracy and Average variance of various imputation techniques for the ESL dataset.

Among the different techniques used for handling missing data, the decision tree approach performed better. The values predicted by the decision tree algorithm produced clusters that closely resembled the original data in terms of intra-cluster and inter-cluster distances. Furthermore, the Sum-Of-Squares Within cluster (SSW) parameter confirmed the suitability of the decision tree method for imputing missing values in the LEV dataset. In contrast, replacing missing data with randomly generated values showed the weakest correlation with the original dataset's cluster formations.

4.2.4. Clustering performance in the SWD dataset

Fig. 10 presents the results derived for the SWD dataset, obtained via the utilization of two pre-established clustering algorithms, PAM and k-means (KM), using four different comparative matrices.

The imputed values using the decision tree and neural network (NN) classifier-based approaches maintain a high degree of resemblance to the clusters constructed with the original dataset, specifically on the SSW measurement criterion. However, when referring to the remaining comparative metrics, the decision tree methodology emerges as the most effective imputation technique for the SWD dataset. The diameter of clusters, the average intra-cluster distances, and the average inter-cluster distances from the decision tree method closely align with the metrics extracted from the clusters of the original dataset. Among all the methods evaluated, the random number imputation method has yielded the least cluster similarity, thereby limiting the dataset's applicability in subsequent analyses.

4.2.5. Clustering performance in the ESL dataset

Fig. 11 demonstrates the clustering efficiency of the various imputation methods on the ESL dataset based on four comparative matrices.

In our analysis of the ESL dataset, the missing instances imputed using the decision tree approach have yielded the most accurate approximation to the original data. Utilizing comparative matrices, we found that the dataset imputed using the decision tree method resulted in clusters that closely matched those formed by the original dataset. This suggests the high efficacy of the decision tree imputation technique in maintaining the structure and characteristics of the ESL dataset, thereby enabling accurate and reliable data analysis.

The mode imputation-based technique proved to be the most suitable for the ESL dataset using other matrices in terms of producing accurate and coherent clusters. The method of imputing missing values using randomly generated numbers proved to be the least effective approach for the ESL dataset. This is particularly noticeable in its performance on the SSW parameter, where it exhibited the most significant variance compared to the other imputation methods.

Table 3 presents the key findings from our investigation, providing a clear and succinct overview of the effectiveness of the different imputation techniques in dealing with missing data in ordinal datasets.

4.3. Classification performance using imputed data

One of the central hurdles stemming from missing data in classification lies in the risk of losing valuable insights, ultimately leading

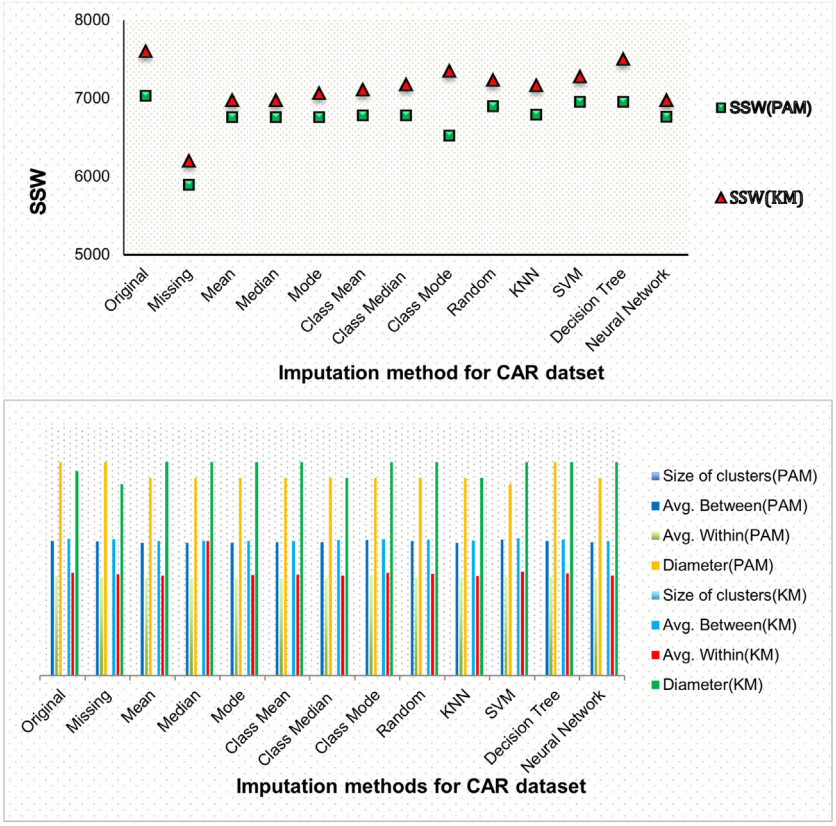

Fig. 7. Clustering performance of various imputation techniques for the CAR dataset.

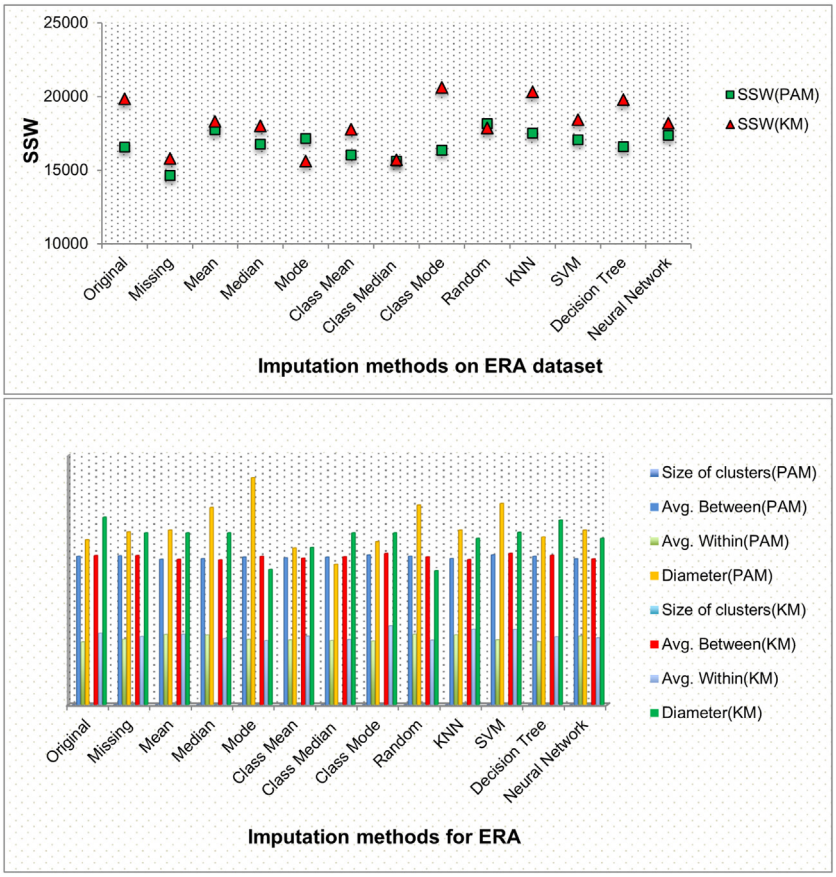

Fig. 8. Clustering performance of various imputation techniques for the ERA dataset.

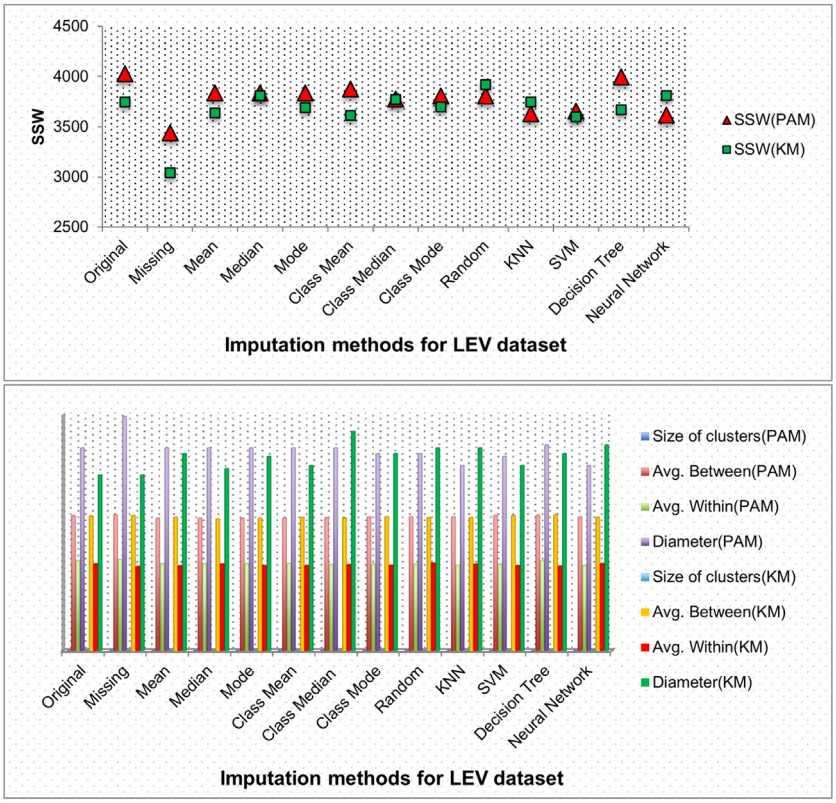

Fig. 9. Clustering performance of various imputation techniques for the LEV dataset.

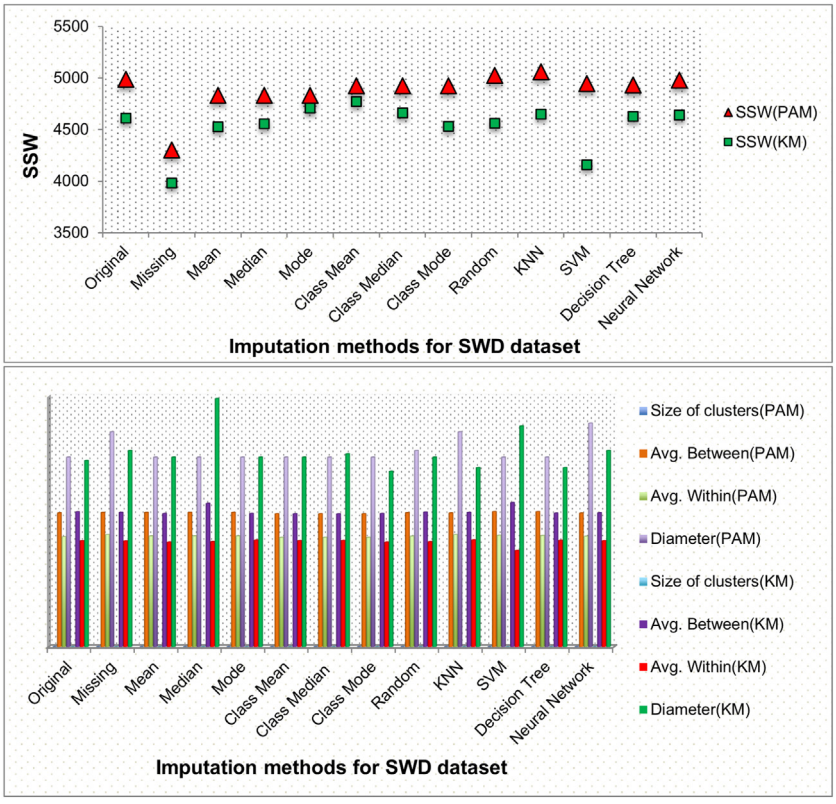

Fig. 10. Clustering performance of various imputation techniques for the SWD dataset.

to skewed or erroneous model predictions. The absence of certain data points can skew the distribution of features, impair the distinctiveness of classes, and alter decision boundaries, all of which pose significant obstacles to the classifier's capacity to generalize effectively. Moreover,

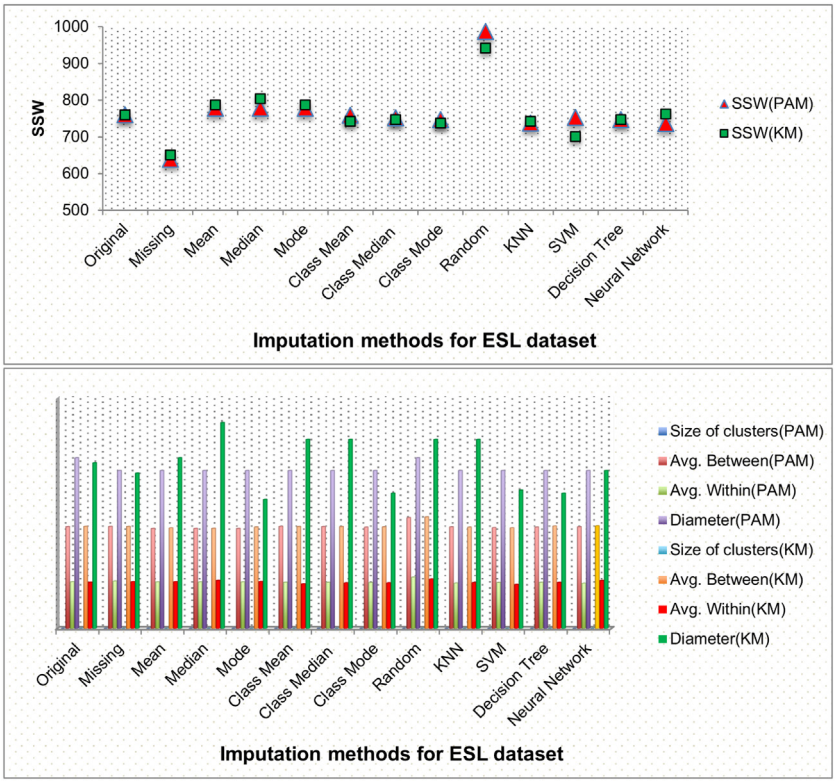

Fig. 11. Clustering performance of various imputation techniques for the ESL dataset.

Table 3
Comparison of the best-suited method for Imputing Missing Values and Clustering Performance using various datasets.

Dataset	Best suitable method for	
	Imputing missing values	Learning clustering performance
CAR	Decision tree	Decision tree
ERA	Decision tree	Decision tree
LEV	Decision tree	Decision tree
SWD	Decision tree	Decision Tree and Neural Network
ESL	Class Mean, Median and Mode	Neural network

missing data often translates into reduced sample sizes, making it a daunting task to construct robust models, especially when dealing with rare or imbalanced classes. In this context, imputation techniques assume a pivotal role in alleviating these challenges by filling in the gaps left by missing values, thereby restoring the dataset's completeness. The integration of imputed values empowers classification algorithms to access a more comprehensive feature space, resulting in enhanced model performance. Beyond that, imputation contributes to upholding the integrity of the dataset's statistical characteristics, thereby ensuring that the classification model yields more precise and dependable predictions.

In addition to the comparison of clustering performance on original as well as imputed data, we have also conducted experiments to evaluate the classification accuracy across various datasets under different training and testing split ratios using three different machine learning algorithms, i.e., k-Nearest Neighbors (kNN), Naive Bayes (NB) and Multilayer Perceptron (MLP). The algorithms used for imputation varied, with the Support Vector Machine, J48 Decision Tree, Logistic Regression, and Random Forest each providing the best accuracy under different conditions. Each of these experiments helped to uncover the potential impact of different algorithms, split ratios, and imputation methods on classification accuracy across the various datasets. Below are the details of each experiment.

Table 4
Summary of the Best Classification Accuracy Results for various datasets with 50% training & 50% testing and 20% training & 80% testing split ratio using kNN algorithm.

Dataset	Split ratio	Best accuracy	Algorithm
CAR	50:50	90.16%	Logistic Regression
CAR	20:80	82.85%	Logistic Regression
ESL	50:50	60.66%	Support Vector Machine
ESL	20:80	57.18%	Logistic Regression
LSV	50:50	62.00%	Random Forest
LSV	20:80	61.00%	Support Vector Machine
ERA	50:50	30.00%	Logistic Regression
ERA	20:80	24.13%	Logistic Regression
SWD	50:50	60.40%	Support Vector Machine
SWD	20:80	56.75%	Random Forest

4.3.1. *Classification performance using k-nearest neighbor algorithm (kNN)*

Table 4 shows the best classification accuracy results for the k-Nearest Neighbors (kNN) algorithm utilizing varying algorithms for five selected datasets, which were randomly split into 50% training & 50% testing data and 20% training & 80% testing data. The highest accuracy for the CAR dataset is achieved using Logistic Regression under both 50:50 and 20:80 split ratios. The effectiveness of the algorithms varied across datasets, reflecting the different characteristics of each dataset.

4.3.2. *Classification performance using Naive Bayes algorithm (NB)*

Table 5 shows the performance of the Naive Bayes (NB) classifier under the same split ratios of 50% training & 50% testing data and 20% training & 80% testing data. Each dataset's best accuracy is shown in comparison to the corresponding imputation algorithm, thereby exploring how different algorithms might impact the outcomes. The J48 Decision Tree showed particularly consistent performance, providing the best accuracy in a number of scenarios.

Table 5
Summary of the Best Classification Accuracy Results for various datasets with 50% training & 50% testing and 20% training & 80% testing split ratio using Naive Bayes (NB) algorithm.

Dataset	Split ratio	Best accuracy	Algorithm
CAR	50:50	86.81%	Support Vector Machine
CAR	20:80	78.25%	J48 Decision Tree
ESL	50:50	63.52%	Support Vector Machine
ESL	20:80	62.38%	Logistic Regression
LSV	50:50	61.60%	Support Vector Machine
LSV	20:80	59.83%	J48 Decision Tree
ERA	50:50	28.40%	Logistic Regression
ERA	20:80	25.86%	Random Forest
SWD	50:50	62.00%	J48 Decision Tree
SWD	20:80	62.18%	J48 Decision Tree

Table 6
Summary of the Best Classification Accuracy Results for various datasets with 50% training & 50% testing and 20% training & 80% testing split ratio using MLP algorithm.

Dataset	Split ratio	Best accuracy	Algorithm
CAR	50:50	97.34%	Original
CAR	20:80	96.95%	Random Forest
ESL	50:50	65.12%	Random Forest
ESL	20:80	64.23%	Random Forest
LSV	50:50	64.80%	Random Forest
LSV	20:80	63.55%	Random Forest
ERA	50:50	27.20%	Logistic Regression
ERA	20:80	26.20%	Logistic Regression
SWD	50:50	59.80%	Random Forest
SWD	20:80	56.93%	Random Forest

4.3.3. Classification performance using multilayer perceptron algorithm (MLP)

Table 6 shows the best classification accuracy results for the selected datasets using 50% training & 50% testing and 20% training & 80% testing split ratios are gauged using the Multilayer Perceptron (MLP) algorithm. For the majority of the tested datasets and split ratios, the Random Forest algorithm appeared to deliver the highest accuracy, with a notable exception for the CAR dataset at a 50:50 split ratio where the original dataset without imputation delivered the best performance. Logistic Regression also proved to be the most effective for the ERA dataset in both split ratios.

5. Conclusion and future work

In this study, we explored various imputation methods for handling missing values in ordinal data using five different datasets. The imputation techniques scrutinized included Mean, Class Mean, Median, Class Median, Mode, Class Mode, k-NN distance-based Median, and classifier-based methods such as SVM, decision tree, and neural network. The efficiency of these techniques is evaluated based on their accuracy and average variance. The findings highlighted that the decision tree imputation technique is the most effective for dealing with missing values in ordinal data. Not only did it exhibit high accuracy, but it also recorded the lowest variance compared to the original data, suggesting that decision tree imputation can accurately fill data gaps while also enhancing the data analysis process.

Moreover, we conducted classification experiments to understand how the imputed values influenced the accuracy of different classification algorithms. These experiments are executed across three different setups: using the k-Nearest Neighbors (kNN) algorithm, utilizing the Naive Bayes (NB) classifier, and implementing the Multilayer Perceptron (MLP) algorithm. These trials provided us with additional insights into the relationship between imputation methods, data-splitting strategies, and the effectiveness of different classification algorithms. In terms of data clustering, our results revealed that clusters formed with the decision tree imputation closely mirrored those derived from the original data, further underscoring the efficacy of this imputation method.

In summary, addressing missing values is a critical aspect of ensuring accurate and unbiased data analysis. Our study emphasizes the efficiency of the decision tree imputation technique for managing missing values in ordinal data. The efficacy of this method is evident not only in its ability to complete datasets accurately but also in its contribution to improving data clustering and enhancing the performance of classification algorithms.

Looking forward, while our study provides useful insights, it is limited by time constraints and a focus on a specific subset of data with a fixed percentage of missing values for evaluating clustering performance. Future work could extend to evaluating a broader range of clustering algorithms and matrices to compare clusters more effectively. Enlarging the pool of ordinal datasets with missing values for our experiments could furnish a more comprehensive analysis. Furthermore, exploring the impact of imputed values on other data analysis tasks for ordinal datasets could open up new directions for future research.

Funding

No specific funding was obtained for this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Used publically available data.

References

[1] S. Tufféry, Data Mining and Statistics for Decision Making, Wiley, John & Sons, 2011.
[2] S.C.W.C. Albright, W. Winston, C. Zappe, Data Analysis and Decision Making, Cengage Learning, 2010.
[3] A. Pantanowitz, T. Marwala, Evaluating the impact of missing data imputation, in: Advanced Data Mining and Applications, Springer Berlin Heidelberg, 2009, pp. 577–586.
[4] A.D. Shah, J.W. Bartlett, J. Carpenter, O. Nicholas, H. Hemingway, Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study, Am. J. Epidemiol. 179 (6) (2014) 764–774.
[5] K.G. Jöreskog, Structural Equation Modeling with Ordinal Variables Using LISREL, Technical Report, Scientific Software International, Inc., Lincolnwood, IL, 2005.
[6] K.J. Lee, J.C. Galati, J.A. Simpson, J.B. Carlin, Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study, Stat. Med. 31 (30) (2012) 4164–4174.
[7] I. Eekhout, R.M. de Boer, J.W. Twisk, H.C. de Vet, M.W. Heymans, Missing data: a systematic review of how they are reported and handled, Epidemiology 23 (5) (2012) 729–732.
[8] M. Huisman, Imputation of missing network data: Some simple procedures, J. Soc. Struct. 10 (1) (2009) 1–29.
[9] E. Acuna, C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: Classification, Clustering, and Data Mining Applications, Springer, 2004, pp. 639–647.
[10] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychol. Meth. 7 (2) (2002) 147.
[11] W.H. Finch, Imputation methods for missing categorical questionnaire data: A comparison of approaches, J. Data Sci. 8 (3) (2010) 361–378.
[12] C.E. Brodley, M.A. Friedl, Identifying mislabeled training data, J. Artif. Intell. Res. 11 (1999) 131–167.
[13] L. Rodwell, K.J. Lee, H. Romaniuk, J.B. Carlin, Comparison of methods for imputing limited-range variables: a simulation study, BMC Med. Res. Methodol. 14 (1) (2014) 57.
[14] X. Su, R. Greiner, T.M. Khoshgoftaar, A. Napolitano, Using classifier-based nominal imputation to improve machine learning, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, Springer, 2011, pp. 124–135.
[15] J.M. Lingeman, D. Shasha, Clustering data, in: Network Inference in Molecular Biology: A Hands-on Framework, Springer, 2012, pp. 11–22.
[16] R. Core, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2014.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.
[18] K. Bache, M. Lichman, UCI Machine Learning Repository, UC Irvine, CA, USA, 2013, URL http://archive.ics.uci.edu/ml.
[19] A.C. Acock, Working with missing values, J. Marriage Fam. 67 (4) (2005) 1012–1028.
[20] S.J. Choudhury, N.R. Pal, Imputation of missing data with neural networks for classification, Knowl.-Based Syst. 182 (2019) 104838.
[21] U. Pujianto, A.P. Wibawa, M.I. Akbar, et al., K-nearest neighbor (k-NN) based missing data imputation, in: International Conference on Science in Information Technology, ICSITech, IEEE, 2019, pp. 83–88.
[22] S. Mercaldo, J.D. Blume, Missing data and prediction: the pattern submodel, Biostatistics 21 (2) (2020) 236–252.
[23] C.-Y. Hung, B.C. Jiang, C.-C. Wang, Evaluating machine learning classification using sorted missing percentage technique based on missing data, Appl. Sci. 10 (14) (2020) 4920.
[24] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, J. Big Data 8 (1) (2021) 1–37.
[25] P.C. Austin, I.R. White, D.S. Lee, S. van Buuren, Missing data in clinical research: a tutorial on multiple imputation, Can. J. Cardiol. 37 (9) (2021) 1322–1331.
[26] A.R. Ismail, N.Z. Abidin, M.K. Maen, Systematic review on missing data imputation techniques with machine learning algorithms for healthcare, J. Robotics Control (JRC) 3 (2) (2022) 143–152.
[27] P.C. Chiu, A. Selamat, O. Krejcar, K.K. Kuok, S.D.A. Bujang, H. Fujita, Missing value imputation designs and methods of nature-inspired metaheuristic techniques: A systematic review, IEEE Access (2022).
[28] U. Shahzad, N.H. Al-Noor, M. Hanif, I. Sajjad, M. Muhammad Anas, Imputation based mean estimators in case of missing data utilizing robust regression and variance–covariance matrices, Comm. Statist. Simulation Comput. 51 (8) (2022) 4276–4295.
[29] W.-C. Lin, C.-F. Tsai, J.R. Zhong, Deep learning for missing value imputation of continuous data and the effect of data discretization, Knowl.-Based Syst. 239 (2022) 108079.
[30] H. Ahn, K. Sun, K. Kim, Comparison of missing data imputation methods in time series forecasting, Comput. Mater. Continua 70 (1) (2022) 767–779.
[31] H.W.H. Hui, W. Kong, H. Peng, W.W.B. Goh, The importance of batch sensitization in missing value imputation, Sci. Rep. 13 (1) (2023) 3003.
[32] K. Psychogyios, L. Ilias, C. Ntanos, D. Askounis, Missing value imputation methods for electronic health records, IEEE Access 11 (2023) 21562–21574.
[33] J. Sim, J.S. Lee, O. Kwon, Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications, in: Mathematical Problems in Engineering, 2014.
[34] B. Wei, F. Yang, X. Wang, Y. Ge, M.B. Wei, Package 'knngarden', 2012, knnGarden.
[35] E.L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, M.D. Cubiles-de-la Vega, Missing value imputation on missing completely at random data using multilayer perceptrons, Neural Netw. 24 (1) (2011) 121–129.
[36] C. Wongkamthong, O. Akande, A comparative study of imputation methods for multivariate ordinal data, J. Surv. Stat. Methodol. 11 (1) (2023) 189–212.
[37] N. Sengupta, M. Udell, N. Srebro, J. Evans, Sparse data reconstruction, missing value and multiple imputation through matrix factorization, Sociol. Methodol. 53 (1) (2023) 72–114.
[38] C. Jacobsen, U. Zscherpel, P. Perner, A comparison between neural networks and decision trees, in: Machine Learning and Data Mining in Pattern Recognition, Springer, 1999, pp. 144–158.
[39] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru, C. Yumei, A SVM regression based approach to filling in missing values, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2005, pp. 581–587.
[40] D. He, Active learning for ordinal classification on incomplete data, Intell. Data Anal. 27 (3) (2023) 613–634.
[41] A.A. Ahmed, New technique for imputing missing item responses for an ordinal variable, 2007, Using Tennessee Youth Risk Behavior Survey as an Example.
[42] A. Palanivinayagam, R. Damaševičius, Effective handling of missing values in datasets for classification using machine learning methods, Information 14 (2) (2023) 92.
[43] S. Pan, S. Chen, Empirical comparison of imputation methods for multivariate missing data in public health, Int. J. Environ. Res. Public Health 20 (2) (2023) 1524.
[44] M. Saar-Tsechansky, F. Provost, Handling missing values when applying classification models, J. Mach. Learn. Res. (2007).
[45] R.T. Rasheed, M.A. Mohammed, N. Tapus, Big data analysis, Mesop. J. Big Data 2021 (2021) 22–25.
[46] Q. Zhao, M. Xu, P. Fr"anti, Sum-of-squares based cluster validity index and significance analysis, in: Adaptive and Natural Computing Algorithms, Springer, 2009, pp. 313–322.