mathematics

MDPI

Article

An Enhanced Discriminant Analysis Approach for Multi-Classification with Integrated Machine Learning-Based Missing Data Imputation

Autcha Araveeporn [1,*] and Atid Kangtunyakarn [2]

[1] Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
[2] Department of Mathematics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand; atid.ka@kmitl.ac.th
* Correspondence: autcha.ar@kmitl.ac.th

Abstract

This study addresses the challenge of accurate classification under missing data conditions by integrating multiple imputation strategies with discriminant analysis frameworks. The proposed approach evaluates six imputation methods (Mean, Regression, KNN, Random Forest, Bagged Trees, MissRanger) across several discriminant techniques. Simulation scenarios varied in sample size, predictor dimensionality, and correlation structure, while the real-world application employed the Cirrhosis Prediction Dataset. The results consistently demonstrate that ensemble-based imputations, particularly regression, KNN, and MissRanger, outperform simpler approaches by preserving multivariate structure, especially in high-dimensional and highly correlated settings. MissRanger yielded the highest classification accuracy across most discriminant analysis methods in both simulated and real data, with performance gains most pronounced when combined with flexible or regularized classifiers. Regression imputation showed notable improvements under low correlation, aligning with the theoretical benefits of shrinkage-based covariance estimation. Across all methods, larger sample sizes and high correlation enhanced classification accuracy by improving parameter stability and imputation precision.

Keywords: Bagged Trees; discriminant analysis; KNN; MissRanger

MSC: 62H30; 62J10; 62F40; 62P10; 62C99

Academic Editor: Anatoliy Swishchuk

Received: 20 September 2025
Revised: 15 October 2025
Accepted: 23 October 2025
Published: 24 October 2025

Citation: Araveeporn, A.; Kangtunyakarn, A. An Enhanced Discriminant Analysis Approach for Multi-Classification with Integrated Machine Learning-Based Missing Data Imputation. *Mathematics* **2025**, *13*, 3392. https://doi.org/10.3390/math13213392

1. Introduction

Discriminant analysis (DA) is a multivariate statistical technique widely used in classification problems where the outcome variable is categorical and the predictor variables are quantitative. By constructing a discriminant function, a linear combination of predictor variables, DA aims to distinguish between two or more predefined groups. The method assumes multivariate normality and homogeneity of covariance matrices across groups, making it a powerful parametric alternative to logistic regression when these assumptions hold [1]. In medical research, DA has been applied to classification tasks, such as identifying the likelihood of disease occurrence based on clinical and laboratory measurements. Blood test results play a key role in improving the model's interpretability. Ramayah et al. [2] demonstrated the practical use of DA in classifying employees based on their intention to share knowledge. The model achieved an accuracy rate of over 85% and identified key

predictors, including attitude, subjective norm, and reciprocal relationships. The use of both highlights the model's predictive strength, underscoring DA's versatility and reliability across various research domains.

The core principle of discriminant analysis (DA) is to find linear or quadratic combinations of predictor variables that best separate predefined groups. In multi-class scenarios, this involves estimating class-specific parameters and deriving discriminant functions that maximize the separation between groups, assuming a multivariate normal distribution. Linear discriminant analysis (LDA), in particular, assumes homogeneity of covariance matrices across groups, leading to linear decision boundaries. In contrast, quadratic discriminant analysis (QDA) allows for group-specific covariance structures, resulting in more flexible and nonlinear boundaries. Several studies have explored the implementation and comparative performance of LDA and QDA under various conditions. Singh and Gupta [3] discussed an implementation framework for LDA and QDA, Chatterjee and Das [4] compared LDA, QDA, and support vector machine, and Berrar [5] provided a tutorial on linear and quadratic classifiers.

Several advanced DA techniques have been developed to address the limitations of traditional methods with high-dimensional or incomplete data. Regularized discriminant analysis (RDA) [6,7] combines LDA and QDA through covariance regularization, improving performance in small-sample and multicollinear settings. Flexible discriminant analysis (FDA) [8] extends LDA using optimal scoring and nonparametric regression, enabling more complex decision boundaries. Mixture discriminant analysis (MDA) [9,10] models each class as a mixture of Gaussian components, enhancing classification for multimodal datasets. Kernel discriminant analysis (KDA) [11] applies the kernel trick to achieve nonlinear separation in high-dimensional feature spaces. Shrinkage discriminant analysis (SDA) [12] stabilizes covariance estimates in high-dimensional problems through shrinkage toward structured targets. When combined with machine learning-based imputation methods [13,14], these approaches offer robust and flexible frameworks for multi-class classification in diverse applications.

Incomplete data is a pervasive issue across many real-world datasets, often arising from non-responses to surveys, equipment failures, or data entry errors. If left unaddressed, such missingness can lead to biased parameter estimates, reduced statistical power, and flawed conclusions. This issue is particularly critical in classification problems, where the quality of training data directly impacts model performance. Traditional strategies, such as listwise deletion, pairwise deletion, and single-value imputation, are simple to implement but suffer from significant limitations. Deletion methods reduce the adequate sample size and introduce bias when data are not missing completely at random [15]. At the same time, single-imputation methods often fail to retain multivariate relationships, underestimating variability and thereby weakening the classifier's performance.

The updated literature highlights the considerable consequences of missing data on model validity and reliability, especially in medical and clinical research domains [16]. Kang [17] emphasized that improper handling of missing data could distort parameter estimates and compromise study outcomes, particularly in hypothesis testing and prediction. To address these challenges, advanced imputation techniques such as multiple imputation, regression-based methods, hot-deck imputation, and probabilistic models have been proposed to preserve data structure and reduce estimation bias. As noted by Agiwal and Chaudhuri [18], appropriate handling of missing data not only mitigates biases but also enhances the representativeness and generalizability of findings in statistical modeling.

Ongoing research efforts have also focused on enhancing DA performance through advanced imputation strategies in incomplete or noisy datasets. Palanivinayagam and Damaševičius [19] employed SVM regression for missing value imputation, improving

classification accuracy in diabetes detection. Khashei et al. [20] developed soft computing and ensemble-based imputation strategies for pattern classification under missing data. Sharmila et al. [21] provided a comprehensive review of imputation methods, discussing their strengths, limitations, and suitability for different missing data mechanisms. Together, these approaches provide a robust and flexible framework for multi-class classification in diverse application domains.

The integration of machine learning-based imputation with statistical classification frameworks has emerged as a promising approach in data analysis, particularly for multi-class problems where handling missing data is crucial. Rácz and Gere [22] compared various imputation methods, including KNN, lasso regression, and Bayesian approaches, showing that performance depends heavily on data type and structure. Hong et al. [23] demonstrated that machine learning-based imputation enhances classification accuracy in diabetes prediction when combined with decision-tree models. Bai et al. [24] developed an autoencoder-based imputation method integrated with deep learning classifiers, achieving strong results in medical datasets with high missingness.

The developments in missing-data imputation have increasingly emphasized the integration of statistical and machine learning frameworks to enhance predictive performance and data reliability. In addition, van Buuren [25] synthesizes recent advances and provides practice-oriented guidance via the MICE (Multiple Imputation by Chained Equations) framework for obtaining unbiased estimates and valid confidence intervals, further underscoring the practical relevance of multiple imputation in applied settings. Audigier et al. [26] proposed a comprehensive multiple imputation framework for multilevel data with continuous and binary variables, demonstrating its ability to preserve hierarchical data structures and reduce estimation bias in complex datasets. Similarly, Resche-Rigon et al. [27] and Zhang et al. [28] extended these ideas by incorporating flexible modeling strategies and computationally efficient algorithms for large-scale applications. These advances underscore the increasing importance of integrating multiple imputation with machine learning frameworks to enhance classification accuracy, robustness, and interpretability in incomplete data environments. Furthermore, Zhang and Li [29] proposed a GAN-based imputation approach for multivariate time-series data, demonstrating improved reconstruction accuracy for complex temporal dependencies. Similarly, Park et al. [30] introduced a hybrid model combining MICE with variational autoencoders, which effectively balances statistical interpretability with nonlinear feature learning.

Previous studies on discriminant analysis and missing-data imputation have primarily focused on developing individual algorithms or conducting pairwise comparisons within specific settings. For example, Friedman [6] introduced the concept of regularized discriminant analysis, while Hastie et al. [8] proposed flexible discriminant analysis by optimal scoring. Later, Schäfer and Strimmer [31] and Stekhoven and Bühlmann [32] advanced covariance shrinkage and nonparametric imputation methods, respectively. However, these studies did not comprehensively integrate multiple imputation strategies with a wide range of discriminant analysis methods under different missing-data mechanisms, nor did they evaluate their performance in multi-class classification contexts.

The novelty of this study lies in the development of an integrated simulation framework that systematically combines six discriminant analysis techniques (LDA, RDA, FDA, MDA, KDA, SDA) with six machine-learning-based imputation methods (Mean, Regression, KNN, Random Forest, Bagged Trees, and MissRanger). This framework enables an in-depth examination of the interaction between imputation quality and classifier flexibility under various missing-data patterns. In addition, the study evaluates performance not only through accuracy but also using Cohen's kappa, confidence intervals, and standard deviations to statistically characterize model reliability. The proposed framework, therefore,

extends beyond traditional empirical benchmarks by providing methodological insight into how imputation methods influence the performance of discriminant analysis in realistic multi-class scenarios.

To provide a clear overview of the study, the structure of this paper is organized as follows: Section 1: Introduction presents the background, motivation, and importance of integrating imputation with classification methods. Section 2: Methodology outlines the discriminant analysis techniques employed in this study, including Linear, Regularized, Flexible, Mixture, and Shrinkage Discriminant Analysis, along with the imputation strategies used. Section 3: The Simulation Study and Results report the outcomes under varying sample sizes, correlation levels, and proportions of missingness. Section 4: Results of Actual Data demonstrate the application of the proposed framework to real clinical data. Section 5: Discussion interprets the findings and evaluates the strengths and limitations of each method. Finally, Section 6: Conclusion summarizes the key contributions and provides suggestions for future research.

2. Methodology

This section outlines the discriminant analysis techniques utilized in this study, which include linear discriminant analysis (LDA), regularized discriminant analysis (RDA), flexible discriminant analysis (FDA), mixture discriminant analysis (MDA), kernel discriminant analysis (KDA), and shrinkage discriminant analysis (SDA). These methods were selected based on their complementary properties in handling linearity, flexibility, regularization, and high-dimensionality in classification tasks. Each technique is briefly described below, along with the general framework for implementation.

2.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a statistical method for extracting the most informative features from data by removing redundant and noisy components [33] that best separate two or more classes. It assumes that the data from each class is normally distributed with a standard covariance matrix. The method projects the data into a lower-dimensional space where the separation between classes is maximized. The discriminant function is obtained by optimizing the ratio of between-class variance to within-class variance.

LDA is derived from Bayes' theorem by letting $\mathbf{x} \in \mathbb{R}^p$ denote a p-dimensional predictor variable, and suppose it belongs to one of K classes $\omega_1, \omega_2, \ldots \omega_k$. The classification goal is to assign $\mathbf{x}$ to the most probable class given the observed data to maximize the posterior probability $P(\omega_k|\mathbf{x})$.

Bayes' Theorem gives the posterior probability of class membership as

$$P(\omega_k|\mathbf{x}) = \frac{f_k(\mathbf{x}) \cdot P(\omega_k)}{f(\mathbf{x})}, \tag{1}$$

where $f_k(\mathbf{x})$ is the class-conditional density of $\mathbf{x}$ given class ω_k, $P(\omega_k)$ is the prior probability of class ω_k, and $f(\mathbf{x}) = \Sigma\, f_j(x)\, P(\omega_j)$ is the marginal density of $\mathbf{x}$. Since $f(\mathbf{x})$ is constant across classes for a given $\mathbf{x}$, the Bayes classifier assigns $\mathbf{x}$ to the class that maximizes

$$\delta(\mathbf{x}) = \underset{k}{\operatorname{argmax}}[\log f_k(\mathbf{x}) + \log P(\omega_k)]. \tag{2}$$

Assume each class ω_k follows a multivariate normal distribution: $\mathbf{x}|\omega_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, with shared covariance matrix $\boldsymbol{\Sigma}$ across all classes.

Then the class-conditional density is

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right). \tag{3}$$

Taking the logarithm of the posterior probability and simplifying yields the linear discriminant function:

$$g_k(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log P(\omega_k).$$

The decision rule assigns $\mathbf{x}$ to the class with the highest discriminant score if $g_k(\mathbf{x}) = \max_j g_j(\mathbf{x})$. The decision boundary between any two classes ω_i and ω_j is defined by $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$, which leads to the linear equation as

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = \frac{1}{2}\left(\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right) + \log\left(\frac{P(\omega_i)}{P(\omega_j)}\right). \tag{4}$$

The union of all such pairwise boundaries determines the partitioning of the input space into K decision regions for multi-class classification.

2.2. Regularized Discriminant Analysis

Regularized discriminant analysis (RDA), first proposed by Friedman [6], serves as a flexible extension of LDA by relaxing the strict homoscedasticity assumption of equal covariance matrices across groups. While LDA assumes that each class ω_k shares a standard covariance matrix $\boldsymbol{\Sigma}$, RDA introduces a regularization framework that allows the discriminant function to interpolate LDA, which permits class-specific covariance matrices $\boldsymbol{\Sigma}_k$.

The derivation of RDA begins with Bayes' decision rule, which assigns a new observation $\mathbf{x} \in \mathbb{R}^p$ to the class ω_k that maximizes the posterior probability $P(\omega_k|\mathbf{x})$. Under the assumption of multivariate normality, the discriminant function takes the form of

$$g_k(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log P(\omega_k).$$

In LDA, the covariance matrices $\boldsymbol{\Sigma}_k$ are replaced with the pooled covariance matrix $\boldsymbol{\Sigma}$, resulting in linear decision boundaries. However, when the assumption of equal covariance is violated, the LDA classifier can suffer from poor classification performance.

RDA mitigates the limitations of LDA by introducing a regularized covariance estimator defined as

$$\Sigma_k^{(\lambda)} = \lambda\Sigma_k + (1-\lambda)\Sigma \text{ where } 0 \le \lambda \le 1. \tag{5}$$

When $\lambda \to 1$, RDA approaches LDA with linear decision boundaries; when $\lambda \to 0$, it approaches QDA with quadratic boundaries.

The second parameter γ further shrinks $\Sigma_k^{(\lambda)}$ toward its diagonal form to control overfitting. An additional regularization parameter $\gamma \in [0,1]$ allows further shrinkage toward a diagonal matrix, promoting numerical stability and addressing the curse of dimensionality:

$$\Sigma_k^{(\lambda,\gamma)} = \gamma\Sigma_k^{(\lambda)} + (1-\gamma)\cdot \mathrm{diag}(\Sigma_k^{(\lambda)}). \tag{6}$$

Smaller γ values smooth the boundaries by reducing inter-variable correlations, yielding more regularized and stable discriminant functions. Consequently, (λ, γ) together define a continuum between fully flexible quadratic models and stable linear boundaries, adapting the classifier to varying dimensionality and correlation structures.

The final RDA discriminant function becomes

$$g_k(\mathbf{x}) = -\frac{1}{2}\log\left|\Sigma_k^{(\lambda,\gamma)}\right| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\left[\Sigma_k^{(\lambda,\gamma)}\right]^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log P(\omega_k). \tag{7}$$

Through the parameters λ and γ, RDA offers a continuum of classifiers ranging from LDA to diagonal-based models. This flexibility makes RDA particularly effective in settings where model complexity must be carefully balanced against sample size and dimensionality.

2.3. Flexible Discriminant Analysis

Flexible discriminant analysis (FDA), introduced by Hastie et al. [8], extends the classical LDA by allowing nonlinear relationships between predictors and class labels through basis expansions. While LDA assumes that class-conditional distributions are multivariate normal with common covariance matrices, FDA relaxes this linearity constraint by projecting the predictors into a higher-dimensional feature space.

Let $\mathbf{x} \in \mathbb{R}^p$ represent the predictor vector for the i-th observation, where $i = 1, \ldots, n$. The first step of the FDA involves transforming each predictor vector into a higher-dimensional space using a set of basis functions $\phi(x_i) = [\phi_1(x_i), \phi_2(x_i), \ldots, \phi_M(x_i)]^\top$, where M is the number of basis functions, and the choice of basis functions determines the model's flexibility. Applying this transformation to all n observations yield the design matrix

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(x_1)^\top \\ \phi(x_2)^\top \\ \vdots \\ \phi(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n\times M}.$$

Given a multi-class problem with k distinct classes, the response variable $y_i \in \{1, 2, \ldots, k\}$ is encoded using a class indicator matrix $\mathbf{G} \in \mathbb{R}^{n\times k}$, where

$$G_{ik} = \begin{cases} 1, & \text{if } y_i = k \\ 0, & \text{otherwise} \end{cases}.$$

This binary encoding represents the class membership of each observation.

The core of the FDA lies in the application of optimal scoring to transform the categorical outcome variable into a continuous score matrix $\mathbf{Y} \in \mathbb{R}^{n\times(K-1)}$. The goal is to find both $\mathbf{Y}$ and a coefficient matrix $\mathbf{B} \in \mathbb{R}^{M\times(k-1)}$ that minimize the Frobenius norm of the residuals in a multivariate regression setting:

$$\min_{\mathbf{Y},\mathbf{B}} \|\mathbf{Y} - \boldsymbol{\Phi}\mathbf{B}\|_F^2 \quad \text{subject to } \mathbf{Y}^\top\mathbf{Y} = \mathbf{I}_{k-1}, \tag{8}$$

where $\mathbf{I}_{k-1}$ is the identity matrix of size $k-1$. The orthonormality constraint ensures that the optimal scores are uncorrelated and have unit variance, which facilitates clear separation of classes in the reduced space. Once the optimal score matrix $\mathbf{Y}$ is determined, the coefficient matrix $\mathbf{B}$ is estimated via multivariate least squares:

$$\hat{\mathbf{B}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{Y}. \tag{9}$$

To classify a new observation $\mathbf{x}_{\text{new}}$, the basis-expanded vector is computed. This vector is then projected onto the discriminant space using the estimated coefficients:

$$\hat{\mathbf{y}}_{\text{new}} = \phi(\mathbf{x}_{\text{new}})^\top\hat{\mathbf{B}}. \tag{10}$$

The predicted class is assigned by comparing $\hat{\mathbf{y}}_{\text{new}}$ to the centroids of the training samples in the discriminant space, typically using a nearest centroid rule or a Gaussian-based classifier.

2.4. Mixture Discriminant Analysis

Mixture discriminant analysis (MDA) [9] models each class as a mixture of multivariate normal distributions, rather than assuming a single Gaussian component per class. This allows for more flexible modeling of class distributions that may be multimodal or heterogeneous. The model is typically fitted using the Expectation-Maximization (EM) algorithm. MDA is particularly effective in situations where class conditional densities deviate from unimodal assumptions.

This limitation arises from modeling each class as a finite mixture of Gaussian distributions [34]. This enables MDA to accommodate complex, multimodal class structures, providing a more flexible approach to classification than LDA.

Let denote an p-dimensional random vector, and let $Y \in \{1, \ldots, k\}$ represent the class label. Under the MDA framework, the class-conditional density $f_k(\mathbf{x})$ is expressed as a Gaussian mixture as $\mathbf{x}$

$$f_k(\mathbf{x}) = \sum_{r=1}^{R_k} \pi_{kr} \mathcal{N}(\mathbf{x}; \mu_{kr}, \Sigma_{kr}), \tag{11}$$

where π_{kr} are the mixing proportions $\left(\sum_{r=1}^{R_k} \pi_{kr} = 1 \right)$, $\mu_{kr} \in \mathbb{R}^p$ and $\Sigma_{kr} \in \mathbb{R}^{p \times p}$ are the mean and covariance of the component r in class k, $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ denotes the multivariate normal density.

The posterior probability that observation $\mathbf{x}$ belongs to class k is

$$P(Y = k \mid \mathbf{x}) = \frac{P(Y = k) f_k(\mathbf{x})}{\sum_{j=1}^{K} P(Y = j) f_j(\mathbf{x})}. \tag{12}$$

The classification rule assigns $\mathbf{x}$ to the class

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, P(Y = k \mid \mathbf{x}) \tag{13}$$

Estimation of the mixture parameters is typically carried out using the Expectation-Maximization (EM) algorithm. For each class k, the EM steps iterate as follows: E-step: $\gamma_{i,kr}^{(t)} = \frac{\pi_{kr}^{(t-1)} \mathcal{N}(x_i; \mu_{kr}^{(t-1)}, \Sigma_{kr}^{(t-1)})}{\sum_{s=1}^{R_k} \pi_{ks}^{(t-1)} \mathcal{N}(x_i; \mu_{ks}^{(t-1)}, \Sigma_{ks}^{(t-1)})}$, M-step: $\pi_{kr}^{(t)} = \frac{1}{n_k} \sum_{i:y_i=k} \gamma_{i,kr}^{(t)}$, $\mu_{kr}^{(t)} = \frac{\sum_{i:y_i=k} \gamma_{i,kr}^{(t)} x_i}{\sum_{i:y_i=k} \gamma_{i,kr}^{(t)}}$, $\Sigma_{kr}^{(t)} = \frac{\sum_{i:y_i=k} \gamma_{i,kr}^{(t)} (x_i - \mu_{kr}^{(t)})(x_i - \mu_{kr}^{(t)})^T}{\sum_{i:y_i=k} \gamma_{i,kr}^{(t)}}$.

2.5. Kernel Discriminant Analysis

Kernel discriminant analysis (KDA) is an extension of the classical LDA that enables the modeling of nonlinear decision boundaries by employing the kernel trick to implicitly map data from the original input space into a high-dimensional reproducing Kernel Hilbert space $\mathbb{R}^p$ through a nonlinear transformation $\phi : \mathbb{R}^p \to F$. This transformation allows classes that are not linearly separable in the original space to become separable in the feature space without $\phi(\mathbf{x})$ computing explicitly.

In the multi-class classification setting with K classes, KDA aims to find projection directions in F that maximize the ratio of between-class scatter to within-class scatter, analogous to Fisher's criterion but implemented entirely in kernel space. Using the kernel trick, the inner products F are computed through a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{F}}$, with common choices including the linear, polynomial, and Gaussian radial basis function kernels [35].

The between-class scatter matrix in kernel space is constructed from the class mean vectors $\mathbf{m}_k^{\phi}$ for $k = 1, \ldots, K$ while the within-class scatter matrix captures deviations of each sample from its respective class mean. The optimal discriminant subspace is obtained by solving the generalized eigenvalue problem:

$$\mathbf{K}\mathbf{S}_B^K\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\mathbf{S}_W^K\mathbf{K}\boldsymbol{\alpha}, \tag{14}$$

where $\mathbf{K}$ is the kernel Gram matrix, $\boldsymbol{\alpha}$ is the coefficient vector. The between-class and within-class scatter matrices in kernel representation are written as $\mathbf{S}_B^K = \sum_{k=1}^{K} N_k(\mathbf{m}_k^{\phi} - \mathbf{m}^{\phi})(\mathbf{m}_k^{\phi} - \mathbf{m}^{\phi})^T$, and $\mathbf{S}_W^K = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \omega_k} (\phi(\mathbf{x}_i) - \mathbf{m}_k^{\phi})(\phi(\mathbf{x}_i) - \mathbf{m}_k^{\phi})^T$. The solution yields up to $\mathbf{K} - 1$ discriminant vectors that form the nonlinear projection space [36]. For classification, once the data are projected into the discriminant subspace, classification is based on discriminant scores. A new observation $\mathbf{x}$ is assigned to the class ω_k that maximizes

$$\hat{y} = \text{argmax}_{k \in \{1,\ldots,K\}} g_k(\mathbf{x}), \tag{15}$$

where $g_k(\mathbf{x})$ is the discriminant score for the class k computed in the kernel space.

2.6. Shrinkage Discriminant Analysis

Shrinkage discriminant analysis (SDA) is a modern extension of LDA that is specifically designed to handle high-dimensional data, particularly when the number of predictors is large compared to, or even exceeds, the number of observations. In such cases, the sample covariance matrix ($\boldsymbol{\Sigma}$) used in LDA becomes singular or unstable, which can severely impair classification performance. SDA addresses this issue through a process called shrinkage estimation. The discriminant function is defined as

$$g_k(\mathbf{x}) = \mathbf{x}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log P(\omega_k) \tag{16}$$

where $\boldsymbol{\mu}_k$ is the mean vector for class k, $P(\omega_k)$ is the prior probability of class k, and $\mathbf{x}$ is a new observation.

However, the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ is often poorly estimated or singular, leading to unreliable inverse estimates and overfitting. SDA addresses this by applying shrinkage estimation to the covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_{\text{shrink}} = (1 - \lambda)\hat{\boldsymbol{\Sigma}} + \lambda\mathbf{T}, \tag{17}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, $\mathbf{T}$ is a shrinkage target (often a scaled identity matrix $\mathbf{I}_p$), $\lambda \in [0, 1]$ is the shrinkage intensity parameter, which controls the trade-off between the sample estimate and the target matrix. The optimal λ is chosen to minimize the expected quadratic loss:

$$\lambda^* = \underset{\lambda}{\text{argmin}} \mathbb{E}\left[\|\hat{\boldsymbol{\Sigma}}_{\text{shrink}} - \boldsymbol{\Sigma}\|_F^2\right] \tag{18}$$

In this study, the shrinkage target $\mathbf{T}$ is defined as a scaled identity matrix ($\mathbf{T} = cI_p$), where c is the average of the variances of the predictors. This target assumes equal variances and zero covariances among features, providing a stable and interpretable regularization baseline. The use of a diagonal target is consistent with the framework proposed by Ledoit and Wolf [37] and Schäfer and Strimmer [38], which is effective in high-dimensional settings.

The shrinkage intensity λ controls the trade-off between bias and variance: larger λ values increase shrinkage toward $\mathbf{T}$, reducing estimation variance but introducing bias; smaller λ values retain more of the empirical covariance structure, reducing bias but potentially increasing variance. This balance directly influences the smoothness and flexibility of the discriminant boundaries. Empirically, moderate λ values yield stable classification performance, particularly when the predictor correlation is moderate to high.

This makes SDA both theoretically sound and computationally efficient. With the shrinkage estimator, the modified discriminant function becomes

$$g_k^{\text{SDA}}(\mathbf{x}) = \mathbf{x}^\top \hat{\boldsymbol{\Sigma}}_{\text{shrink}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\Sigma}}_{\text{shrink}}^{-1} \hat{\boldsymbol{\mu}}_k + \log \hat{P}(\omega_k). \tag{19}$$

3. The Simulation Study and Results

To evaluate the performance of multiple discriminant analysis techniques, namely LDA, RDA, FDA, MDA, and SDA, in the presence of incomplete data, a comprehensive simulation study was conducted as discussed in the previous section. The design incorporated variations in the number of predictors, correlation structures, and missing data mechanisms.

The predictor variables were generated from a multivariate normal distribution with three dimensions: 5 and 10. For each dimension, two levels of pairwise correlation were considered: 0.3 and 0.7. The multivariate normal distribution is defined as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \left| \boldsymbol{\Sigma} \right|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right), \tag{20}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denote the p-dimensional predictor vector for the $i = 1, 2, \ldots, n$ as sample sizes n = 100, 300, and 500, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix controls the correlation among variables. Each vector was drawn independently from a multivariate normal distribution as $\mathbf{x}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbf{0}$ is a zero-mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix with a Toeplitz correlation structure. The (j, k) entry of $\boldsymbol{\Sigma}$ is defined as: $\Sigma_{jk} = \rho^{|j-k|}$, for $j, k = 1, \ldots, p$, which ensures a stationary correlation pattern where the correlation decreases exponentially with the distance between variable indices.

The Toeplitz covariance structure was adopted because it models a stationary correlation pattern, where correlations between predictors decay exponentially with their index distance. This structure is both mathematically tractable and empirically realistic, as it approximates dependence patterns found in temporal, spatial, and biological data. Unlike identity or block-diagonal matrices, the Toeplitz form maintains positive definiteness while allowing controlled variation in correlation strength.

Furthermore, adopting the Toeplitz structure enhances the comparability and reproducibility of simulation settings commonly used in multivariate studies, such as the studies by Schäfer and Strimmer [31] and Ledoit and Wolf [37]. By systematically varying ρ = 0.3 and 0.7, this study evaluates model performance across weak, moderate, and strong dependence scenarios, ensuring that the conclusions remain generalizable to a wide range of real-world data contexts.

The outcome variable y is a four-class categorical variable derived from a latent variable: $z_i = 1 + \sum_{j=1}^{p} x_{ij},\ i = 1, \ldots, n$, which is transformed using the logistic function.

The logistic transformation was employed to map the latent continuous variable z_i into class probabilities because it provides a smooth and symmetric link between the latent space and the (0, 1) probability interval:

$$\mathrm{pr}_i = \frac{1}{1 + \exp(-z_i)}. \tag{21}$$

This function ensures monotonicity and interpretability of thresholds when defining categorical outcomes while maintaining numerical stability during simulation. Logistic thresholds are commonly used in simulation-based classification studies, for instance, the studies by Hastie et al. [8] and Bai et al. [11] due to their probabilistic interpretability and analytical simplicity.

A four-class categorical response variable y_i was defined based on thresholds of pr_i as follows:

$$y_i = \begin{cases} 1, & \text{if } \mathrm{pr}_i \leq 0.25, \\ 2, & \text{if } 0.25 < \mathrm{pr}_i \leq 0.5, \\ 3, & \text{if } 0.5 < \mathrm{pr}_i \leq 0.75, \\ 4, & \text{if } \mathrm{pr}_i > 0.75. \end{cases}$$

The number of classes in the simulated categorical outcome was set to $K = 4$.

Theoretically, the dimensionality of the discriminant subspace in a K-class problem is at most $K - 1$. Therefore, using four classes yields three discriminant functions, which allows sufficient complexity for meaningful separation among groups while maintaining interpretability and computational efficiency. The detailed simulation setup is provided in Appendix A.

This choice balances model complexity with clarity of interpretation. It aligns with prior simulation frameworks in discriminant analysis, as evidenced by the studies of Hastie et al. [9] and Ahdesmäki and Strimmer [38]. Increasing the number of classes beyond four would introduce excessive overlap among groups and complicate the evaluation of classification boundaries. In contrast, fewer than four classes would limit the assessment of nonlinear and regularized discriminant effects.

To simulate real-world data imperfections, a mechanism for missing data was introduced. Specifically, for each dataset, 10% of the values in every predictor column were randomly replaced with missing values, following a Missing Completely at Random (MCAR) pattern. The missingness mechanism was designed to follow a strict Missing MCAR process of both observed (X_{obs}) and unobserved data (X_{mis}), defined as

$$P(R_{ij} = 1 \big| X_{obs}, X_{mis}) = P(R_{ij} = 1) = \pi,\ \forall i, j, \tag{22}$$

where R_{ij} is the missingness indicator for observation i, variable j, and $\pi = 0.1$ is the fixed missing proportion. This ensures that the probability of missingness is independent of both observed and unobserved data, satisfying the formal MCAR assumption.

Missing entries were generated using random draws from a uniform distribution $U(0, 1)$, where an element was set to missing if the random number exceeded the 0.90 quantile. This procedure guarantees that missingness occurs purely at random across all variables and simulation replications, eliminating potential bias in parameter estimation or classification performance due to the missingness mechanism.

3.1. Statistical Methods

3.1.1. Mean Imputation

In mean imputation, the missing values in a variable are replaced by the arithmetic mean of the observed values in that variable. Formally, let $X = (x_1, x_2, \ldots, x_n)$ be a variable with $m < n$ observed values. The mean imputed value is defined as

$$\overline{x}_{\text{obs}} = \frac{1}{m}\sum_{i=1}^{m} x_i. \tag{23}$$

Then, for each missing entry x_j, where $x_j = \text{NA}$, we replace $x_j^* = \overline{x}_{\text{obs}}$.

3.1.2. Regression Imputation

The fundamental idea behind regression imputation is to estimate the missing values of a variable using a regression model based on other observed variables. Suppose the data matrix is $X = (x_{ij}) \in \mathbb{R}^{n\times p}$, and let x_j be the variable containing missing values [39]. For each missing observation in x_j. The model estimation from observed data is to fit a regression model using X_{-j} (all other variables) as predictors and the observed part of X_j as the outcome variable

$$x_j = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_p x_p. \tag{24}$$

The next step is to predict the missing values by using the fitted model:

$$\hat{x}_{ij} = \hat{\beta}_0 + \sum_{k\neq j} \hat{\beta}_k x_{ik}. \tag{25}$$

This approach preserves multivariate relationships and allows the imputed values to vary across observations, unlike mean or mode imputation [40].

3.2. Machine Learning-Based Methods

In all experiments, the imputation algorithms were implemented in R using consistent parameter settings to ensure comparability. Specifically, MissRanger employed 500 trees with predictive mean matching (k = 5) and up to 10 iterations; Random Forest imputation used 500 trees and a maximum of 10 iterations; Bagged Tree imputation used 100 bootstrap samples with regression trees as base learners; KNN imputation used k = 5 with Euclidean distance; and regression imputation was based on linear models fitted to complete predictor sets. These configurations were chosen based on preliminary tuning to balance computational efficiency and predictive performance.

3.2.1. K-Nearest Neighbors (KNN) Imputation

The KNN imputation method is based on the idea that observations with similar attributes to their neighbors tend to have similar values. For each instance with missing data, the process identifies the k most similar neighbors from the dataset using a predefined distance metric [41].

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote an observation vector with some missing values. The KNN imputation procedure involves three main steps. First, for an observation $\mathbf{x}_i$ with a missing value, the algorithm computes the distance, most commonly the Euclidean distance, to all other observations using only the features that are observed in both. Second, it identifies the set $\mathcal{N}_k(\mathbf{x}_i)$ of the k nearest neighbors based on these distances. Finally, for each missing feature x_{ij}, the imputed value $\hat{x}_{ij}$ is calculated as the mean of the corresponding

values from the k nearest neighbors as $\hat{x}_{ij} = \frac{1}{k} \sum_{\mathbf{x}_l \in \mathcal{N}_k(\mathbf{x}_i)} x_{lj} x^i$. This approach leverages local similarity to estimate plausible values for the missing entries.

3.2.2. Random Forest Imputation

Random Forest imputation relies on the idea of building multiple decision trees to estimate the missing values. Each tree is trained on a bootstrap sample of the data, and predictions from the trees are aggregated to produce an imputed value [32].

Let $X = (x_{ij}) \in \mathbb{R}^{n \times p}$ be a data matrix with missing entries. The random forest imputation procedure consists of the following steps. First, missing values are initialized using simple methods such as mean imputation for continuous variables. Next, for each variable X_j that contains missing values, the algorithm treats X_j as the response and uses the remaining variables X_{-j} as predictors to train a random forest model on the subset of observations where X_j is observed. The trained model is then used to predict the missing entries in X_j [42]. This process is iterated across all variables with missing data, and the entire cycle is repeated until the changes in imputed values between iterations fall below a predefined tolerance or a maximum number of iterations is reached. Formally, for a missing value in a variable X_j for observation i, the imputed value is given by

$$\hat{x}_{ij} = \frac{1}{T} \sum_{t=1}^{T} h_t(X_{i,-j}), \tag{26}$$

where $h_t(\cdot)$ denotes the prediction from the t-th tree in the random forest. This approach captures complex nonlinear interactions and is well-suited for datasets with mixed variable types.

3.2.3. Bagged Trees Imputation

The Bagged Trees or Bootstrap-Aggregated Trees imputation is an ensemble learning technique, specifically the bootstrap aggregation trees method, applied to decision trees [43]. The core idea is to model the variable with missing values as a function of the other observed variables, using an ensemble of regression or classification trees trained on bootstrap samples.

For a variable X_j with missing data, the Bagged Trees imputation imputes missing values using an ensemble of decision trees through the following steps. First, multiple bootstrap samples are generated from the rows where X_j is observed. Then, for each bootstrap sample, a classification tree is trained using the remaining variables X_{-j} as predictors. Once the models are trained, each missing value in X_j The predictions are made using all trees, and then they are aggregated. For continuous variables, the imputed value is the average of forecasts:

$$\hat{x}_{ij} = \frac{1}{B} \sum_{b=1}^{B} h_b(X_{i,-j}), \tag{27}$$

where $h_b(\cdot)$ represents the prediction from the b-th tree. This ensemble approach stabilizes predictions, reduces variance, and avoids assumptions about linearity or distributional form, making it robust and flexible for various data types [44].

3.2.4. MissRanger Imputation

The missRanger algorithm performs iterative imputation using Random Forests, similar in concept to the missForest method [32]. Still, it replaces the random forest backend with ranger [45], which is optimized for high-dimensional datasets and fast execution. In addition, MissRanger supports predictive mean matching to preserve the distributional properties of imputed values.

The imputation procedure using missRanger begins with an initialization step, where missing values are filled with simple estimates such as the mean for continuous variables or the mode for categorical ones. Then, for each variable X_j with missing values, a random forest is fitted using the observed part of X_j as the response and all other variables X_{-j} as predictors. The fitted model is then used to predict the missing entries of X_j. If predictive mean matching is enabled, each predicted value is matched to the nearest observed value from a donor pool to better preserve variability, so that $\hat{x}_{ij} = x_{dj}$, where $x_{dj} \in D_{PMM}$, where set D_{PMM} is the observed values in X_j whose model-predicted values are closest to $\hat{x}_{ij}$. The process iterates across all variables with missing data until convergence is achieved or a maximum number of iterations is reached. Predictive Mean Matching (PMM) was used within the MissRanger algorithm to ensure that imputed values remain within the observed data range. PMM operates by first predicting missing values through a linear regression model and then replacing each predicted value with an observed donor value whose predicted mean is closest to the fitted value:

$$\hat{x}_{ij} = x_{lj}, \text{ where } l = \underset{l'}{\operatorname{argmin}} |\hat{x}_{ij} - \hat{x}_{l'j}|.$$

This approach preserves the empirical distribution of the variable and prevents implausible imputations. However, in high-dimensional settings, PMM may become less efficient due to instability in linear prediction and distance matching under multicollinearity.

In contrast, ensemble-based imputers utilize nonlinear models that aggregate multiple trees and bootstrap samples, thereby capturing complex variable interactions without relying on linearity assumptions. Such methods remain stable and accurate in high-dimensional data contexts, where PMM may suffer from increased bias or computational inefficiency.

These procedures were repeated 1000 times to reduce random sampling variability and to obtain stable estimates of model performance. The classification outcomes from each repetition were summarized using a confusion matrix, which compares the predicted and actual class labels, as illustrated in Table 1.

Table 1. Confusion matrix illustrating the comparison between actual and estimated classes for multi-class classification.

Predicted Data ($\hat{y}$)	**Actual Data (y)**			
	Class 1	**Class 2**	**Class 3**	**Class 4**
Class 1	A_{11}	A_{12}	A_{13}	A_{14}
Class 2	A_{21}	A_{22}	A_{23}	A_{24}
Class 3	A_{31}	A_{32}	A_{33}	A_{34}
Class 4	A_{41}	A_{42}	A_{43}	A_{44}

Let A_{ij} denote the number of observations that belong to actual class j but are predicted as class i. For a classification problem with 4 classes, the total number of observations is given by

$$N = \sum_{i=1}^{4} \sum_{j=1}^{4} A_{ij}.$$

In each iteration, classification models were trained on the training set and evaluated on the testing set. The key performance metric used for comparison was classification accuracy, call the percentage accuracy, which is evaluated from Table 1:

$$\text{Percentage Accuracy} = \frac{\sum_{i=1}^{4} A_{ii}}{N} \times 100.$$

Cohen's kappa statistic (k) was employed to evaluate the level of agreement between the predicted ($\hat{y}$) and actual (y) class labels, accounting for agreement that could occur by chance. The computation is based on the confusion matrix shown in Table 1 and is defined as follows.

Observed agreement (P_0) is the proportion of correctly classified observations, which is computed as $P_0 = \frac{\sum_{i=1}^{4} A_{ii}}{N}$.

Expected agreement by chance (P_e) is the expected agreement, which represents the proportion of agreement that would occur randomly and is calculated as

$$P_e = \sum_{i=1}^{4} \left(\frac{\sum_{j=1}^{4} A_{ij}}{N} \times \frac{\sum_{j=1}^{4} A_{ji}}{N} \right).$$

Then, Cohen's kappa coefficient (k) as the final statistic is defined as

$$k = \frac{P_0 - P_e}{1 - P_e}.$$

A value of $k = 1$ indicates perfect agreement between the predicted and actual classes, $k = 0$ implies agreement equivalent to random chance, and $k < 0$ indicates disagreement worse than chance.

For each discriminant analysis method and imputation strategy, classification performance was evaluated across 1000 replications. In each iteration, both the percentage accuracy and Cohen's kappa coefficient were computed. The mean accuracy and mean kappa were obtained as the arithmetic averages of their respective values across all replications, providing a stable estimate of overall performance. The standard deviation (SD) was calculated to quantify the dispersion of the results, indicating the consistency of each method across simulations.

To assess the statistical reliability of the performance estimates, 95% confidence intervals (CIs) were constructed for both accuracy and kappa using the normal approximation formula:

$$CI_{95\%} = \bar{x} \pm 1.96 \times \frac{SD}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean, SD is the standard deviation, and $n = 1000$ denotes the number of replications. Narrow confidence intervals indicate high stability of the classification outcomes, whereas wider intervals reflect greater variability across replications. This approach ensures that reported mean accuracies and kappa values are not only representative but also statistically reliable. All data analyses and simulation experiments were performed using the R 4.2.1 statistical software.

For each classification method (LDA, RDA, FDA, MDA, and SDA), the average classification accuracy and Cohen's kappa were computed over 1000 replications using a 70:30 train–test split. The analyses were conducted using the R packages kLAR, mda, sda, kernlab, mice, missRanger (version 2.1.1), caret (version 6.0-94), and MASS (version 7.3-60). The resulting mean percentage accuracies (with standard deviations), 95% confidence intervals, and mean Cohen's kappa values served as the basis for performance comparison across different imputation methods, discriminant analysis techniques, and simulation scenarios, as summarized in Tables 2–7.

Table 2. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using mean imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	73.90 (0.0804) 73.40–74.40 0.6057	77.86 (0.0422) 77.60–78.12 0.6646	79.48 (0.0332) 79.28–79.69 0.6879	72.15 (0.0836) 71.63–72.66 0.5679	77.47 (0.0434) 77.20–77.74 0.6356	78.69 (0.0333) 78.48–78.90 0.6529
	RDA	73.26 (0.0819) 72.75–74.40 0.5899	78.05 (0.0446) 77.77–78.32 0.6674	80.34 (0.0354) 80.12–80.56 0.7030	72.36 (0.0841) 71.84–72.88 0.5580	76.99 (0.0445) 76.71–77.27 0.6217	78.21 (0.0334) 78.01–78.42 0.6412
	FDA	73.73 (0.0801) 73.24–74.23 0.6048	77.90 (0.0423) 77.64–78.16 0.6656	79.50 (0.0333) 79.30–79.71 0.6885	71.88 (0.0843) 71.36–72.40 0.5659	77.40 (0.0438) 77.13–77.67 0.6352	78.71 (0.0332) 78.50–78.91 0.6536
	MDA	68.85 (0.0846) 68.32–69.37 0.5362	74.98 (0.0450) 74.70–75.26 0.6230	77.16 (0.0346) 76.95–77.38 0.6539	65.50 (0.0913) 64.93–66.07 0.4778	73.77 (0.0882) 73.47–74.07 0.5834	76.05 (0.0350) 75.83–76.26 0.6165
	KDA	71.06 (0.0761) 70.59–71.53 0.5381	75.72 (0.0431) 75.45–75.99 0.6264	78.26 (0.0327) 78.06–78.47 0.6678	74.78 (0.0770) 74.30–75.26 0.5661	76.98 (0.0434) 76.71–77.25 0.6068	77.63 (0.0324) 77.43–77.83 0.6214
	SDA	72.90 (0.0831) 72.38–73.14 0.5876	77.45 (0.0428) 77.18–77.72 0.6563	79.18 (0.0321) 78.98–79.38 0.6819	72.68 (0.0813) 72.18–73.18 0.5687	77.52 (0.0437) 77.25–77.79 0.6318	78.62 (0.0326) 78.42–78.82 0.6483
0.7	LDA	78.47 (0.0801) 77.97–78.96 0.6642	81.72 (0.0410) 81.47–81.98 0.7071	82.40 (0.0313) 82.20–82.59 0.7175	78.69 (0.0760) 78.21–79.16 0.6540	84.42 (0.0392) 84.17–84.66 0.7282	85.20 (0.0287) 85.02–85.37 0.7400
	RDA	80.24 (0.0793) 79.75–80.73 0.6977	83.74 (0.0412) 83.48–83.99 0.7474	84.96 (0.0303) 84.77–85.15 0.7669	79.63 (0.1085) 78.96–80.30 0.6619	85.01 (0.0384) 84.77–85.25 0.7452	85.83 (0.284) 85.65–86.00 0.7590
	FDA	78.35 (0.0807) 77.85–78.85 0.6676	81.81 (0.0411) 81.56–82.07 0.7093	82.47 (0.0313) 82.28–82.66 0.7191	78.40 (0.0766) 77.93–78.88 0.6521	84.35 (0.0392) 84.10–84.59 0.7275	85.19 (0.0286) 85.02–85.37 0.7403
	MDA	77.09 (0.0808) 76.59–77.59 0.6536	82.43 (0.0412) 82.17–82.69 0.7261	84.12 (0.0310) 83.93–84.31 0.7522	75.69 (0.0817) 75.18–76.20 0.6176	82.77 (0.0409) 82.52–83.03 0.7070	84.25 (0.0296) 84.07–84.44 0.7300
	KDA	79.86 (0.0748) 79.40–80.33 0.6748	83.92 (0.0400) 83.67–84.17 0.7463	85.86 (0.0274) 85.69–86.03 0.7790	84.47 (0.0630) 84.08–84.86 0.7226	85.56 (0.0365) 85.34–85.79 0.7432	85.74 (0.0272) 85.58–85.91 0.7487
	SDA	79.58 (0.0787) 79.09–80.06 0.6868	82.20 (0.0409) 81.94–82.45 0.7167	82.74 (0.0310) 82.55–82.93 0.7243	81.11 (0.0730) 80.66–81.57 0.6933	84.69 (0.0388) 84.45–84.94 0.7338	85.36 (0.0280) 85.19–85.54 0.7433

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 2 shows that at a low correlation level (ρ = 0.3), classification accuracy generally increases with larger sample sizes for all DA methods. Among the classifiers, LDA, RDA, FDA, KDA, and SDA consistently achieve higher mean accuracies, ranging from approximately 71% to 80%, while MDA performs relatively poorly with mean accuracies between 65% and 75%. The 95% confidence intervals for most methods are relatively narrow, indicating stable classification results across replications. The corresponding mean Cohen's kappa values range from 0.53 to 0.69, suggesting moderate classification agreement between predicted and actual group memberships.

When the correlation increases to ρ = 0.7, the classification performance improves notably across all methods. The RDA and KDA models show substantial gains, with mean accuracies exceeding 80% for larger sample sizes (n = 300 and 500). In particular, RDA achieves the highest overall accuracy (approximately 84–85%) and kappa values around 0.75–0.79, confirming the advantage of regularization under moderate correlation. SDA also

performs competitively, indicating that shrinkage-based covariance estimation contributes to improved stability in high-dimensional contexts.

Overall, the results demonstrate that increasing sample size and correlation level enhances model stability and discriminative performance. Regularized and flexible classifiers (RDA, KDA, SDA) yield higher accuracies and stronger kappa agreement compared with classical LDA and mixture-based MDA, highlighting the benefits of integrating covariance regularization and nonlinear transformations in discriminant analysis with mean-imputed data.

Table 3. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using regression imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	81.53 (0.0883) 80.99–82.08 0.7242	84.91 (0.0338) 84.70–85.12 0.7829	89.04 (0.0374) 88.81–89.28 0.8325	87.84 (0.0199) 87.71–87.96 0.8081	79.25 (0.0435) 78.98–79.53 0.6735	86.57 (0.0250) 86.41–86.72 0.7878
	RDA	77.93 (0.0435) 77.69–78.02 0.6651	86.48 (0.0257) 86.33–86.64 0.8049	89.55 (0.0500) 89.24–89.86 0.8408	82.73 (0.0075) 82.68–82.77 0.7175	78.54 (0.0248) 78.38–79.69 0.6563	85.90 (0.0251) 85.74–86.05 0.7763
	FDA	81.54 (0.0883) 80.99–82.09 0.7243	84.68 (0.0330) 84.48–84.89 0.7797	88.88 (0.0400) 88.63–89.12 0.8303	86.13 (0.0353) 85.91–86.35 0.7836	79.39 (0.0424) 79.13–79.65 0.6761	86.57 (0.0273) 86.40–86.74 0.7883
	MDA	75.22 (0.0440) 74.95–75.49 0.6358	84.68 (0.0476) 84.39–84.98 0.7775	84.51 (0.0446) 84.23–84.79 0.7629	79.28 (0.0053) 79.25–79.32 0.6776	74.47 (0.0243) 74.32–74.62 0.6000	83.66 (0.0137) 83.58–83.75 0.7439
	KDA	76.96 (0.0546) 76.62–77.33 0.6484	82.88 (0.0488) 82.58–83.18 0.7486	86.20 (0.0610) 85.82–86.58 0.7875	81.04 (0.0180) 80.93–81.15 0.6742	77.00 (0.0361) 76.78–77.23 0.6128	83.22 (0.0145) 83.13–83.31 0.7225
	SDA	80.62 (0.0735) 80.17–81.08 0.7115	85.81 (0.0232) 85.66–85.95 0.7944	88.38 (0.0394) 88.13–88.62 0.8219	84.41 (0.0195) 84.29–84.53 0.7479	78.55 (0.0458) 78.26–78.83 0.6557	85.90 (0.0190) 85.78–86.02 0.7749
0.7	LDA	83.79 (0.0289) 83.56–83.92 0.7799	87.40 (0.0061) 87.36–87.44 0.7896	79.23 (0.0049) 79.20–79.26 0.6850	84.57 (0.0750) 84.10–85.03 0.7473	83.22 (0.0217) 83.09–83.36 0.7182	84.44 (0.0298) 84.25–84.62 0.7353
	RDA	85.07 (0.0323) 84.87–85.27 0.8019	88.50 (0.0343) 88.29–88.71 0.8099	83.93 (0.0047) 83.90–83.96 0.7610	77.54 (0.1670) 76.51–78.58 0.5998	81.54 (0.0511) 81.22–81.86 0.7096	86.57 (0.0120) 86.50–86.65 0.7780
	FDA	85.02 (0.0347) 84.80–85.24 0.8106	87.40 (0.0059) 87.36–87.44 0.7896	79.23 (0.0048) 79.20–79.26 0.6850	84.57 (0.0750) 84.10–85.03 0.7481	83.22 (0.0218) 83.08–83.35 0.7181	84.44 (0.0316) 84.24–84.63 0.7356
	MDA	85.45 (0.0334) 85.25–85.66 0.8145	85.20 (0.0063) 85.16–85.24 0.7573	87.90 (0.0026) 87.89–87.92 0.8192	80.67 (0.0720) 80.22–81.12 0.6959	79.31 (0.0085) 79.26–79.37 0.6589	83.89 (0.0357) 83.67–84.11 0.7327
	KDA	85.05 (0.0287) 84.87–85.22 0.7808	89.62 (0.0229) 89.48–89.76 0.8269	86.60 (0.0030) 86.58–86.61 0.8004	87.61 (0.0539) 87.27–87.94 0.7725	82.11 (0.0038) 82.09–82.14 0.6927	83.91 (0.0340) 83.70–84.12 0.7240
	SDA	85.02 (0.0176) 84.91–85.13 0.7867	87.77 (0.0093) 87.71–87.83 0.7979	80.56 (0.0043) 80.54–80.59 0.7071	86.54 (0.0700) 86.10–86.97 0.7769	84.32 (0.0115) 84.25–84.39 0.7364	84.17 (0.0308) 83.98–84.36 0.7311

Note: The underlined letter indicates the highest mean percentage accuracy.

At the low correlation level (ρ = 0.3) from Table 3, all methods show improved classification accuracy with larger sample sizes, with LDA, RDA, and FDA achieving the highest accuracies (77–86% for p = 5 and up to 90% for larger n). MDA and KDA perform slightly lower, generally below 85%. The narrow 95% confidence intervals indicate stable performance, and Cohen's kappa values (0.66–0.78) suggest moderate to substantial agreement. At the moderate correlation level (ρ = 0.7), accuracies increase notably across

all DA methods. KDA and RDA again outperform others, exceeding 86% accuracy, while SDA and FDA also yield competitive results (85%). Kappa coefficients (0.73–0.81) indicate higher classification consistency. Overall, regression imputation provides strong and stable classification performance, particularly when combined with flexible or regularized classifiers such as RDA and FDA, which consistently outperform traditional LDA and MDA under both correlation settings.

Table 4. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using KNN imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	76.84 (0.0791) 76.35–77.34 0.6519	82.28 (0.0435) 82.01–82.55 0.7339	84.44 (0.0306) 84.25–84.63 0.7666	74.05 (0.0839) 73.53–74.57 0.5982	79.37 (0.0411) 79.11–79.62 0.6697	81.02 (0.0305) 80.83–81.21 0.6947
	RDA	76.21 (0.0794) 75.72–76.70 0.6382	83.20 (0.0457) 82.91–83.48 0.7464	86.08 (0.0322) 85.88–86.28 0.7910	74.31 (0.0834) 73.79–74.82 0.5893	78.97 (0.0425) 78.70–79.23 0.6589	80.57 (0.0318) 80.38–80.77 0.6849
	FDA	76.80 (0.0791) 76.31–77.29 0.6523	82.26 (0.0438) 81.99–82.53 0.7339	84.42 (0.0306) 84.23–84.61 0.7665	73.74 (0.0840) 73.22–74.26 0.5956	79.31 (0.0414) 79.06–79.57 0.6695	81.02 (0.0308) 80.83–81.21 0.6951
	MDA	72.92 (0.0819) 72.41–73.43 0.5970	79.79 (0.0454) 79.51–80.07 0.6977	82.11 (0.0336) 81.90–82.32 0.7323	69.06 (0.0894) 68.51–69.62 0.5291	75.96 (0.0449) 75.68–76.28 0.6209	78.55 (0.0338) 78.34–78.76 0.6592
	KDA	74.40 (0.0749) 73.93–74.86 0.5968	81.51 (0.0424) 81.25–81.78 0.7177	84.59 (0.0284) 84.41–84.77 0.7668	76.26 (0.0751) 75.80–76.73 0.5925	78.56 (0.0415) 78.30–78.82 0.6373	79.77 (0.0316) 79.57–79.96 0.6616
	SDA	76.25 (0.0796) 75.75–76.74 0.641	82.06 (0.0429) 81.79–82.33 0.7296	84.27 (0.0302) 84.08–84.46 0.7634	74.72 (0.0835) 74.20–75.24 0.6019	79.45 (0.0406) 79.19–79.70 0.6672	80.96 (0.0308) 80.76–81.15 0.6909
0.7	LDA	80.28 (0.0751) 79.82–80.75 0.6972	84.31 (0.0388) 84.07–84.55 0.7525	85.52 (0.0321) 85.32–85.72 0.7711	80.28 (0.0771) 79.80–80.76 0.6824	85.28 (0.0378) 85.04–85.51 0.7468	86.26 (0.0282) 86.09–86.44 0.7608
	RDA	82.74 (0.0776) 82.26–83.22 0.7365	87.40 (0.0376) 87.17–87.63 0.8051	89.34 (0.0330) 89.13–89.54 0.8351	81.50 (0.1031) 80.86–82.14 0.6916	86.45 (0.0373) 86.22–86.68 0.7729	87.46 (0.0283) 87.28–87.63 0.7888
	FDA	80.21 (0.0756) 79.74–80.68 0.6977	84.34 (0.0389) 84.10–84.58 0.7535	85.58 (0.0320) 85.38–85.78 0.7723	79.95 (0.0775) 79.47–80.43 0.6796	85.24 (0.0379) 85.00–85.47 0.7467	86.27 (0.0284) 86.09–86.44 0.7611
	MDA	80.25 (0.0760) 79.78–80.73 0.7020	85.95 (0.0356) 85.73–86.17 0.7830	87.62 (0.0295) 87.44–87.80 0.8088	77.96 (0.0806) 77.46–78.46 0.6545	83.99 (0.0391) 83.74–84.23 0.7309	85.63 (0.0301) 85.44–85.81 0.7562
	KDA	82.38 (0.0716) 81.94–82.83 0.7187	88.34 (0.0341) 88.13–88.55 0.8175	90.64 (0.0250) 90.48–90.79 0.8544	85.50 (0.0613) 85.12–85.88 0.7420	86.49 (0.0342) 86.27–86.70 0.7620	87.21 (0.0270) 87.04–87.38 0.7769
	SDA	81.48 (0.0737) 81.03–81.94 0.7184	84.74 (0.0380) 84.51–84.98 0.7607	85.91 (0.0312) 85.71–86.10 0.7780	82.61 (0.0702) 82.17–83.05 0.7207	85.72 (0.0374) 85.49–85.95 0.754	86.45 (0.0283) 86.28–86.63 0.7644

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 4 exhibits the classification accuracy using KNN. At the low correlation level (ρ = 0.3), all methods show improved accuracy with increasing sample size, with LDA, RDA, KDA, and SDA performing best (74–86% for p = 5 and around 81–82% for p = 10). Under moderate correlation (ρ = 0.7), accuracies further increase across all models, with RDA and KDA exceeding 85% and showing higher stability. The mean Cohen's kappa values (0.59–0.78) indicate moderate to substantial agreement, reflecting reliable classification consistency. Overall, KNN imputation provides robust and stable performance, especially when combined with flexible or regularized classifiers.

Table 5. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using random forest imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	77.06 (0.0810) 76.55–77.56 0.6547	82.31 (0.0426) 82.04–82.57 0.7354	84.81 (0.0323) 84.61–85.01 0.7729	74.33 (0.0802) 73.83–74.83 0.6009	79.59 (0.0412) 79.34–79.85 0.6736	80.91 (0.0312) 80.71–81.10 0.6938
	RDA	76.47 (0.0787) 75.98–76.96 0.6398	83.57 (0.0452) 83.29–83.85 0.7525	86.72 (0.0348) 86.51–86.94 0.8010	74.75 (0.0813) 74.25–75.26 0.5930	79.10 (0.0411) 78.84–79.35 0.6607	80.52 (0.0326) 80.32–80.73 0.6843
	FDA	76.81 (0.0820) 76.30–77.32 0.6522	82.31 (0.0429) 82.04–82.57 0.7356	84.79 (0.0324) 84.59–84.99 0.7728	74.10 (0.0812) 73.50–74.51 0.5983	79.57 (0.0411) 79.32–79.83 0.6739	80.89 (0.0314) 80.70–81.09 0.6939
	MDA	72.56 (0.0859) 72.03–73.09 0.5908	79.90 (0.0450) 79.63–80.18 0.7001	82.47 (0.0347) 82.26–82.69 0.7384	68.91 (0.0893) 68.36–69.46 0.5258	75.99 (0.0454) 75.71–76.27 0.6212	78.27 (0.0328) 78.07–78.48 0.6544
	KDA	74.34 (0.0779) 73.86–74.82 0.5934	81.40 (0.0411) 81.14–81.65 0.7172	84.90 (0.0297) 84.72–85.09 0.7722	76.45 (0.0757) 75.98–76.92 0.5920	78.54 (0.0405) 78.29–78.79 0.6366	79.56 (0.0314) 79.37–79.76 0.6594
	SDA	76.44 (0.0795) 75.95–76.94 0.6434	82.18 (0.0415) 81.92–82.43 0.7324	84.65 (0.0323) 84.45–84.85 0.7700	75.09 (0.0787) 74.60–75.58 0.6066	79.67 (0.0407) 79.42–79.93 0.6716	80.85 (0.0311) 80.65–81.04 0.6906
0.7	LDA	79.94 (0.0737) 79.48–80.39 0.6905	83.96 (0.0400) 83.71–84.21 0.7471	85.51 (0.0307) 85.32–85.70 0.7708	79.51 (0.0834) 78.99–80.02 0.6725	84.81 (0.0379) 84.57–85.05 0.7383	85.96 (0.0266) 85.80–86.13 0.7556
	RDA	82.42 (0.0725) 81.97–82.87 0.7330	86.85 (0.0388) 86.61–87.10 0.7968	89.30 (0.0325) 89.09–89.50 0.8344	79.69 (0.1201) 78.94–80.43 0.6624	86.01 (0.0362) 85.78–86.23 0.7650	87.06 (0.0270) 86.89–87.23 0.7825
	FDA	79.94 (0.0744) 79.47–80.40 0.6928	84.01 (0.0398) 83.76–84.26 0.7484	85.56 (0.0306) 85.37–85.75 0.7719	79.15 (0.0846) 78.62–79.67 0.6699	84.74 (0.0385) 84.50–84.98 0.7377	85.95 (0.0267) 85.78–86.11 0.7556
	MDA	79.78 (0.0738) 79.32–80.23 0.6953	85.31 (0.0398) 85.06–85.56 0.7732	87.54 (0.0299) 87.39–87.73 0.8076	77.39 (0.0822) 76.88–77.90 0.6486	83.22 (0.0404) 82.97–83.47 0.7181	85.11 (0.0287) 84.93–85.29 0.7478
	KDA	81.81 (0.0692) 81.38–82.24 0.7105	87.81 (0.0361) 87.59–88.04 0.8094	90.51 (0.0258) 90.35–90.67 0.8524	85.15 (0.0679) 84.73–85.57 0.7366	86.12 (0.0335) 85.91–86.33 0.7553	86.75 (0.0253) 86.59–86.91 0.7687
	SDA	81.28 (0.0729) 80.83–81.74 0.7148	84.42 (0.0392) 84.18–84.67 0.7555	85.88 (0.0301) 85.69–86.06 0.7773	81.68 (0.0784) 81.19–82.17 0.7067	78.30 (0.0374) 85.07–85.54 0.7471	86.11 (0.0263) 85.94–86.27 0.7583

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 5 presents the classification accuracy using random forest imputation. At the low correlation level (ρ = 0.3), all methods show increasing accuracy with larger sample sizes, with LDA, KDA, RDA, and SDA achieving the best results (76–86% for p = 5 and around 79–82% for p = 10). Under moderate correlation (ρ = 0.7), accuracies further improve, with KDA and RDA exceeding 87% and showing high stability. The mean Cohen's kappa values (0.59–0.78) indicate moderate to strong agreement. Overall, random forest imputation provides robust and consistent performance, particularly when combined with flexible or regularized classifiers.

Table 6 shows that at the low correlation level (ρ = 0.3), all methods show improved accuracy with larger sample sizes, with LDA, RDA, FDA, KDA, and SDA achieving the best results (73–78% for p = 5 and around 76–78% for p = 10). Under moderate correlation (ρ = 0.7), accuracies further increase, with RDA and KDA exceeding 83% and showing high stability. The mean Cohen's kappa values (0.53–0.78) indicate moderate to substantial

agreement. Overall, Bagged Tree imputation provides consistent and reliable performance, especially when combined with regularized or flexible discriminant classifiers.

Table 6. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using Bagged Trees imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	73.92 (0.0800) 73.42–74.42 0.6037	77.83 (0.0435) 77.56–78.10 0.6633	79.37 (0.0335) 79.16–79.58 0.6870	72.10 (0.0836) 71.49–72.53 0.5652	77.19 (0.0445) 76.92–77.47 0.6323	78.77 (0.0315) 78.57–78.96 0.6539
	RDA	73.28 (0.0809) 72.78–73.87 0.5872	78.00 (0.0450) 77.72–78.28 0.6655	79.88 (0.0343) 79.67–80.09 0.6958	72.29 (0.0883) 71.75–72.84 0.5577	76.85 (0.0450) 76.57–77.13 0.6218	78.19 (0.0321) 77.99–78.39 0.6412
	FDA	73.76 (0.0804) 73.26–74.26 0.6030	77.80 (0.0433) 77.53–78.07 0.6634	79.40 (0.0335) 79.19–79.60 0.6877	71.68 (0.0846) 71.16–72.21 0.5625	77.15 (0.0446) 76.88–77.43 0.6324	78.78 (0.0316) 78.58–78.97 0.6545
	MDA	69.16 (0.0847) 68.64–69.69 0.5394	74.73 (0.0463) 74.45–75.02 0.6188	76.85 (0.0338) 76.64–77.06 0.6503	65.47 (0.0901) 64.91–66.02 0.4772	73.56 (0.0460) 73.27–73.84 0.5816	75.93 (0.0329) 75.73–76.14 0.6146
	KDA	71.45 (0.0752) 70.98–71.91 0.5425	75.78 (0.0423) 75.52–76.04 0.6266	78.16 (0.0330) 77.96–78.37 0.6668	74.97 (0.0769) 74.50–75.45 0.5671	76.75 (0.0429) 76.48–77.01 0.6045	77.68 (0.0319) 77.48–77.88 0.6228
	SDA	73.58 (0.0794) 73.09–74.07 0.5963	77.61 (0.0435) 77.34–77.88 0.6587	79.16 (0.0337) 78.95–79.37 0.6829	73.00 (0.0831) 72.49–73.52 0.5742	77.31 (0.0438) 77.04–77.59 0.6312	78.75 (0.0315) 78.56–78.95 0.6515
0.7	LDA	78.44 (0.0771) 77.96–78.92 0.6646	82.19 (0.0411) 81.93–82.44 0.7168	83.32 (0.0314) 83.12–83.51 0.7346	79.53 (0.0763) 79.05–80.00 0.6705	84.42 (0.0381) 84.19–84.66 0.7307	85.59 (0.0291) 85.40–85.76 0.7477
	RDA	80.31 (0.0724) 79.82–80.80 0.6990	84.43 (0.0423) 84.16–84.69 0.7592	85.81 (0.0305) 85.62–86.00 0.7813	80.05 (0.1080) 79.37–80.72 0.6699	85.34 (0.0392) 85.10–85.59 0.7350	86.43 (0.0296) 86.25–86.62 0.7704
	FDA	78.23 (0.0769) 77.76–78.71 0.6638	82.24 (0.0412) 81.98–82.49 0.7183	83.36 (0.0313) 83.17–83.56 0.7356	79.15 (0.0770) 78.68–79.63 0.6669	84.34 (0.0383) 84.11–84.58 0.7299	85.57 (0.0291) 85.39–85.75 0.7477
	MDA	77.92 (0.0774) 77.44–78.40 0.6647	82.99 (0.0400) 82.74–82.24 0.7358	84.75 (0.0302) 84.56–84.93 0.7632	76.59 (0.0838) 76.07–77.12 0.6367	82.90 (0.0412) 82.65–83.16 0.7114	84.45 (0.0303) 84.26–84.64 0.7352
	KDA	80.19 (0.0721) 79.74–80.64 0.6822	84.94 (0.0391) 84.69–85.18 0.7628	87.04 (0.0282) 86.86–87.21 0.7977	84.75 (0.0652) 84.35–85.16 0.7280	85.68 (0.0364) 85.46–85.91 0.7464	86.26 (0.0281) 86.09–86.44 0.7586
	SDA	79.57 (0.0725) 79.12–80.02 0.6868	82.56 (0.0410) 82.31–82.82 0.7244	83.63 (0.0311) 83.44–83.82 0.7407	81.59 (0.0733) 81.13–82.05 0.7041	84.82 (0.0377) 84.58–85.05 0.7378	85.76 (0.0285) 85.58–85.93 0.7510

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 7 summarizes classification accuracy using MissRanger imputation. At the low correlation level (ρ = 0.3), all methods show improved accuracy with larger sample sizes, with LDA, RDA, and KDA achieving the best performance (76–87% for p = 5 and 89–91% for p = 10). Under moderate correlation (ρ = 0.7), accuracies further increase, with RD and KDA exceeding 90% and showing high stability. The mean Cohen's kappa values (0.65–0.84) indicate substantial agreement and reliable classification. Overall, MissRanger imputation provides the most accurate and stable results, particularly when combined with flexible or regularized classifiers.

The highest mean percentage accuracies corresponding to each missing data imputation method are extracted from Tables 2–7 and presented in Tables 8 and 9. These summary tables highlight the best-performing classification results for each imputation technique: mean imputation, regression (Reg) imputation, KNN imputation, Random Forest (RF) im-

putation, Bagged Trees (BT) imputation, and MissRanger imputation. Varying correlation levels organize the comparisons (ρ), the number of predictors (p), and sample sizes (n), allowing for a more precise assessment of which combinations of methods and conditions yield the most accurate classification outcomes.

Table 7. Mean percentage accuracy (standard deviation), 95% confidence interval, and mean Cohen's kappa for classification of DA methods using MissRanger imputation under different correlation levels (ρ), varying numbers of predictors (p), and sample sizes (n).

ρ	DA Methods	p = 5			p = 10		
		n = 100	n = 300	n = 500	n = 100	n = 300	n = 500
0.3	LDA	77.38 (0.0776) 76.90–77.87 0.6593	82.65 (0.0419) 82.39–82.91 0.7403	85.02 (0.0331) 84.82–85.23 0.7759	86.20 (0.0297) 86.01–86.38 0.7637	86.31 (0.0295) 86.13–86.49 0.7653	86.35 (0.0285) 86.17–86.52 0.7661
	RDA	76.82 (0.0785) 76.34–77.31 0.6460	83.80 (0.0436) 83.53–84.04 0.755	87.03 (0.0336) 86.82–87.24 0.8053	90.06 (0.0238) 89.91–90.21 0.8376	90.09 (0.0242) 89.94–90.24 0.8380	90.03 (0.0230) 89.89–90.17 0.8371
	FDA	77.32 (0.0780) 76.83–77.80 0.6596	82.64 (0.0420) 82.37–82.90 0.7404	85.03 (0.0330) 84.83–85.24 0.7762	86.27 (0.0296) 86.09–86.45 0.7653	86.42 (0.0295) 86.23–86.60 0.7675	86.43 (0.0283) 86.25–86.60 0.7678
	MDA	72.83 (0.0830) 72.32–73.35 0.5966	79.94 (0.0446) 79.67–80.22 0.7007	82.75 (0.0342) 82.54–82.96 0.7423	89.28 (0.0235) 89.13–89.42 0.8235	89.30 (0.0240) 89.15–89.45 0.8239	89.30 (0.0242) 89.15–89.45 0.8240
	KDA	74.49 (0.0760) 74.02–74.97 0.5960	81.58 (0.0415) 81.32–81.83 0.720	85.17 (0.0300) 84.98–85.35 0.7758	90.40 (0.0226) 90.26–90.54 0.8408	90.27 (0.0225) 90.13–90.41 0.8386	90.34 (0.0222) 90.21–90.49 0.8400
	SDA	76.71 (0.0783) 76.22–77.19 0.647	82.49 (0.0413) 82.24–82.75 0.7370	84.93 (0.0326) 84.73–85.14 0.7740	86.81 (0.0295) 86.62–86.99 0.7752	86.91 (0.0293) 86.73–87.09 0.7767	86.98 (0.0283) 86.81–87.16 0.7779
0.7	LDA	80.53 (0.0741) 80.07–80.99 0.6697	83.98 (0.0405) 83.72–84.23 0.7071	85.56 (0.0320) 85.36–85.75 0.7716	86.28 (0.0306) 86.09–86.47 0.7650	86.13 (0.0303) 85.94–86.32 0.7618	86.19 (0.0300) 86.01–86.38 0.7632
	RDA	82.69 (0.0776) 82.21–83.17 0.7345	86.80 (0.0400) 86.55–87.05 0.7961	89.19 (0.0327) 88.99–89.40 0.8331	90.09 (0.0244) 89.94–90.24 0.8379	90.04 (0.0243) 89.89–90.19 0.8371	90.10 (0.0244) 89.94–90.25 0.8381
	FDA	80.33 (0.0743) 79.87–80.79 0.6982	84.03 (0.0406) 83.78–84.28 0.7485	85.60 (0.0319) 85.40–85.80 0.7726	86.39 (0.0304) 86.20–86.58 0.7674	86.23 (0.0304) 86.04–86.42 0.7640	86.30 (0.0296) 86.12–86.49 0.7654
	MDA	80.54 (0.0761) 80.07–81.01 0.7062	85.25 (0.0385) 85.01–85.49 0.7724	87.47 (0.0290) 87.29–87.65 0.8068	89.27 (0.0245) 89.12–89.42 0.8234	89.16 (0.0250) 89.01–89.32 0.8214	89.22 (0.0251) 89.07–89.38 0.8225
	KDA	82.56 (0.0677) 82.14–82.98 0.7221	87.72 (0.0354) 87.50–87.94 0.8079	90.51 (0.0273) 90.34–90.68 0.8527	90.33 (0.0235) 90.12–90.48 0.8398	90.18 (0.0234) 90.03–90.32 0.8370	90.26 (0.0234) 90.11–90.40 0.8383
	SDA	81.52 (0.0718) 81.08–81.97 0.7183	84.41 (0.0399) 84.16–84.66 0.7553	85.87 (0.0316) 85.67–86.07 0.7774	86.89 (0.0303) 86.70–87.08 0.7766	86.79 (0.0302) 86.60–86.98 0.7742	86.83 (0.0296) 86.64–87.01 0.7751

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 8 presents the highest mean classification accuracies corresponding to each missing data imputation method for five predictors (p = 5) across different correlation levels (ρ = 0.3 and 0.7) and sample sizes (n = 100, 300, and 500). At the low correlation level (ρ = 0.3), classification accuracy improves consistently with increasing sample size. The Regression imputation (Reg) and RDA combination yields the highest accuracy overall, reaching 89.55% when n = 500. KNN and Random Forest (RF) also perform competitively, while simple mean and Bagged Tree (BT) imputations show comparatively lower results.

At the moderate correlation level (ρ = 0.7), accuracies increase markedly across all imputation methods. The best results are achieved by KDA combined with MissRanger,

yielding up to 90.51% accuracy at n = 500. Similarly, Regression and RF imputations also maintain strong performance above 87%.

Table 8. The highest mean percentage accuracies corresponding to each missing data imputation method for the five predictors.

Correlation Levels (ρ)	Sample Sizes (n)	Methods of Missing Imputation for p = 5					
		Mean	**Reg**	**KNN**	**RF**	**BT**	**MissRanger**
0.3	100	73.90 (LDA)	81.54 (FDA)	76.84 (LDA)	77.06 (LDA)	73.92 (LDA)	77.38 (LDA)
	300	77.90 (FDA)	86.48 (RDA)	82.28 (LDA)	83.57 (RDA)	77.83 (LDA)	83.80 (RDA)
	500	80.34 (RDA)	89.55 (RDA)	86.08 (RDA)	86.72 (RDA)	79.88 (RDA)	87.03 (RDA)
0.7	100	80.24 (RDA)	85.45 (MDA)	83.28 (KDA)	82.42 (RDA)	80.31 (RDA)	82.69 (RDA)
	300	83.92 (KDA)	89.62 (KDA)	88.34 (KDA)	87.81 (KDA)	84.94 (KDA)	87.72 (KDA)
	500	85.56 (KDA)	87.90 (MDA)	90.64 (KDA)	90.51 (KDA)	87.04 (KDA)	90.51 (KDA)

Note: The underlined letter indicates the highest mean percentage accuracy.

From Table 8, the highest mean percentage accuracies of the imputation–classification methods were statistically compared using the Friedman test. The results indicated a significant overall difference among the six imputation techniques (p-value < 0.05). Accordingly, the pairwise Wilcoxon signed-rank test revealed that the Mean and Bagged Trees (BT) imputations showed significant differences when compared with the other imputation methods, including Regression, KNN, RF, and MissRanger (p-value < 0.05). These results suggest that the Mean and BT imputations yielded relatively lower classification accuracies than the ensemble-based approaches.

The results indicate that advanced imputation techniques such as Regression, KNN, RF, and MissRanger substantially improve classification accuracy, particularly when used with flexible or regularized discriminant methods (e.g., RDA and KDA). Accuracy tends to increase with both correlation strength and sample size, confirming the stability and robustness of ensemble-based imputation in high-dimensional discriminant analysis.

Table 9. The highest mean percentage accuracies corresponding to each missing data imputation method for the ten predictors.

Correlation Levels (ρ)	Sample Sizes (n)	Methods of Missing Imputation for p = 10					
		Mean	**Reg**	**KNN**	**RF**	**BT**	**MissRanger**
0.3	100	74.78 (KDA)	87.84 (LDA)	76.26 (KDA)	76.45 (KDA)	74.97 (KDA)	90.40 (KDA)
	300	77.52 (SDA)	79.39 (FDA)	79.45 (SDA)	77.67 (SDA)	77.61 (SDA)	90.27 (KDA)
	500	78.69 (LDA)	86.57 (FDA)	81.02 (LDA)	80.85 (SDA)	78.78 (FDA)	90.34 (KDA)
0.7	100	84.47 (KDA)	87.61 (KDA)	85.50 (KDA)	85.15 (KDA)	84.75 (KDA)	90.33 (KDA)
	300	85.56 (KDA)	84.32 (SDA)	84.49 (KDA)	86.12 (KDA)	85.68 (KDA)	90.18 (KDA)
	500	85.83 (RDA)	86.57 (RDA)	87.46 (LDA)	87.06 (RDA)	86.26 (KDA)	90.26 (KDA)

Note: The underlined letter indicates the highest mean percentage accuracy.

At the low correlation level (ρ = 0.3), accuracy consistently improves with increasing sample size. The KDA classifier combined with MissRanger imputation achieves the highest performance across all sample sizes, reaching 90.40% accuracy when n = 100 and maintaining strong results (above 90%) for larger samples. Regression and KNN imputations also perform well, though slightly lower than MissRanger, while Mean and Bagged Tree (BT) imputations yield comparatively weaker accuracies.

At the moderate correlation level (ρ = 0.7), classification performance improves across all imputation methods, with the KDA classifiers achieving accuracies above 90% in all cases. The MissRanger method again provides the best overall results, peaking at 90.33% (n = 100) and 90.18% (n = 300), indicating both high predictive power and consistency. The results confirm that ensemble-based imputation methods such as MissRanger yield superior performance, particularly when combined with flexible or regularized classifiers like KDA.

From Table 9, the highest mean percentage accuracies of the imputation methods were statistically compared using the Friedman test, which revealed a significant overall difference among the six imputation techniques (p-value < 0.05). Subsequently, the pairwise Wilcoxon signed-rank test was conducted to identify specific differences among the imputation methods. Based on the p-values, several method pairs exhibited statistically significant differences (p-value < 0.05). In particular, the MissRanger imputation method showed significant differences compared with Mean, Regression, KNN, RF, and Bagged Trees (BT), indicating that ensemble-based imputations tended to achieve higher classification accuracies.

To evaluate the computational efficiency of each missing data imputation approach, both average runtime and peak memory usage were recorded during the simulation experiments. Runtime was measured in seconds as the average processing time required to complete one imputation cycle, while peak memory usage (in megabytes, MB) represents the maximum memory allocation during computation. These metrics provide insight into the trade-off between computational cost and imputation accuracy, highlighting the practicality of each method when applied to large or high-dimensional datasets. The results are summarized in Table 10.

Table 10. Average runtime and peak memory usage for each imputation method.

Imputation Method	Average Runtime (s)	Peak Memory (MB)
Mean	0.05	12
Regression	0.08	18
KNN	2.47	65
Random Forest	8.12	120
Bagged Trees	10.35	180
MissRanger	15.28	250

Note: The evaluation was performed on a dataset with n = 500, p = 10, ρ = 0.7, and 10% missing data.

From Table 10, while mean and regression imputations are nearly instantaneous, ensemble-based methods such as Bagged Trees and MissRanger are computationally more intensive due to iterative model fitting and aggregation. However, the substantial gains in accuracy and robustness justify the additional computational cost for practical applications involving complex or high-dimensional data.

4. Results of Actual Data

This study utilizes real-world clinical data concerning liver disease progression, specifically cirrhosis caused by chronic liver conditions such as hepatitis and prolonged alcohol abuse. The dataset analyzed in this research is publicly available from the title Cirrhosis

Prediction Dataset and can be accessed at https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset (accessed on 9 August 2025). The data originally stem from a clinical trial conducted by the Mayo Clinic on primary biliary cirrhosis between 1974 and 1984.

The Cirrhosis Prediction Dataset consists of 418 patient records, of which 312 complete cases were used for analysis after excluding observations with missing outcome labels. The outcome variable, Stage, represents the histologic stage of liver disease with four classes: Stage 1 (n = 69), Stage 2 (n = 121), Stage 3 (n = 97), and Stage 4 (n = 25). The dataset includes nine continuous biochemical predictor variables, Age (in years), Bilirubin (serum bilirubin in mg/dL), Cholesterol (serum cholesterol in mg/dL), Albumin (serum albumin in g/dL), Copper (urine copper in µg/day), Alkaline Phosphatase (Alk_Phos) in U/L, SGOT (serum glutamic-oxaloacetic transaminase in U/mL), Triglycerides (in mg/dL), Platelets (platelet count per 1000 cells/mL), and Pro-thrombin Time (in seconds). The missing rates vary across variables: Cholesterol (8.97%), Triglycerides (9.61%), Copper (0.64%), and Platelets (1.28%). In contrast, other variables are fully observed. Table 11 presents the descriptive statistics of the key continuous variables, including minimum, quartiles, mean, maximum, and the number of missing observations per variable. These statistics offer an overview of the data distribution and help identify variables requiring imputation before classification modeling.

Table 11. The descriptive statistics of the data.

Data	Min	Q1	Median	Mean	SD	Q3	Max	Missing
Age [years]	26.30	42.27	49.83	50.05	10.53	56.75	78.49	-
Bilirubin [mg/dL]	0.30	0.80	1.35	3.256	4.60	3.425	12.00	-
Cholesterol [mg/dL]	120	249.5	309.5	369.5	234.78	400	1775.0	28 (8.97%)
Albumin [g/dL]	1.96	3.31	3.55	3.52	0.40	3.80	4.64	-
Copper [µg/day]	4.00	41.25	73.00	97.65	88.26	123.00	588.00	2 (0.61%)
Alk_Phos [U/liter]	289.0	871.5	1259.0	1982.7	2115.47	1980.0	13862.4	-
SGOT [mg/dL]	26.35	80.60	114.70	122.56	56.71	151.90	457.25	-
Triglycerides [mg/dL]	33.00	84.25	108.00	124.70	65.28	151.00	598.00	30 (9.61%)
Platelets [1000 cells/mL]	62.0	199.8	257.0	261.9	93.12	322.5	563.0	4 (1.28%)
Prothrombin [second]	9.00	10.0	10.60	10.73	1.00	11.10	17.10	-

Table 11 presents the reported statistics, including the minimum, quartiles, median, mean, standard deviation, maximum, and the number of missing values. While most variables have complete data, some, such as cholesterol, triglycerides, copper, and platelets, contain missing entries, with triglycerides having the highest number of missing values (30). Certain variables, such as alkaline phosphatase and cholesterol, exhibit strong right-skewed distributions with extremely high maximum values, indicating the presence of potential outliers. In contrast, others, such as age and albumin, appear more symmetrically distributed. These summary statistics provide an overview of the data structure and emphasize the need for appropriate imputation techniques before classification modeling.

To evaluate the missingness mechanism, Little's MCAR test was performed. It yielded a non-significant result (p-value > 0.05), suggesting that the missing data can be reasonably considered Missing Completely at Random (MCAR). Therefore, the application of the same imputation framework as in the simulation study is justified. This ensures methodological consistency and avoids assumption violations among the imputation techniques employed.

The classification results obtained using various discriminant methods and imputation strategies, including mean, regression, K-Nearest Neighbors (KNN), random forest (RF), Bagged Trees (BT), and MissRanger, are presented in Table 12. This table reports the percentage accuracy for each combination of methods.

Table 12. The percentage accuracy, 95% confidence interval, and Cohen's kappa for multi-classification of the data.

Methods	LDA	RDA	FDA	MDA	KDA	SDA
Mean	51.08 40.44–61.65 0.2516	51.08 40.44–61.65 0.2516	51.08 40.44–61.65 0.2516	53.26 42.56–63.74 0.2955	51.08 40.44–61.65 0.2434	44.56 34.19–55.29 0.169
Reg	48.91 38.34–59.55 0.2029	48.91 38.34–59.55 0.2029	48.91 38.34–59.55 0.2029	43.47 33.16–54.21 0.1623	52.17 41.50–62.70 0.2695	46.73 36.25–57.43 0.1920
KNN	48.91 38.34–59.55 0.1760	40.21 38.34–59.55 0.1760	47.82 37.29–57.49 0.1615	51.08 40.44–61.65 0.2157	47.82 37.29–58.49 0.1697	45.56 35.22–56.36 0.4565
RF	57.60 46.86–67.85 0.5760	<u>57.60</u> 46.86–67.85 0.3517	57.60 46.86–67.85 0.3517	54.34 43.63–64.77 0.2987	53.26 42.56–63.74 0.2762	46.73 36.25–57.43 0.1972
BT	52.17 41.50–62.70 0.2737	54.34 41.50–62.70 0.2737	52.17 41.50–62.70 0.2737	52.17 41.50–62.70 0.2877	59.78 49.04–69.87 0.3781	48.91 38.34–59.55 0.2400
MissRanger	<u>58.69</u> 47.94–68.86 0.3388	52.17 47.94–68.86 0.3388	<u>58.69</u> 47.94–68.86 0.3393	<u>60.86</u> 50.13–70.88 0.4019	<u>63.04</u> 52.34–72.87 0.4190	<u>54.34</u> 43.63–64.77 0.5434

Note: The underlined letter indicates the highest mean percentage accuracy.

Table 12 presents the percentage accuracy, 95% proportion confidence intervals, and Cohen's kappa coefficients for six discriminant analysis (DA) methods applied to real clinical data using different missing data imputation techniques.

Among the imputation methods, MissRanger achieves the highest classification accuracies across most DA models, particularly for KDA (63.04%), MDA (60.86%), and LDA (58.69%), indicating that ensemble-based imputation substantially improves predictive performance. The corresponding Cohen's kappa values (0.3388–0.5434) suggest moderate agreement between predicted and valid class memberships. Bagged Tree (BT) and KNN imputations perform competitively, though their accuracies are slightly lower (approximately 52–54%). In contrast, simpler methods such as Mean and Regression imputations yield the weakest results, with accuracies below 52%.

These results confirm that ensemble-based imputation approaches, particularly MissRanger, enhance model stability and classification reliability across all discriminant analysis frameworks. The improvement in both accuracy and kappa indicates that MissRanger preserves the multivariate data structure more effectively than traditional imputation techniques, making it well-suited for real-world datasets with incomplete information.

To further evaluate the classification performance of the discriminant analysis models on actual clinical data, confusion matrices were constructed for the three best-performing methods (LDA, MDA, and KDA) using the MissRanger imputation technique. Each confusion matrix presents the number of correctly and incorrectly classified observations across the four disease stages (Class 1–Class 4), allowing for a detailed assessment of model accuracy and misclassification patterns. The results are summarized in Table 13.

Table 13 displays the confusion matrices and percentage accuracies for the LDA, MDA, and KDA models applied to the Cirrhosis dataset after MissRanger imputation. The KDA model achieves the highest classification accuracy (63.04%), correctly identifying most observations in Classes 3 and 4, which represent more advanced disease stages. The MDA model follows with an accuracy of 60.86%, showing balanced but slightly less precise predictions across classes. The LDA model attains a lower accuracy of 58.69%, with greater misclassification among Classes 2 and 3.

Table 13. Confusion matrices and percentage classification accuracies for LDA, MDA, and KDA methods.

DA Methods	Predicted Data ($\hat{y}$)	Actual Data (y)				Percentage Accuracy
		Class 1	Class 2	Class 3	Class 4	
LDA	Class 1	0	0	0	0	58.69
	Class 2	1	1	0	0	
	Class 3	4	16	33	12	
	Class 4	1	1	3	20	
MDA	Class 1	1	0	0	0	60.86
	Class 2	3	6	3	2	
	Class 3	1	11	29	10	
	Class 4	1	1	4	20	
KDA	Class 1	0	0	0	0	63.04
	Class 2	1	5	0	0	
	Class 3	5	13	29	8	
	Class 4	0	0	7	24	

5. Discussion

This study investigated the classification performance of five discriminant analysis techniques, LDA, RDA, FDA, MDA, KDA, and SDA, under various missing data strategies, including mean, regression, KNN, RF, BT, and MissRanger, using both real clinical data and simulated datasets. The findings, summarized in Tables 1–9, demonstrate the significant impact that handling missing data has on classification accuracy, particularly when dealing with complex, high-dimensional correlation and partially observed data.

The diagnostic analysis of multivariate normality and covariance homogeneity indicated that while the simulated datasets adhered closely to theoretical assumptions, the real imputed dataset exhibited mild violations due to skewness and unequal variances among groups. Nevertheless, the robust nature of flexible classifiers, particularly RDA, FDA, KDA, and SDA, allowed them to maintain stable performance under such conditions. This confirms that ensemble-based imputations, such as MissRanger, effectively preserve multivariate dependencies, thereby mitigating the impact of assumption violations on classification accuracy.

LDA tends to yield low-variance boundaries under its parametric assumptions but may be biased under covariance heterogeneity. QDA relaxes these assumptions, lowering bias at the expense of higher variance. RDA and SDA reduce variance through covariance shrinkage, which is particularly beneficial in high-dimensional or small-sample settings. FDA and MDA primarily reduce bias by increasing model flexibility (optimal scoring with flexible regression; class-mixture modeling), though this can increase variance unless adequately regularized. KDA reduces bias by allowing nonlinear boundaries; kernel regularization is essential for controlling variance. Our results align with this view: shrinkage-based methods (RDA and SDA) are comparatively more stable (smaller SD, narrower CI), whereas more flexible methods (FDA, MDA, and KDA) often achieve higher central performance in misspecified settings but exhibit larger variability unless tuning is sufficiently regularized.

The findings from this study align closely with existing literature on the impact of missing data handling in classification problems, reaffirming that the choice of imputation method can substantially influence model performance [17,18]. Both the simulation experiments and the analysis of the Cirrhosis Prediction Dataset revealed that ensemble-based imputations, particularly MissRanger, consistently outperformed simpler methods such as mean and regression imputation across a range of discriminant analysis (DA) tech-

niques. This advantage is consistent with the results of Stekhoven and Bühlmann [32], who demonstrated that random forest-based imputations can effectively capture nonlinear dependencies and maintain the multivariate structure of mixed-type data.

In simulated settings, regression, MissRanger and KNN repeatedly achieved the highest classification accuracies under varying sample sizes, correlation levels, and predictor dimensionalities. For example, under high correlated and high-dimensional conditions, MissRanger combined with KDA achieved accuracies approaching 90%, underscoring the synergy between advanced imputation and flexible classifiers, as also emphasized by Schäfer and Strimmer [31] in their work on preserving covariance structure in high-dimensional analysis. Regression imputation, while comparatively less effective under low correlation, improved markedly when correlation increased and was paired with regularized methods such as RDA, consistent with the theoretical framework of Ledoit and Wolf [37] on shrinkage-based covariance estimation for stabilizing classification in ill-conditioned settings.

From a theoretical perspective, the simulation findings are consistent with the asymptotic properties of discriminant estimators under missing-data mechanisms. Under standard regularity conditions, estimators in LDA and QDA are asymptotically unbiased and consistent as $n \to \infty$ provided that covariance estimates are unbiased. When missing data are imputed, an additional variance component is introduced.

Imputation methods such as regression and ensemble-based approaches (MissRanger, Bagged Trees) mitigate this issue by stabilizing covariance estimation and maintaining asymptotic efficiency, whereas simpler methods like mean imputation may introduce deterministic bias. These theoretical insights explain why the proposed framework yielded stable and accurate classification across correlation structures and missingness levels, confirming its asymptotic robustness under MCAR conditions.

The real-data analysis confirmed the simulation results: MissRanger achieved the highest average accuracies for LDA, FDA, MDA, KDA, and SDA, while Random Forrest imputation performed best for RDA. These findings are in agreement with Hong et al. [23], who demonstrated that machine learning-based imputations improve classification in medical data, and Bai et al. [24], who reported robust performance of autoencoder-based imputation in high-missingness clinical datasets. Although computationally inexpensive, simple imputation methods were unable to model complex variable interactions, leading to reduced classification performance, a limitation also noted by Little and Rubin [39] and van Buuren and Groothuis-Oudshoorn [40].

While this study employed a single training/testing split to ensure consistency with the simulation design, implementing cross-validation would provide a more comprehensive assessment of model stability and predictive generalization across multiple data partitions. This approach could also help confirm whether the observed ranking of imputation–classifier combinations remain consistent under repeated resampling, thereby enhancing the reliability and external validity of the findings.

Both simulations and real data show that larger sample sizes consistently enhance classification performance for all imputation–classifier combinations, reflecting improvements in parameter stability and imputation accuracy. This observation supports the conclusions of Palanivinayagam and Damaševičius [19], who noted that sample size plays a critical role in the success of imputation-driven classification frameworks.

The superior performance of MissRanger and Bagged Trees can be attributed to their ability to preserve multivariate covariance structures and capture nonlinear dependencies among variables. Unlike single-value imputations that treat each variable independently, ensemble-based approaches jointly model relationships across predictors, allowing the imputed data to better reflect the original data geometry. In particular, MissRanger com-

bines Random Forest prediction with predictive mean matching, maintaining the empirical distribution of continuous variables and reducing bias. Bagged Trees, through bootstrap aggregation, stabilize the imputation process and mitigate random variation in estimated values, which in turn enhances covariance estimation and classification boundary stability. These mechanisms collectively explain why ensemble-based imputations yield higher discriminant accuracy, particularly in high-correlation and high-dimensional settings, as also observed by Stekhoven and Bühlmann [42] and Schäfer and Strimmer [44].

Overall, the results underscore that optimal classification performance in incomplete-data scenarios depends on aligning the imputation strategy with the assumptions and flexibility of the classifier. Ensemble-based approaches, such as MissRanger, when integrated with flexible or regularized discriminant classifiers, offer a robust and scalable framework for multi-class classification in both theoretical and applied contexts. This is consistent with prior recommendations by Khashei et al. [20] and Sharmila et al. [21] for adopting advanced and context-appropriate imputation methods.

6. Conclusions

This study evaluated the impact of various missing data imputation techniques on the classification performance of five discriminant analysis methods using both real-world clinical data and controlled simulation datasets. The analysis of the Cirrhosis Prediction Dataset revealed that ensemble-based imputation methods, particularly MissRanger and Random Forrest, significantly improved classification accuracy compared to simpler approaches such as mean and regression imputation. Flexible and regularized classifiers such as RDA, FDA, and MDA were more responsive to advanced imputation methods, while classical LDA showed only marginal improvement. Simulation results further reinforced these findings under varying sample sizes, correlation levels, and predictor dimensions. MissRanger consistently yielded high classification accuracies, especially in high-dimensional settings and under strong correlation. Moreover, regression-based imputation performed better under low correlation, particularly when used with regularized models like KDA and MDA. The effectiveness of each imputation method was also found to depend on its compatibility with the underlying classifier assumptions. Notably, larger sample sizes consistently enhanced performance across all settings, reinforcing the importance of sufficient data for both imputation accuracy and model stability.

The findings emphasize that selecting an appropriate imputation strategy is critical for maximizing classification performance in the presence of missing data. Ensemble-based methods such as MissRanger and regression, when paired with flexible discriminant classifiers, provide robust and scalable solutions for both clinical and simulated data scenarios. Future research could extend this framework by exploring deep learning-based imputers, evaluating model interpretability, or benchmarking under different missing data mechanisms to further improve generalizability and real-world applicability.

Author Contributions: Conceptualization, A.A. and A.K.; methodology, A.A.; software, A.K.; validation, A.A. and A.K.; formal analysis, A.A.; investigation, A.K.; resources, A.A.; writing—original draft, A.A. and A.K.; writing—review and editing, A.A. and A.K.; visualization, A.A.; supervision, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: The research titled "An Enhanced Discriminant Analysis Approach for Multi-Classification with Integrated Machine Learning-Based Missing Data Imputation" (grant number RE-KRIS/FF68/51) by King Mongkut's Institute of Technology, Ladkrabang, School of Science, Department of Statistics, has received funding support from the NSRF.

Data Availability Statement: Data are available at https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset (accessed on 9 August 2025).

Acknowledgments: This research was supported by King Mongkut's Institute of Technology Ladkrabang and NSRF.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DA	Discriminant Analysis
LDA	Linear Discriminant Analysis
RDA	Regularized Discriminant Analysis
FDA	Flexible Discriminant Analysis
MDA	Mixture Discriminant Analysis
SDA	Shrinkage Discriminant Analysis
KNN	k-Nearest Neighbors
RF	Random Forest
BT	Bagged Trees

Appendix A

Appendix A.1

To ensure reproducibility and transparency, the simulation setup was described in detail. The data generation process, parameter settings, and randomization control are summarized below. These additions clarify how predictors, covariance structures, missingness, and classification processes were generated and replicated across all simulation experiments.

Table A1. Summary of key simulation parameters.

Parameter	Symbol	Values/Description
Number of predictor variables	p	5, 10
Sample sizes	n	100, 300, 500
Correlation levels	ρ	0.3, 0.7
Covariance structure	Σ	Toeplitz: $\Sigma_{jk} = \rho^{\lvert j-k \rvert}$
Missingness mechanism	–	10% missing
Number of replications	-	1000
Class distribution	–	Balanced (4 classes)
Random seed setup	99	Fixed master seed

References

1. Jain, S.; Kuriakose, M. Discriminant analysis—Simplified. *Int. J. Contemp. Dent. Med. Rev.* **2020**, *2019*, 031219.
2. Ramayah, T.; Ahmad, N.H.; Halim, H.A.; Zainal, S.R.M.; Lo, M.C. Discriminant analysis: An illustrated example. *Afr. J. Bus. Manag.* **2010**, *4*, 1654–1667.
3. Singh, A.; Gupta, S. Implementation of linear and quadratic discriminant analysis incorporating costs of misclassification. *Processes* **2021**, *9*, 1382.
4. Chatterjee, A.; Das, D. Comparative Study of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) in Dataset. *Int. J. Comput. Appl.* **2020**, *975*, 8887.
5. Berrar, D. Linear vs. quadratic discriminant analysis classifier: A tutorial. *Mach. Learn. Bioinform.* **2019**, *106*, 1–13.
6. Friedman, J.H. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **1989**, *84*, 165–175. [CrossRef]
7. Di Franco, C.; Palumbo, F. A RDA-based clustering approach for structural data. *Stat. Appl.* **2022**, *34*, 249–272.
8. Hastie, T.; Tibshirani, R.; Buja, A. Flexible discriminant analysis. *J. Am. Stat. Assoc.* **1994**, *89*, 1255–1270. [CrossRef]
9. Hastie, T.; Tibshirani, R. Discriminant analysis by Gaussian mixtures. *J. R. Stat. Soc. B* **1996**, *58*, 155–176. [CrossRef]
10. Notice, D.; Soleimani, H.; Pavlidis, N.G.; Kheiri, A.; Muñoz, M.A. Instance Space Analysis of the Capacitated Vehicle Routing Problem with Mixture Discriminant Analysis. In Proceedings of the GECCO '25: Proceedings of the Genetic and Evolutionary Computation Conference, Málaga, Spain, 14–18 July 2025; ACM: New York, NY, USA, 2025; pp. 1–9.

11. Bai, X.; Zhang, M.; Jin, Z.; You, Y.; Liang, C. Fault Detection and Diagnosis for Chiller Based on Feature-Recognition Model and Kernel Discriminant Analysis. *Sustain. Cities Soc.* **2022**, *79*, 103708. [CrossRef]
12. Bickel, P.J.; Levina, E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **2004**, *10*, 989–1010. [CrossRef]
13. Vo, T.H.; Nguyen, L.T.; Vo, B.N.; Vo, A.H. Weighted missing linear discriminant analysis. *arXiv* **2024**, arXiv:2407.00710. [CrossRef]
14. Nguyen, D.; Yan, J.; De, S.; Liu, Y. Efficient parameter estimation for multivariate monotone missing data. *arXiv* **2020**, arXiv:2009.11360.
15. Pepinsky, T.B. A Note on Listwise Deletion versus Multiple Imputation. *Polit. Anal.* **2018**, *26*, 480–488. [CrossRef]
16. Ibrahim, J.G.; Molenberghs, G. Missing Data Methods and Applications. In *The Oxford Handbook of Applied Bayesian Analysis*; O'Hagan, A., West, M., Eds.; Oxford University Press: Oxford, UK, 2010.
17. Kang, H. The prevention and handling of the missing data. *Korean J. Anesthesiol.* **2013**, *64*, 402–406. [CrossRef]
18. Agiwal, V.; Chaudhuri, S. Methods and Implications of Addressing Missing Data in Health-Care Research. *Curr. Med.* **2024**, *22*, 60–62. [CrossRef]
19. Palanivinayagam, A.; Damaševičius, R. Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. *Information* **2023**, *14*, 92. [CrossRef]
20. Khashei, M.; Najafi, F.; Bijari, M. Pattern classification with missing data: A review and future research directions. *Appl. Soft Comput.* **2023**, *136*, 110141.
21. Sharmila, R.; Sundararajan, V.; Krishnamoorthy, S. Classification Techniques for Datasets with Missing Data: A Comprehensive Review. In Proceedings of the 2022 International Conference on Computing, Communication and Green Engineering (CCGE), Coimbatore, India, 2–4 December 2022; pp. 252–256.
22. Rácz, A.; Gere, A. Comparison of missing value imputation tools for machine learning models based on product development case studies. *LWT-Food Sci. Technol.* **2025**, *221*, 117585. [CrossRef]
23. Hong, J.; Lee, H.; Kim, D. Enhancing missing data imputation using machine learning techniques for diabetes classification. *Health Inform. J.* **2020**, *26*, 2671–2685.
24. Bai, T.; Liang, X.; He, L.; Zhang, H. Deep learning-based imputation with autoencoder for medical data with missing values. *IEEE Access* **2022**, *10*, 59301–59313.
25. van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; Chapman and Hall/CRC: New York, NY, USA, 2018.
26. Audigier, V.; White, I.R.; Jolani, S.; Debray, T.P.A.; Quartagno, M.; Carpenter, J.; van Buuren, S.; Resche-Rigon, M. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Stat. Sci.* **2018**, *33*, 160–183. [CrossRef]
27. Resche-Rigon, M.; White, I.R.; Bartlett, J.W.; Carpenter, J.R.; van Buuren, S. Multiple Imputation for Missing Data in Multilevel Models: A Practical Guide. *Stat. Methods Med. Res.* **2020**, *29*, 1348–1364.
28. Zhang, F.; Liu, S.; Li, J. A Machine Learning-Based Multiple Imputation Method for Incomplete Medical Data. *Information* **2023**, *10*, 77. [CrossRef]
29. Zhang, Y.; Li, H. GAN-Based Imputation Framework for Multivariate Time-Series Data. *Pattern Recognit. Lett.* **2024**, *184*, 56–64.
30. Park, J.; Kim, S.; Lee, D. A Hybrid Missing Data Imputation Model Combining MICE and Variational Autoencoders. *Knowl.-Based Syst.* **2025**, *298*, 112056.
31. Schäfer, J.; Strimmer, K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4*, 32. [CrossRef]
32. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
33. Zhao, S.; Zhang, B.; Yang, J.; Zhou, J.; Xu, Y. Linear Discriminant Analysis. *Nat. Rev. Methods Primers* **2024**, *4*, 70. [CrossRef]
34. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley: Hoboken, NJ, USA, 2004.
35. Baudat, G.; Anouar, F. Generalized Discriminant Analysis Using a Kernel Approach. *Neural Comput.* **2000**, *12*, 2385–2404. [CrossRef]
36. Cai, D.; He, X.; Han, J. Speed Up Kernel Discriminant Analysis. *VLDB J.* **2011**, *20*, 21–33. [CrossRef]
37. Ledoit, O.; Wolf, M. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [CrossRef]
38. Ahdesmäki, M.; Strimmer, K. Feature Selection in Omics Prediction Problems Using CAT Scores and False Non-Discovery Rate Control. *Ann. Appl. Stat.* **2010**, *4*, 503–519. [CrossRef]
39. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2002.
40. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
41. Zhang, S. Nearest Neighbor Selection for Iteratively kNN Imputation. *J. Syst. Softw.* **2012**, *85*, 2541–2552. [CrossRef]
42. Golino, H.F.; Gomes, C.M.A. Random Forest as an Imputation Method for Education and Psychology Research: Its Impact on Item Fit and Difficulty of the Rasch Model. *J. Appl. Stat.* **2016**, *43*, 401–421. [CrossRef]

43. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
44. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
45. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]