

Mask-Aware MLP and Entropy-Weighted MCDM for Robust Multiclass Paper Ranking

Masoud Kargar

kargar@iau.ac.ir

Ta.C, Islamic Azad University

Shahin Sharbaf Movassaghpour

Ta.C, Islamic Azad University

Habib Izadkhah

University of Tabriz

Research Article

Keywords: Classification, Evaluation Metrics, Mask-aware Multilayer Perceptron, Multi-criteria Decision Making

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10829779/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Mask-Aware MLP and Entropy-Weighted MCDM for Robust Multiclass Paper Ranking

Masoud Kargar¹, Shahin Sharbaf Movassaghpour¹, Habib Izadkhah²

¹Department of Computer Engineering, Ta.C., Islamic Azad University, Tabriz, Iran

²Department of Computer Science, University of Tabriz, Tabriz, Iran

Email: kargar@iaui.ac.ir, sh.movassaghpour@iaut.ac.ir, izadkhah@tabrizu.ac.ir

ABSTRACT

Comprehensive evaluation and ranking of scientific articles are essential for assessing the validity of reported findings and enabling meaningful comparisons across studies. A major challenge is missing or erroneous values in evaluation metrics, which can undermine the reliability of validation and ranking procedures. Researchers have proposed various approaches based on advanced deep learning techniques, statistical methods, and multi-criteria decision-making frameworks to address these limitations. In this study, we generated simulated datasets using statistical distributions and confusion matrices. The datasets included complete, noisy, and incomplete observations, emulating the outputs of classification models applied to problems with 2–10 classes. We provided the resulting data, along with corresponding binary masks indicating missing-value locations, as inputs to a mask-aware multilayer perceptron. We developed separate models for each class configuration and considered both macro and weighted-average metrics, resulting in 17 distinct models. The proposed mask-aware framework imputes missing values and provides confidence estimates for screening potentially inconsistent evaluation metrics. We then determine final article rankings using an MCDM method with entropy-based criterion weights. Experimental results show that the validity-classification models achieve accuracies ranging from 0.8961 to 0.9274 and F1 scores ranging from 0.90 to 0.93 across classification settings involving 2–10 classes. In missing-value imputation experiments, we compared the proposed MaskAwareMLP with mean imputation, KNN, MICE, and MissForest. The proposed method performed competitively at lower missingness levels and achieved the lowest MAE and RMSE and the highest R^2 among the evaluated methods at 30% and 40% missingness. In ranking-preservation experiments, MaskAwareMLP achieved the highest Kendall's τ -b at 30% and 40% missingness, with values of 0.7879 and 0.7165, respectively, while maintaining competitive ranking agreement at lower missingness levels. These results indicate that explicitly incorporating missingness patterns can improve neural imputation and downstream MCDM-based ranking robustness, particularly under moderate-to-high missingness. Finally, we implemented the proposed framework as a user-friendly application that supports automated and visual analysis of scientific articles.

Keywords: Classification, Evaluation Metrics, Mask-aware Multilayer Perceptron, Multi-criteria Decision Making

1. INTRODUCTION

One key machine learning task is classification, which spans diverse domains to address complex problems. In medicine, this technique helps detect diseases such as cancer and cardiac disorders early. In image processing, it enables object recognition without human intervention. Furthermore, in large-scale dataset analysis and recommender systems, it is used to suggest items on commercial platforms [1,2]. Researchers evaluate classification models using metrics such as accuracy, precision, recall, specificity, and F1-score, which enable rigorous, precise scientific comparisons among methods and models [3]. These metrics are derived from the confusion matrix and help researchers meticulously examine the strengths and weaknesses of different approaches [4]. Reliable reporting of such performance measures is therefore essential for reproducible evaluation, fair comparison, and informed selection of classification models.

Despite substantial advances in classification algorithms, the incomplete or inaccurate reporting of performance metrics remains a persistent concern in the scientific literature [5]. Many studies report only a subset of essential metrics or omit fundamental information, such as the confusion matrix, thereby hindering the independent verification of performance calculations and comprehensive comparisons among competing methods [3]. In medical diagnostics, for example, failing to report either the confusion matrix or the values needed to reconstruct it may prevent independent calculation of recall, despite the metric's critical importance for quantifying a model's ability to identify patients with a disease [6]. Moreover, some reported values may be unreliable because of calculation errors, implementation flaws, or, in rare cases, deliberate data manipulation [4]. These deficiencies impede assessment of reported results and reliable ranking of classification methods, particularly when performance metrics differ marginally across studies [7]. Consequently, they may compromise the fairness of cross-study comparisons and diminish confidence in scientific findings [8].

However, the lack of automated, integrated approaches for handling incomplete or potentially erroneous performance metrics hinders efficient identification of suitable models and may impede the development of improved methodological solutions [9]. Missing-value imputation techniques include conventional approaches such as mean imputation [50], k-nearest-neighbor (KNN) imputation [51], multiple imputation by chained equations (MICE) [52], and Random-Forest-based iterative approaches such as the MissForest methodology [10]. Researchers have also investigated neural-network-based approaches, including conventional multilayer perceptrons (MLPs), for missing-data reconstruction. In parallel, multi-criteria decision-making methods, including entropy-weighted approaches, are commonly used to rank and select candidate models based on multiple performance criteria [53,54]. Recent studies have also investigated missing and incomplete information in multi-criteria decision-making [55–57], while the consistency of reported machine-learning performance measures has also received attention [58]. These studies demonstrate that missing-data handling, metric consistency assessment, and multi-criteria ranking are established research areas. Nevertheless, these lines of research have largely evolved independently, and the specific problem of reconstructing incomplete classification-performance profiles while explicitly accounting for the pattern of missingness and its effect on downstream ranking remains insufficiently addressed.

A particularly important limitation of conventional imputation approaches is that they generally treat the available numerical values as the primary source of information and do not explicitly represent which values are observed and which are missing. Although neural networks have been applied to missing-data reconstruction, using a multilayer perceptron explicitly augmented with a missingness mask to reconstruct classification-performance metrics has received limited attention. To the best of our knowledge, the use of a mask-aware MLP specifically designed to exploit both observed performance values and their explicit missingness pattern for classification-performance reconstruction and subsequent model ranking has not been systematically examined. This motivates the development of a MaskAwareMLP, which explicitly incorporates a binary missingness mask into the learning process, enabling the model to distinguish observed values from missing or initially imputed values.

The need for such a missingness-aware reconstruction approach becomes particularly important when imputed metrics are subsequently used for model selection or ranking. Accurate reconstruction of individual missing cells does not necessarily preserve the relative ordering of competing classification models. Thus, evaluating imputation methods solely through cell-level measures such as MAE, RMSE, or R^2 may not fully capture their usefulness in scientific model comparison. Accordingly, we integrate the proposed MaskAwareMLP into an entropy-based multi-criteria decision-making framework and evaluate the resulting rankings against rankings obtained from complete reference information. To enable the development and controlled evaluation of the proposed approach, this study also creates a dedicated simulated dataset containing internally consistent classification-performance profiles with known ground-truth values. We generate the dataset from confusion matrices for binary and multiclass classification scenarios, including different numbers of classes and commonly reported performance metrics. We introduce controlled noise to simulate potentially erroneous or inconsistent reported values, and we use random feature masking to generate incomplete profiles at different missingness levels. This dataset provides a controlled environment for training and evaluating the proposed MaskAwareMLP and for objectively measuring reconstruction and ranking performance under known conditions.

The principal methodological novelty of this study is therefore the MaskAwareMLP architecture and its explicit use of the missingness pattern as an additional source of information for performance-metric reconstruction. We conduct an ablation study to isolate the contribution of the proposed mask-aware mechanism. The study further examines the behavior of the proposed and competing imputation methods across different missingness levels, recognizing that no single method is necessarily optimal under all conditions. The results indicate that the advantage of the MaskAwareMLP becomes more pronounced under moderate-to-high missingness, particularly at 30% and 40% missingness.

The study also contributes by jointly evaluating cell-level reconstruction and downstream ranking preservation. We compare established imputation methods, including mean imputation, KNN, MICE, and an iterative Random-Forest-based approach, with the proposed MaskAwareMLP. Reconstruction quality is assessed using MAE, RMSE, and R^2 , while ranking preservation is evaluated using Kendall's τ -b and Spearman's ρ . This dual evaluation determines whether improvements in metric reconstruction translate into more reliable classification-model rankings.

Finally, the proposed framework integrates missing-value estimation, identification of potentially erroneous performance profiles, reliability assessment, and entropy-based MCDM ranking within a single automated workflow. A multi-stage search procedure refines missing metric estimates, and the complete methodology is implemented through a graphical user interface that enables researchers to enter or upload performance data, estimate missing values, assess reliability, and rank scientific articles.

Thus, this study extends beyond applying existing imputation and MCDM techniques by providing a missingness-aware reconstruction architecture, a dedicated controlled dataset, an ablation-based evaluation, and a decision-oriented framework for comparing incomplete classification-performance information.

The remainder of this article is organized as follows. Section 2 reviews the relevant literature; Section 3 presents the proposed methodology; Section 4 reports and discusses the experimental results; and Section 5 summarizes the principal contributions, outlines the study limitations, and identifies directions for future research.

2. PREVIOUS WORKS

Several key challenges arise in evaluating and ranking machine learning models for classification tasks. These include selecting and reporting appropriate performance metrics [3] and effectively handling missing values, which can compromise the accuracy and reliability of model evaluations [2]. Multi-criteria decision-making (MCDM) methods can further support systematic comparisons and model ranking by integrating multiple, potentially conflicting evaluation criteria [12]. Existing studies relevant to this problem can be broadly organized into five research directions: (1) machine-learning-based and statistical methods for imputing missing values in datasets and performance metrics; (2) MCDM methods for ranking alternative models; (3) decision-making under incomplete information; (4) consistency assessment of reported machine-learning performance metrics; and (5) hybrid approaches that integrate machine learning and MCDM to support ranking and decision-making in complex settings characterized by incomplete data [11]. Table 1 summarizes representative methods within these research directions. The following subsections critically review these research areas and identify the remaining methodological gaps.

Missing values in classification performance metrics may introduce uncertainty into the evaluation process and reduce the reliability of model comparisons [3]. Conventional approaches, such as mean imputation and constant-value substitution, are widely used because of their simplicity [50]. However, such approaches may fail to preserve relationships among variables and may therefore provide less accurate estimates when the underlying data structure is heterogeneous or nonlinear [4]. K-nearest-neighbor (KNN) imputation provides an alternative by exploiting similarities between observations and has been widely used for missing-value estimation [51]. Multiple imputation by chained equations (MICE) provides a more flexible iterative approach for multivariate missing-data problems by modeling incomplete variables conditional on other observed variables [52]. Similarly, Stekhoven and Bühlmann [10] proposed MissForest, a Random-Forest-based nonparametric imputation method that models nonlinear relationships and interactions in mixed-type data. To address the limitations of conventional imputation approaches, several studies have investigated neural-network-based methods for reconstructing incomplete observations. Silva-Ramírez et al. [13] used multilayer perceptron networks to impute data missing completely at random (MCAR), reporting improved imputation performance for highly noisy datasets. Pan et al. [14] subsequently proposed an improved MLP-based method (IMLP) that incorporates a momentum-based gradient optimization mechanism and achieves favorable performance across multiple missingness patterns. Choudhury et al. [15] introduced an augmented MLP designed to process incomplete datasets and showed it can maintain classification performance under challenging conditions. These studies demonstrate the potential of neural networks for reconstructing incomplete data; however, they primarily address general data imputation or classification and do not explicitly use a binary mask to distinguish observed and missing data in classification-performance metrics.

Researchers have also developed alternative imputation strategies using ensemble learning and generative modeling. Yoon et al. [16] introduced generative adversarial imputation networks (GAIN), which exploit adversarial learning to reconstruct missing observations. Wan et al. [17] proposed a score-based diffusion approach for tabular data imputation that incorporates local and global information. These approaches offer flexible alternatives for missing-data reconstruction; however, their primary objective is recovering incomplete observations rather than preserving rankings across multiple evaluation criteria. MCDM methods, including the Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS), Elimination and Choice Translating Reality (ELECTRE), and Vlsekriterijumska Optimizacija I Kompromisno Resenje (VIKOR), provide systematic frameworks for ranking classification models against multiple evaluation criteria [12]. By explicitly defining criteria, weights, and preference structures, these methods can support transparent comparisons when several performance indicators must be considered simultaneously. Depending on the method employed, alternatives may be evaluated based on their relative proximity to an ideal solution, outranking relationships, or a compromise between group utility and individual regret [22]. When criterion weights are derived using data-driven procedures, such as entropy weighting, subjective judgments may also be reduced.

Peng et al. [18] applied an entropy-weighted TOPSIS framework to rank classifiers for software defect prediction and examined the robustness of the resulting rankings through sensitivity analysis. Tunç and Solmaz [19] integrated entropy weighting with TOPSIS and applied sensitivity analysis to evaluate the robustness of the resulting rankings under different weighting scenarios. Sharma et al. [20] compared a hybrid fuzzy analytic hierarchy process–entropy–TOPSIS (FAHP–Entropy–TOPSIS) framework with the conventional analytic hierarchy process (AHP) for industrial ranking. Mahmoudi et al. [21] proposed a hybrid TOPSIS–ordinal priority approach (OPA) for project selection under incomplete information, incorporating an optimization-based procedure for deriving criterion weights. Tzeng and Huang [22] compared ELECTRE and TOPSIS in multi-criteria decision environments, while Liu et al. [23] compared TOPSIS and VIKOR under a unified MCDM framework. These studies demonstrate the usefulness of MCDM methods for integrating multiple criteria and also indicate that ranking outcomes may depend on the selected decision-making method, criterion-weighting scheme, and characteristics of the decision matrix. Such findings emphasize the importance of robustness and sensitivity analysis in multi-criteria ranking. MCDM has also been applied directly to evaluating and ranking machine learning classifiers. Kou et al. [54] evaluated classification algorithms using multiple MCDM methods and examined the agreement between the resulting rankings using rank-correlation measures. Chowdhury et al. [53] combined entropy-based criterion weighting with TOPSIS to rank machine-learning models, illustrating the applicability of data-driven weighting and MCDM to classifier selection. These studies establish the relevance of MCDM for comparing classification models. However, they focus mainly on the ranking problem and do not specifically address reconstructing and validating missing classification-performance metrics before ranking.

A related line of research has explicitly considered incomplete information in MCDM. Ma et al. [55] developed a model for decision-making with missing, imprecise, and uncertain evaluations across multiple criteria. Labreuche and Destercke [56] investigated methods for handling missing values in multi-criteria decision aiding, while Topal et al. [57] proposed a method for handling missing values in multi-criteria sorting problems based on dominance rough sets. These studies demonstrate that incomplete decision matrices constitute an established problem in MCDM. Nevertheless, they generally address missing evaluations at the decision-matrix level rather than the specific reconstruction and consistency assessment of classification-performance metrics such as accuracy, precision, recall, specificity, and F1-score.

Another relevant research direction concerns the consistency of reported machine-learning performance metrics. Reported values such as accuracy, precision, recall, specificity, and F1-score are related through the underlying confusion matrix and therefore cannot always be treated as independent quantities. Kovács and Fazekas [58] introduced *mlscorecheck*, a framework for testing the numerical consistency of reported machine-learning performance scores and experiments. Related studies have also investigated the consistency of reported performance scores in classification problems using numerical feasibility and consistency checks. This line of research is particularly relevant to evaluating reported classification results because a set of individually plausible metric values may still be mutually incompatible. Consistency checking and missing-value reconstruction, however, represent related but distinct problems. A consistency-checking method can identify potentially incompatible reported metrics without necessarily estimating unavailable values. Conversely, an imputation method can reconstruct a missing value without determining whether the resulting metric profile is structurally consistent with the remaining reported values. Therefore, jointly considering metric consistency, missing-value reconstruction, and subsequent model ranking remains a relevant methodological challenge.

Integrating machine learning with MCDM offers a promising strategy for supporting complex decisions with incomplete or uncertain data. For instance, Alqaysi et al. [24] proposed a hybrid framework for autism spectrum disorder detection that employed mean imputation to address missing values and the fuzzy decision by opinion score method (FDOSM) to rank alternatives. Baryannis et al. [11] integrated interpretable machine learning with TOPSIS to develop a supplier-selection framework that combines predictive modeling with multi-criteria decision analysis. Ulutaş et al. [25] proposed a hybrid grey MCDM–machine learning framework for resilient supplier selection, using machine learning to model supplier-related criteria. Lim et al. [26]

proposed an uncertainty-aware imputation framework that estimates missing values together with their associated uncertainty for decision-making under incomplete data. Jinbo et al. [27] introduced an XGBoost–multiple imputation by chained equations (XGBoost–MICE) method for imputing missing values in mine-ventilation datasets. Kang et al. [28] proposed a TOPSIS-based framework for decision-making with incomplete evaluation matrices, using a similarity-based strategy to estimate missing values and improve ranking under incomplete information. These studies illustrate that machine learning and MCDM can be combined effectively. However, they address different application-specific decision problems and do not jointly focus on classification-performance metrics, structural consistency, missingness patterns, and preserving classification-model rankings.

Despite considerable advances in missing-data imputation [50–52], neural-network-based imputation [13–17], MCDM-based ranking [53,54], incomplete-information decision-making [55–57], and performance-score consistency assessment [58], the literature still has several important limitations. First, many proposed frameworks are tailored to specific application domains, such as medical diagnosis, software defect prediction, and supplier selection, which may limit their direct generalization to broader classification-model evaluation problems [11]. Second, numerous studies either assume that performance data are sufficiently complete or focus primarily on estimating missing values without considering their subsequent effect on model ranking. Third, consistency-checking methods and imputation methods are generally developed as separate procedures, although both can be relevant when reported classification metrics are incomplete or potentially inconsistent. Fourth, existing studies generally evaluate imputation quality separately from its downstream effect on multi-criteria ranking. Thus, accurate cell-level reconstruction does not necessarily preserve the resulting model ranking.

Accordingly, a generalizable, automated framework is still needed to identify potentially inconsistent performance metrics, estimate missing values, provide model-based confidence information for validity screening, and support ranking classification models using multiple evaluation criteria. Addressing these requirements would help bridge the gap between missing-data imputation, performance-metric consistency assessment, and MCDM-based ranking while facilitating more transparent and reproducible comparisons across classification studies. Importantly, the contribution of the present study should not be understood as the introduction of a new imputation algorithm or a new MCDM method in isolation. Rather, the methodological contribution lies in integrating mask-aware learning, metric-consistency assessment, missing-value reconstruction, and entropy-weighted MCDM within a single workflow. In particular, explicitly representing the missingness pattern provides information unavailable to conventional imputation models, while linking the subsequent ranking task connects cell-level data reconstruction to the decision-level objective of comparative model evaluation. From a practical perspective, integrated, user-friendly tools that automatically process incomplete or potentially inconsistent classification-performance data and support subsequent model ranking remain relatively limited [9].

Table 1. Comparison of Data Imputation and Model Ranking Methods

Category	Field / Application	Methods / Combination	Data Type	Main Limitation / Scope
Missing-value imputation	General missing-data imputation [50]	Mean imputation / conventional substitution	Numerical data	Simple approach; may not preserve relationships among variables
	MCAR data imputation [13]	MLP for data completion	Numerical data	Does not specifically address consistency of reported classification metrics
	Missing-value estimation [51]	KNN imputation	Numerical / tabular data	Depends on similarity structure and selected neighborhood
	Missing-value completion [52]	MICE	Multivariate data	Iterative modeling can increase computational cost
	Classification of incomplete data [14,15]	IMLP / Augmented MLP	Incomplete data	Does not explicitly use a missingness mask for classification-performance metrics
	Mixed-data completion [10]	MissForest	Mixed data	General-purpose imputation; does not address downstream model ranking
	Incomplete-data completion [16]	GAIN	General / tabular data	Requires adversarial training and parameter tuning

	Tabular-data imputation [17]	Score-based diffusion model	Tabular data	Relatively high computational complexity
MCDM methods	Multi-criteria decision making [12]	TOPSIS	Multi-criteria data	Standard formulation generally requires a specified decision matrix
	Classifier selection [53,18]	Entropy + TOPSIS	Classification metrics	Does not specifically reconstruct missing classification criteria
	Classification algorithm evaluation [54]	Multiple MCDM methods + rank correlation	Classification metrics	Focuses on ranking rather than missing-metric reconstruction
	Risk prioritization [19]	Entropy-weighted TOPSIS + sensitivity analysis	Multi-criteria data	Does not specifically consider missing classification criteria
	Industrial ranking [20]	FAHP–Entropy–TOPSIS + AHP comparison	Industrial data	Does not focus on incomplete classification metrics
	Project selection [21]	TOPSIS–OPA	Project data	Application-specific decision problem
	Multi-criteria decision making [22]	ELECTRE vs. TOPSIS	Multi-criteria data	Focuses on alternative MCDM formulations
	Industrial ranking [23]	TOPSIS vs. VIKOR	Multi-criteria data	Ranking may vary with methodological assumptions
Incomplete MCDM	Missing, imprecise, and uncertain evaluations [55]	MCDM with incomplete/uncertain evaluations	Incomplete multi-criteria data	General MCDM setting rather than classification metrics
	Multi-criteria decision aiding [56]	Missing-value handling in MCDM	Incomplete multi-criteria data	General MCDM setting rather than classification metrics
	Multi-criteria sorting [57]	Dominance rough sets for missing values	Incomplete decision matrices	Does not specifically model classification-metric relationships
Metric consistency	Machine-learning performance reporting [58]	Numerical consistency checking	Reported ML metrics	Identifies potential inconsistencies but does not reconstruct missing values
Hybrid ML & MCDM	Autism detection [24]	ML + MCDM with mean imputation	Medical data	Application-specific and uses simple imputation
	Supplier selection [11]	Interpretable ML + TOPSIS	Industrial data	Does not focus on classification-performance metrics
	Supplier selection [25]	Grey MCDM–ML	Industrial data	Application-specific supplier-selection problem
	Uncertainty-aware imputation [26]	U-VAE	Tabular data	Does not specifically rank classification models
	Mining data completion [27]	XGBoost–MICE	Mining data	Does not specifically assess classification-metric consistency
	Decision making with incomplete information [28]	Incomplete-matrix TOPSIS	Incomplete decision matrices	Does not specifically model relationships among classification metrics

Building on these findings, the proposed framework integrates mask-aware metric reconstruction, consistency screening, and entropy-weighted TOPSIS to evaluate models with incomplete and potentially inconsistent performance data.

3. PROPOSED METHOD

This study presents an automated framework to identify potentially erroneous values, impute missing values, and rank scientific articles based on numerical performance metrics reported for classification tasks. The framework begins by generating a comprehensive simulated dataset from confusion matrices representing the outputs of classification models across diverse scenarios. We then derive standard performance metrics, including accuracy, precision, recall, specificity, and the F1 score, from the corresponding confusion matrix components. We introduce controlled patterns of missingness and inconsistency into the generated data and then use them to train a neural network to estimate missing values and flag potentially erroneous observations. The proposed framework comprises three main stages: (1) generating simulated datasets subject to predefined validity constraints, (2) preprocessing the data and training the predictive models, and (3) assessing the reliability of the reported metrics and ranking

scientific articles using a MCDM method. The complete workflow is implemented through a graphical user interface that enables researchers to enter or upload performance data, estimate missing values, identify observations requiring further scrutiny, assess data reliability, and visualize the resulting article rankings.

We construct valid simulated datasets by reconstructing classification-model confusion-matrix outputs in both binary and multiclass scenarios. This procedure enables the internally consistent calculation of standard performance metrics. To approximate the heterogeneity observed in empirical studies, we generate the underlying classification outcomes using a range of prespecified probability distributions. We then simulate erroneous and missing observations by perturbing selected values with controlled noise and randomly masking a proportion of the data, respectively. We label each observation as valid or invalid according to predefined consistency criteria, enabling the model to distinguish internally consistent metric values from deliberately corrupted ones. Following feature scaling to ensure comparability across variables, the resulting dataset is used to train a mask-aware multilayer perceptron.

The trained model estimates missing or potentially inconsistent classification performance metrics reported across studies, and the corresponding confidence scores indicate the reliability of these estimates. The resulting metric set is subsequently used to compare and rank studies through an MCDM procedure. A graphical user interface enables users to enter data, visualize prediction results, inspect reliability scores, and interactively retrieve study rankings. By combining missing-value estimation, reliability assessment, and multi-criteria ranking in a single workflow, the proposed method supports more consistent and transparent comparisons of classification models. The method comprises four main steps, as outlined below.

3.1 Create a valid simulated dataset to train the Mask-aware MLP model

To train the mask-aware multilayer perceptron, we generate simulated datasets by constructing confusion matrices that represent plausible outputs from classification models in binary and multiclass scenarios, with the number of classes in multiclass settings ranging from 3 to 10. For multiclass scenarios, we calculate both macro-averaged and weighted-averaged metrics to reflect common reporting schemes. We then derive standard performance metrics, including accuracy, precision, recall, specificity, and the F1 score, from the corresponding confusion matrix components. The primary objective is to generate datasets that approximate the distributional characteristics of performance metrics reported in scientific studies, thereby providing controlled training data for estimating missing values and flagging potentially erroneous or inconsistent entries. We parameterize the data-generation process using probabilistic distributions, including Dirichlet, multinomial, and beta distributions. To emulate reporting errors or inconsistencies commonly found in published studies, we introduce Gaussian noise into selected portions of the generated data. We randomly draw the total sample size from the interval [800, 1000] to increase variability across simulated datasets. The following subsections describe the procedures for generating binary and multiclass datasets.

3.1.1 Valid Simulated Binary Dataset

In the binary classification setting, as illustrated in Figure 1, data are generated by constructing a confusion matrix with four components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The proportion of positive-class samples, denoted by π , is sampled from a uniform distribution over the interval [0.3, 0.7] to represent approximately balanced to moderately imbalanced class distributions. Given the total sample size N , the numbers of positive and negative samples are calculated as $N_{pos} = \pi \cdot N$ and $N_{neg} = (1 - \pi) \cdot N$, respectively. The class-specific probability of a correct prediction is sampled from a beta distribution with parameters $\alpha = 4$ and $\beta = 1$, with a mean of 0.8. In this study, this sampling strategy is used to generate a broad distribution of classifier accuracies, primarily within the range [0.6, 1.0]. The confusion matrix entries are then calculated as follows:

$$TP = \lfloor P_{pos} * N_{pos} \rfloor, P_{pos} \sim \text{Beta}(4, 1) \quad (1)$$

$$FN = N_{pos} - TP \quad (2)$$

$$FP = \lfloor (1 - P_{neg}) * N_{neg} \rfloor, P_{neg} \sim \text{Beta}(4, 1) \quad (3)$$

$$TN = N_{neg} - FP \quad (4)$$

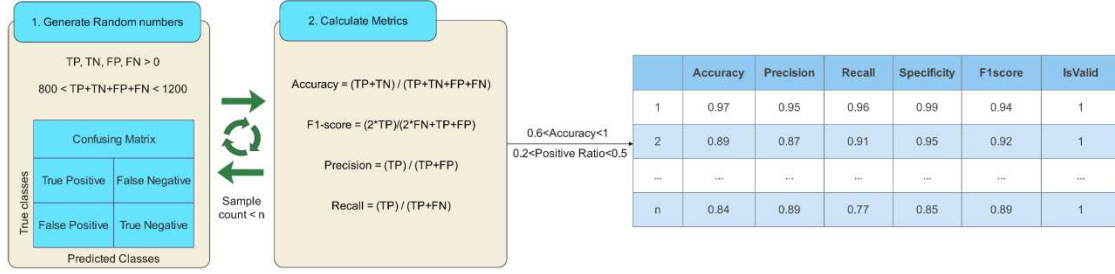


Figure 1: Process for Generating Valid Simulated Binary Dataset

The Beta distribution is defined by the following probability density function:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, x \in [0, 1] \quad (5)$$

Here, $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and Γ is the Gamma function. With $\alpha = 4$ and $\beta = 1$, the mean of the distribution is $\alpha/(\alpha + \beta) = 4/5 = 0.8$, which produces diverse accuracies in the range [0.6–1.0]. Then, the evaluation metrics are calculated using the following formulas [29,30]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 - Score = \frac{2 * TP}{2 * TP + FP + FN} \quad (9)$$

Unique metric combinations are retained, duplicate entries are removed, and the IsValid column is set to True for all generated records.

3.1.2 Valid Simulated Multiclass Dataset

For the multiclass setting, as illustrated in Figure 2, a confusion matrix is generated for K classes (K ranging from 3 to 10), along with the corresponding numerical values.

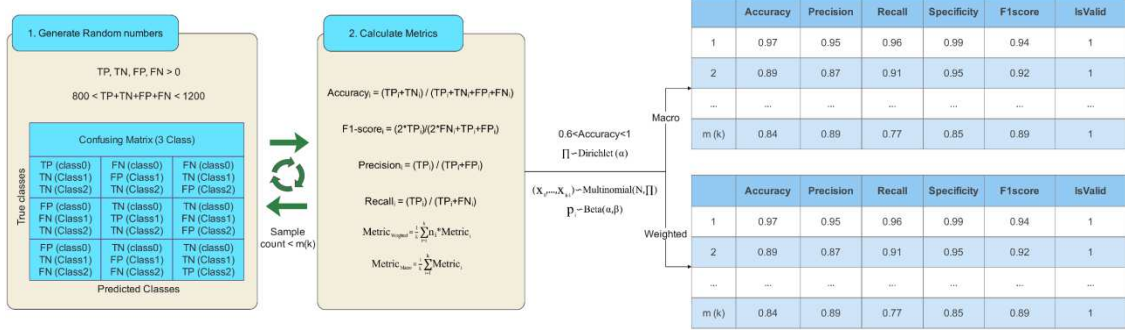


Figure 2: Process for Generating Valid Simulated Multiclass Dataset

This procedure is applied consistently across all class counts from 3 to 10. For example, in the three-class setting, class proportions and confusion matrix entries are generated for three classes; similarly, the same procedure is repeated for four classes, and so forth, while the underlying formulas and distributions remain unchanged. The class proportion vector, $\pi = (\pi_0, \pi_1, \dots, \pi_{K-1})$, is sampled from a Dirichlet distribution to produce balanced yet realistic class distributions. In this study, the Dirichlet distribution is parameterized with $\alpha_i = 5$ for each class. Its probability density function is given by:

$$f(\pi_0, \pi_1, \pi_{k-1}, \alpha_0, \alpha_{k-1}) = \frac{1}{B(\alpha_0, \dots, \alpha_{K-1})} \prod_{i=0}^{K-1} \pi_i^{\alpha_i-1} \quad (10)$$

Here, $\sum_{i=0}^{K-1} \pi_i = 1$ and $\pi_i \geq 0$, and $B(\alpha_0, \dots, \alpha_{K-1}) = \frac{\prod_{i=0}^{K-1} \Gamma(\alpha_i)}{\Gamma(\sum_{i=0}^{K-1} \alpha_i)}$ is the multivariate Beta function. With $\alpha_i = 5$, the distribution is balanced but has enough diversity. This process is the same for all $k=3$ to 10. The number of samples for each class is calculated using a multinomial distribution:

$$(X_0, X_1, \dots, X_{K-1}) \sim \text{Multinomial}(N, \pi) \quad (11)$$

The Multinomial distribution is defined by the following probability:

$$P(X_0, \dots, X_{K-1}; N, \pi) = \frac{N!}{X_0! \dots X_{K-1}!} \prod_{i=0}^{K-1} \pi_i^{X_i} \quad (12)$$

Here, $N = \sum_{i=0}^{K-1} X_i$ and $\mathbb{E}[X_i] = N \cdot \pi_i$. These calculations are the same for all class numbers from 3 to 10. For each class i , the probability of correct prediction p_i is sampled from a Beta(4,1) distribution, and the confusion matrix values are calculated as follows:

$$TP_i = \lfloor p_i \cdot X_i \rfloor, p_i \sim \text{Beta}(4,1) \quad (13)$$

$$FN_i = X_i - TP_i \quad (14)$$

$$FP_i = \sum_{j \neq i} [(1 - p_j) \cdot \pi_j \cdot X_j] \quad (15)$$

$$TN_i = \sum_{j \neq i} (X_j - FP_j) \quad (16)$$

The Beta distribution with $\alpha = 4$ and $\beta = 1$ has the density $f(x; 4,1) = 4x^3$ and a mean of $4/(4+1) = 0.8$, which produces diverse accuracies in the range [0.6, 1.0]. This distribution is the same for all class numbers ranging from 3 to 10. The metrics for each class are calculated as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (17)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (18)$$

$$Specificity_i = \frac{TN_i}{TN_i + FP_i} \quad (19)$$

$$F1-Score_i = \frac{2 * TP_i}{2 * TP_i + FP_i + FN_i} \quad (20)$$

These calculations are repeated for all classes in each model, with class numbers ranging from 3 to 10. Finally, to provide a single overall value for the whole model, two averaging approaches are used:

Macro Average: The simple mean of the values calculated for all classes, without considering the relative size of each class:

$$Metric_{Macro} = \frac{1}{K} \sum_{i=1}^k Metric_i \quad (21)$$

Weighted Average: The weighted mean of the metrics based on the proportion of real samples in each class to the total data, reflecting the effect of class imbalance.

$$Metric_{weighted} = \sum_{i=1}^k \frac{n_i}{N} * Metric_i \quad (22)$$

n_i is the number of real samples in class i . Only samples with Accuracy in the range [0.6, 1.0] and unique combinations are kept, and the IsValid column is set to True.

3.1.3 Noisy Data Generation

The noise-free datasets are partitioned into 2 equal subsets. One subset is left unchanged, and its IsValid attribute is set to True. The other half is perturbed with Gaussian noise to simulate study errors, such as calculation errors or incorrect reporting, as shown in Figure 3.

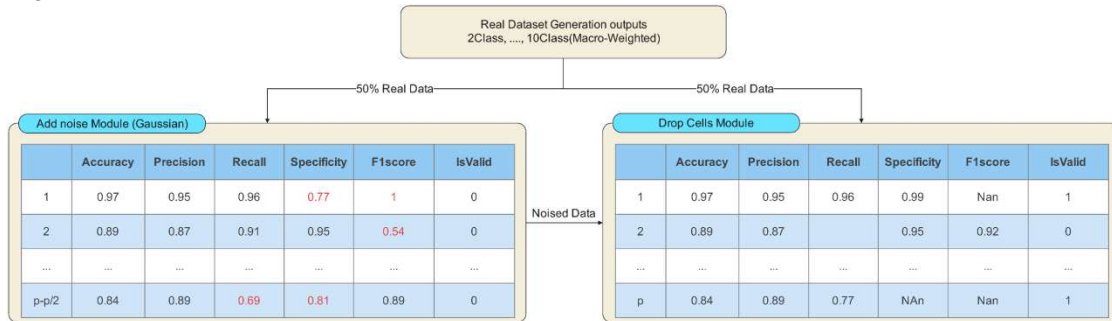


Figure 3: Simulation of Noisy Datasets with Random Feature Removal

For each noisy sample, between 1 and 5 performance metrics are randomly selected, and Gaussian noise with a mean of 0 and a standard deviation of 0.1, 0.2, or 0.3 is added to the selected metrics. The following probability density function defines the Gaussian distribution:

$$f(x; \mu, \delta) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left(-\frac{(x - \mu)^2}{2\delta^2}\right) \quad (23)$$

Here, $\mu = 0$ and $\sigma \in \{0.1, 0.2, 0.3\}$. The values are clipped to the range $[0, 1]$ after adding noise, and the samples are labeled with `IsValid = False`. All clean and noisy data are saved in a CSV file with the `IsValid` label.

3.1.4 Random Feature Removal

To simulate incomplete data and train the MaskAwareMLP model to estimate missing feature values, selected features are randomly masked in a subset of the samples. This procedure is applied to both valid and noise-perturbed samples. For each selected sample, the number of features to be masked is determined according to the following probability distribution:

- 50% probability of no change (no features are removed)
- 30% probability of removing one feature
- 15% probability of removing two features
- 5% probability of removing three features

The features Accuracy, Precision, Recall, Specificity, and F1-score are randomly selected, and their values are set to NaN. This procedure enables the model to learn how to estimate missing or unreported performance metrics in published studies. To simulate incompleteness without altering the records' validity status, the `IsValid` label is left unchanged.

3.2 Preprocessing and Model Training

With the clean, noise-perturbed, and incomplete datasets prepared, the study proceeds to the preprocessing and model training phases. This stage prepares the data for the MaskAwareMLP network, enabling it to impute missing values and classify records as valid or erroneous. The following subsections describe the protocols for data preprocessing, model training, and performance assessment.

3.2.1 Data Preprocessing

After generation, we stored the clean, noise-perturbed, and feature-missing datasets as CSV files, merged them into a single dataset, and randomly shuffled it. We then divided the resulting dataset into training and testing subsets using an 80:20 stratified split to preserve the distribution of the `IsValid` labels across both subsets. For each sample, we constructed a binary observation mask to encode feature availability, assigning observed non-NaN values 1 and missing values (NaN) 0. The resulting mask was provided as an auxiliary input to the MaskAwareMLP network, enabling the model to explicitly identify missing-value locations and distinguish between observed and subsequently imputed feature values during training and inference.

3.2.2 Training of Mask-Aware MLP Models

Because neural networks require numerical inputs and preprocessing procedures such as standardization are more stable on complete feature matrices, we first imputed missing values using the KNN Imputer method. In this approach, each missing value is estimated from the values of its nearest neighboring samples, thereby producing a complete numerical dataset suitable for subsequent model input. Importantly, we retained the binary observation mask as an auxiliary input, allowing the MaskAwareMLP model to distinguish between originally observed and imputed feature values during training and inference. This design enables the model to incorporate missingness information explicitly, rather than treating imputed values as fully equivalent to observed measurements. After imputation, we standardized the features using StandardScaler to achieve zero mean and unit variance, ensuring all variables contributed on a comparable scale during model training.

As shown in Figure 4, we developed mask-aware MLP networks using evaluation metrics and observation masks as inputs. We trained a single model for binary datasets, while we trained two models for multiclass datasets with 3 to 10 classes using macro- and weighted-average metrics. This architecture supports missing-value estimation and classification of valid and erroneous records.

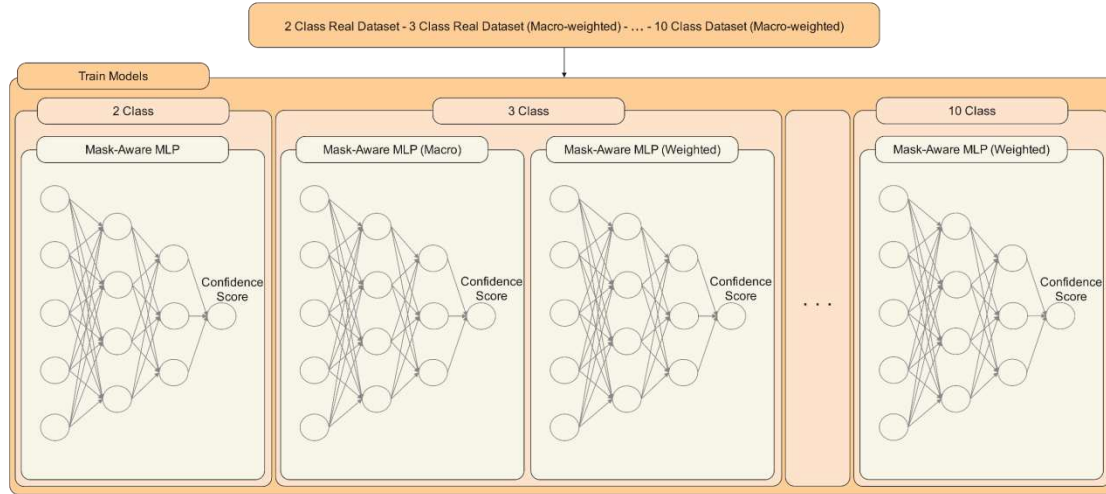


Figure 4: Mask-aware MLP Models

We trained the models in a supervised manner using preprocessed feature data and corresponding *IsValid* labels, enabling them to learn relationships between evaluation metrics and sample validity. For each model, we generated a confusion matrix to provide a detailed assessment of classification performance across datasets with class cardinalities ranging from 2 to 10. Figure 5 presents the resulting confusion matrices, providing a visual representation of class-specific prediction patterns, including correctly classified and misclassified samples. Following training, each model was serialized in HDF5 (.h5) format, while the fitted standardization scaler and imputation object were stored as pickle (.pkl) files. Preserving these preprocessing objects ensures that the same transformations applied during model development can be consistently reproduced during deployment within the application.

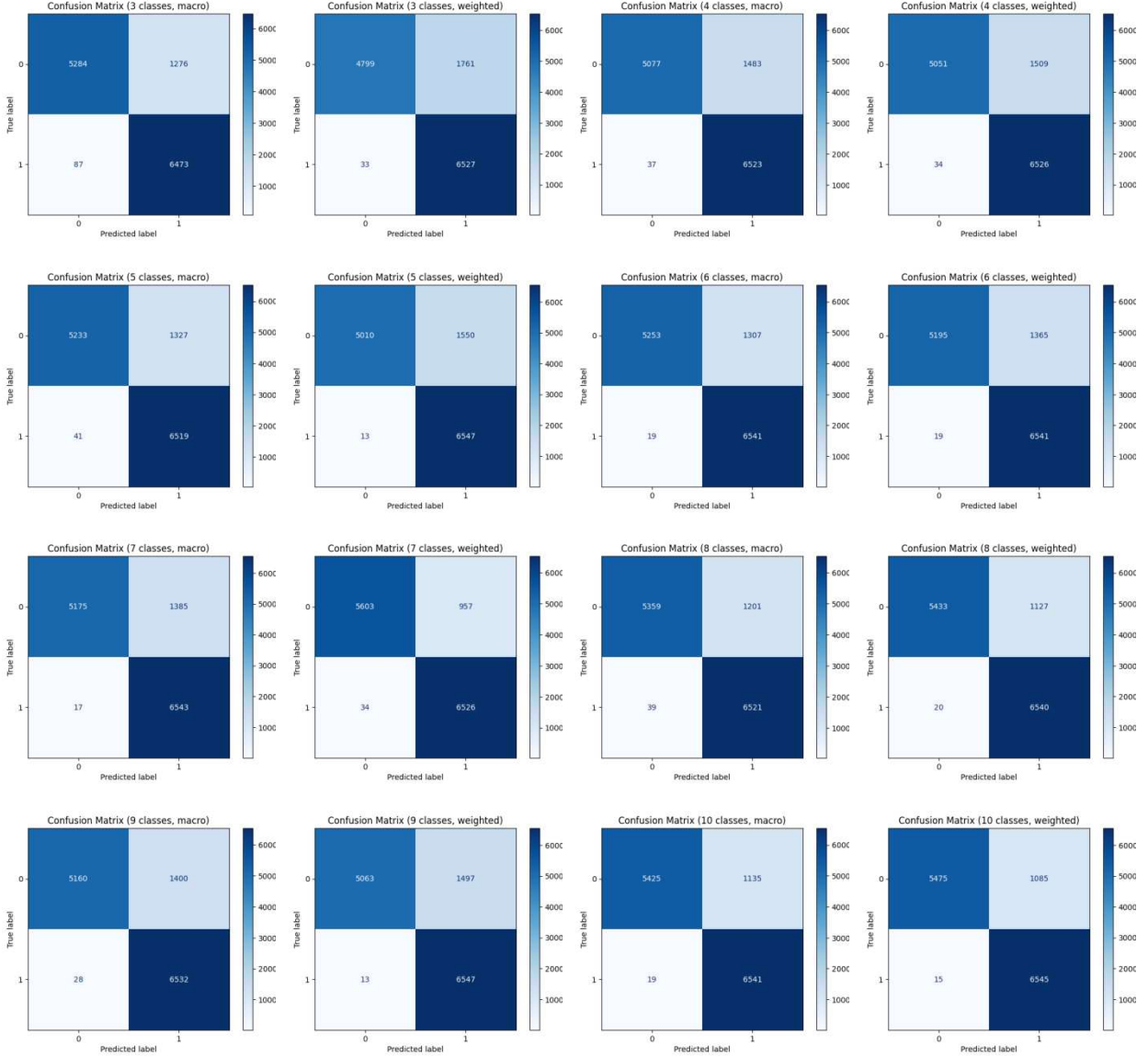


Figure 5: Confusion Matrices of the Models

3.3 Prediction and Ranking of Articles Using MCDM

After training, the MaskAwareMLP model was used to estimate missing metric values and determine article validity. To enable the comparative assessment and ranking of multiple articles, a MCDM framework was then employed. The following subsection describes the procedural steps involved in prediction, evaluation, and ranking.

3.3.1 Prediction of Article Validity and Imputation of Missing Values

As depicted in Figure 6, the evaluation metrics are initially fed into the model.

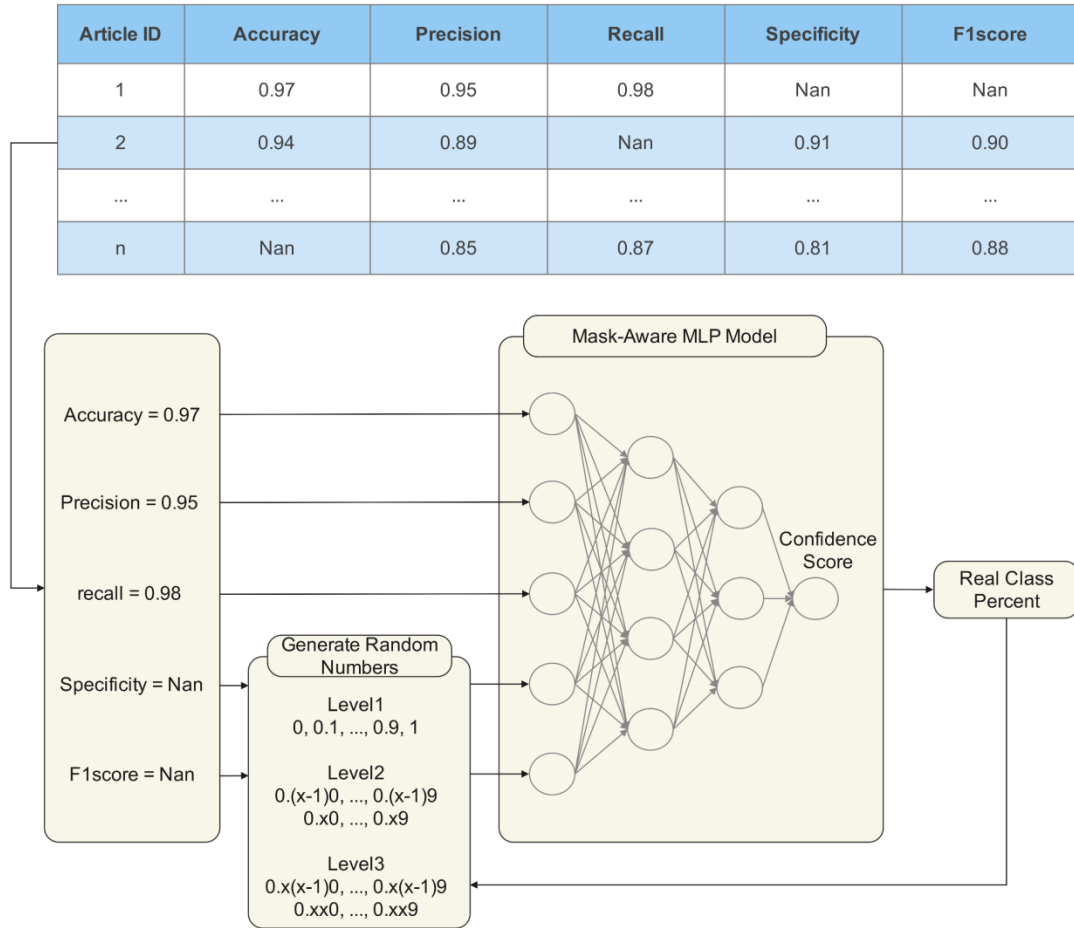


Figure 6: Steps of Validity Prediction

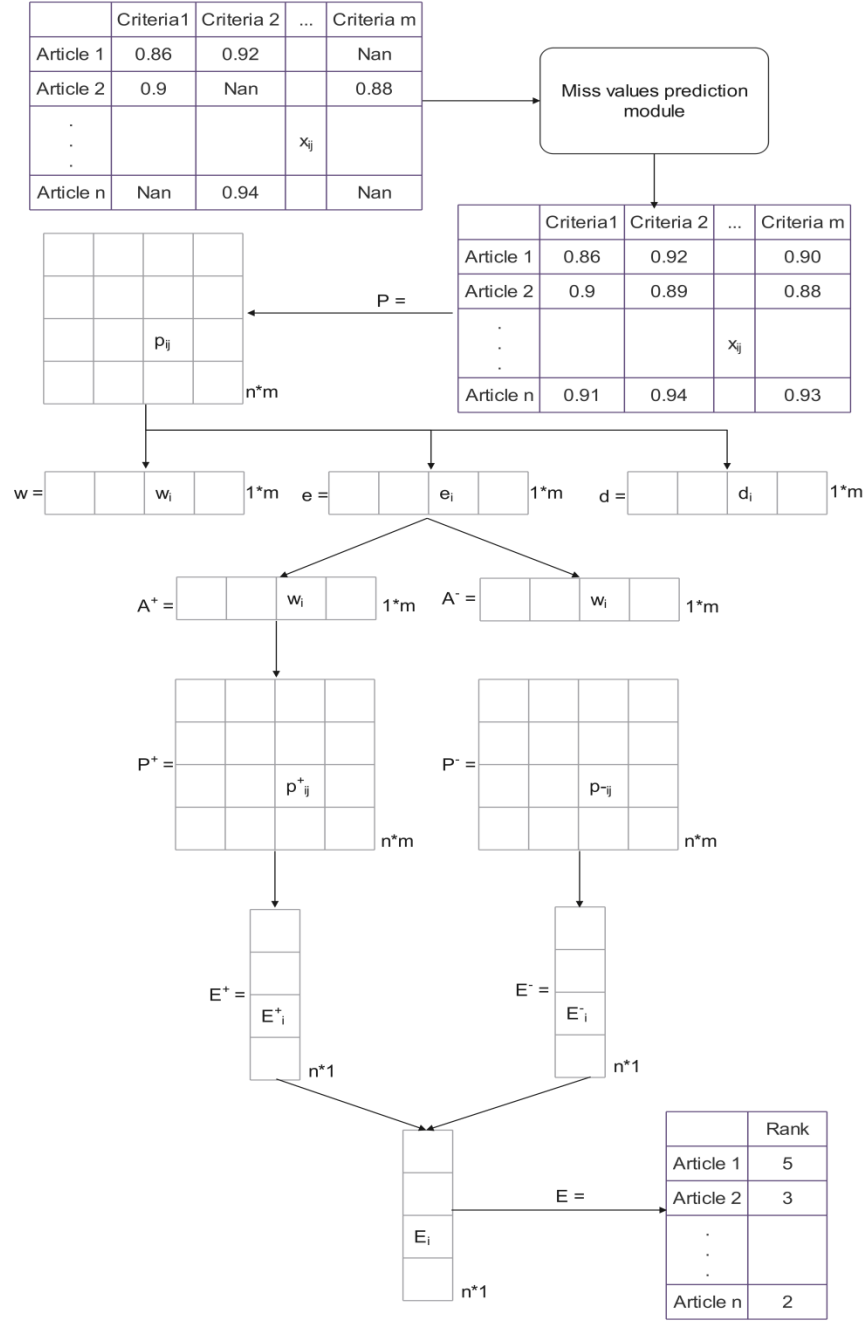


Figure 7: MCDM

If missing values (NaN) are present, the MaskAwareMLP neural network estimates them using a multi-stage, multi-dimensional search procedure. This procedure consists of three successive phases: coarse, fine, and super-fine. Across these phases, the neural network generates candidate values for the missing features. It evaluates them using the model's confidence scores, identifying candidates closer to the optimal solution. In the coarse phase, the search is initialized with relatively large step sizes, considering 11 candidate values per missing feature. The procedure is then progressively refined in the fine and super-fine phases using smaller step sizes, with 20 candidate values per feature at each stage, to obtain a more precise final estimate. As reported in Table 2, the computational cost of this imputation procedure increases with the number of missing features. Specifically, imputing one missing feature requires 51 model evaluations, whereas imputing two, three, and four missing features requires 921, 17,331, and 334,641

evaluations, respectively. After imputation, the resulting data are fed into the MaskAwareMLP model to predict whether the article is valid or invalid.

Table 1: Number of Steps Required for Estimating Missing Values

Number of Missing Values	Coarse Stage	Fine Stage	Extra-Fine Stage	Total Runs
1	11	20	20	51
2	121 (11 ²)	400 (20 ²)	400 (20 ²)	921
3	1,331 (11 ³)	8,000 (20 ³)	8,000 (20 ³)	17,331
4	14,641 (11 ⁴)	160,000 (20 ⁴)	160,000 (20 ⁴)	334,641

3.3.2 Comparison and Ranking of Articles

To compare and rank the articles, we first evaluate each article using the MaskAwareMLP model. Only articles whose evaluation-metric profiles are classified as valid are retained for the subsequent ranking stage. As shown in Figure 7, we then rank the eligible articles using the proposed MCDM framework. Within this framework, the entropy-weighting method objectively derives criterion weights by quantifying each criterion's dispersion across the set of articles. Criteria with greater discriminatory power are assigned higher weights, reflecting their relative contribution to the ranking process. The resulting entropy weights are then combined with the evaluation metric values to compute an aggregate score and determine each article's final rank.

3.4 Proposed Tool for Evaluation and Ranking

These processes were implemented in an interactive graphical user interface that allows users to enter evaluation metric values, predict article validity, and rank articles. The interface presents the resulting scores and ranks in tabular form, facilitating direct comparison of articles based on their evaluation performance. The application consists of the following components:

3.4.1 Data Input

The user enters performance metrics, such as accuracy, precision, recall, specificity, and F1-score, for a given article. When one or more values are missing, the application automatically estimates them using a mask-aware multilayer perceptron neural network combined with a multi-stage search procedure. This procedure is guided by the number of classes and the selected averaging strategy, while simultaneously assessing the reliability of the reported performance-metric profile.

3.4.2 Article Ranking

As illustrated in Figure 8, when multiple articles are compared, the application first estimates any missing metric values and highlights them in red. It then identifies articles whose evaluation-metric profiles are classified as valid and retains them for the ranking stage. Criterion weights are subsequently derived using the entropy-weighting method, after which the eligible articles are ranked using the selected MCDM method. The resulting numerical scores and ordinal ranks are presented for each eligible article.

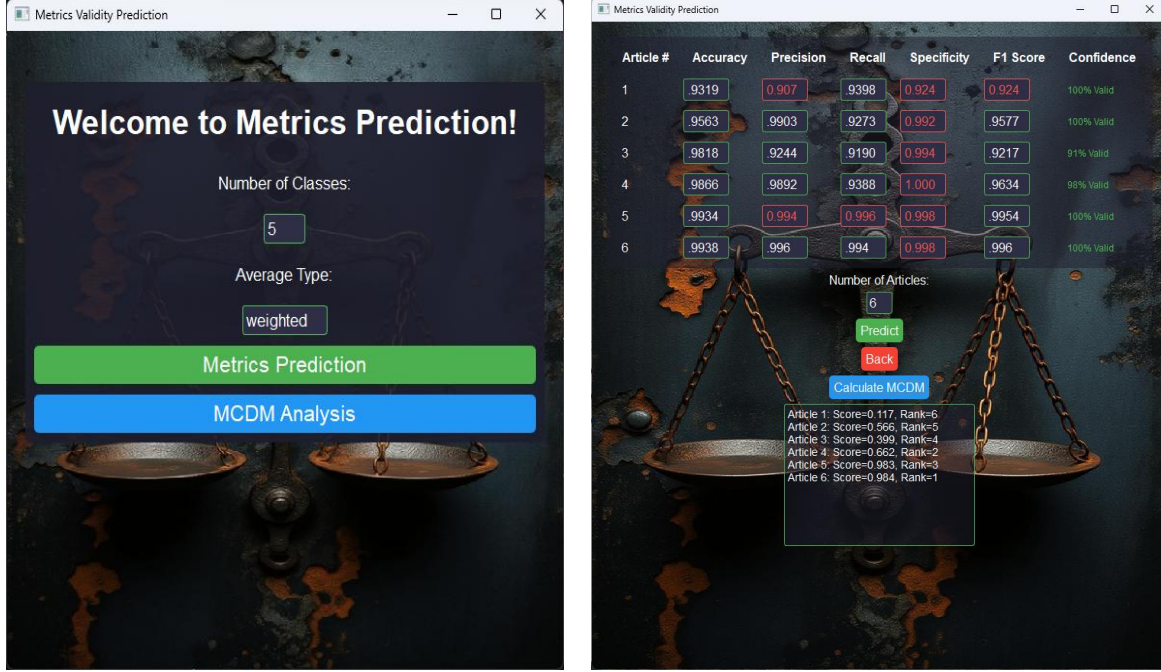


Figure 8: Developed Application Schema

4. RESULTS

This section presents the empirical results of the proposed framework using simulated classification-performance datasets and selected classification studies from medical and image-processing domains. The evaluation covers binary and multiclass classification settings and considers complete, noisy, and incomplete metric profiles. The results are presented in five main stages: (i) evaluation of the validity-classification component, (ii) analysis of reported classification metrics from published studies, (iii) comparison of missing-value imputation methods, (iv) evaluation of ranking preservation under controlled missingness, and (v) assessment of ranking consistency across different classification configurations. In addition, we conduct an ablation study to isolate the contribution of the explicit missingness mask in the proposed neural imputation architecture. This design allows evaluation of both the reconstruction of missing metrics and the extent to which that reconstruction preserves the ranking obtained from complete reference data.

4.1 Hyperparameters

Table 3 summarizes the hyperparameters used for data preprocessing, model architecture, training, and evaluation, together with the sample sizes of the generated clean and noise-perturbed datasets. The preprocessing configuration included an 80:20 training–test split and K-nearest-neighbor (KNN) imputation to produce complete numerical feature matrices for the validity-classification component. A binary observation mask, consisting of 0/1 indicators, was provided to the MaskAwareMLP alongside the imputed feature values, enabling the model to distinguish originally observed entries from values reconstructed during preprocessing. Rather than treating imputed values as equivalent to originally observed measurements, the mask explicitly indicates feature availability and allows the network to incorporate missing-data patterns during prediction.

The validity-classification model comprised five dense layers, with batch normalization and LeakyReLU activations incorporated to support stable optimization and nonlinear representation learning. The model was trained for 50 epochs using the Adam optimizer with a learning rate of 0.001. Class weights were applied during training to reduce the influence of class imbalance during model optimization.

For the missing-value reconstruction experiments, we also evaluated the proposed MaskAwareMLP against mean imputation, KNN, MICE, and an iterative Random-Forest-based (MissForest-style) approach. We evaluated these methods using identical

missingness realizations at each missingness level and assessed performance using cell-level reconstruction errors and preservation of the complete-data MCDM ranking.

Table 3: Model Hyperparameter and Dataset Summary

Category	Parameter	Value / Description
Data Processing	Train/Test Split	80/20
	Validation Split	10% of training data
	Imputation	KNN Imputer (k=5)
	Normalization	StandardScaler
	Class Weighting	Balanced
	Missing Mask	Binary vector (0/1)
Model Structure	Dense Layers	5
	Layer Dimensions	2048 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 64 $\rightarrow$ 16
	Batch Normalization	After each Dense
	Activation Function	LeakyReLU ($\alpha=0.1$)
	Dropout	0.2, 0.2, 0.2, 0.1
	Output Layer	Softmax (2 classes: Valid/Invalid)
Model Training	Loss Function	Categorical Crossentropy
	Optimizer	Adam
	Learning Rate	0.001
	Batch Size	32
	Epochs	50
	Early Stopping	None
Number of Samples in Valid Simulated Dataset	Clean Data	32,799
	Noisy Data	32,799
	Total	65,598

4.2 Performance of the Validity-Classification Component

We evaluated the validity-classification component using simulated datasets generated from binary and multiclass classification settings. We developed separate models for the binary case and for multiclass settings with 3–10 classes, using both macro- and weighted-average metrics for the multiclass configurations. This resulted in 17 classification configurations. Across these configurations, the validity-classification models achieved classification accuracies ranging from 0.8961 to 0.9274, while the corresponding F1 scores ranged from approximately 0.90 to 0.93. These results indicate that the MaskAwareMLP-based validity classifier distinguished clean and noise-perturbed metric profiles with relatively consistent predictive performance across different class configurations and averaging strategies.

Interpret the results as performance on simulated validity labels rather than as direct evidence that a reported result in the scientific literature is objectively correct or incorrect. In the simulated data, the validity label is determined by the predefined data-generation and corruption procedures. Accordingly, the model's output represents a model-based confidence estimate regarding consistency with the simulated validity patterns.

4.3 Analysis of Results from Scientific Articles

We applied the proposed framework to metric profiles reported in several published studies, as described in the following subsections. In the corresponding tables, values highlighted in red were not reported in the original studies and were therefore estimated using the proposed framework. The resulting confidence scores represent the output of the validity-classification model and should be interpreted as screening indicators rather than formal statistical probabilities of correctness.

Table 4, adapted from Ref. [35], compares several methods for cardiac arrhythmia detection. Applying the proposed framework revealed a potential inconsistency in the evaluation metrics reported in Ref. [32]. Specifically, that study did not provide a confusion matrix, preventing independent verification of the reported metrics from the underlying classification outcomes. Moreover, recalculation of the F1-score from the reported precision (96.29%) and recall (85.38%) yields approximately 90.51%, whereas the study reports an F1-score of 80.37%. This discrepancy suggests the reported metrics may have come from different averaging

strategies, datasets, experimental runs, or metric definitions. If these factors do not explain the difference, the reported F1-score requires clarification or correction by the original authors.

Table 4: Binary Classification Metrics

Article	Accuracy	Precision	Recall	Specificity	F1-Score	Confidence	Score (Rank)
[31]	88.59	87.34	90.07	0.867	88.69	100% Valid	0 (4)
[32]	95.79	96.29	85.38	1.00	80.37	100% Invalid	–
[33]	98.15	97.31	96.85	0.989	97.79	98% Valid	0.847 (3)
[34]	98.28	99.90	97.72	0.996	0.988	100% Valid	0.935 (2)
[35]	99.24	99.09	99.02	0.999	99.05	100% Valid	0.968 (1)

Application of the proposed framework to Table 5, adapted from the article [39] on glioma classification in 3D MRI images, indicates that all evaluated metric profiles were classified as valid according to the learned screening criteria.

Table 5: Binary Classification Metrics

Article	Accuracy	Precision	Recall	Specificity	F1-Score	Confidence	Score (Rank)
[36]	0.9730	0.961	0.9732	0.9734	0.967	100% Valid	0 (4)
[37]	0.9750	1.00	0.9743	1.00	0.985	98% Valid	0.575 (2)
[38]	0.9890	0.979	0.9810	0.9870	0.9810	99% Valid	0.464 (3)
[39]	0.9993	1.00	0.9987	1.00	0.9993	100% Valid	1.00 (1)

Table 6, adapted from article [44], which focuses on breast tumor segmentation and classification in ultrasound images, indicates that all evaluated metric profiles satisfy the internal consistency criteria defined by the proposed screening framework.

Table 6: Ternary Classification Metrics

Article	Accuracy	Precision	Recall	Specificity	F1-Score	Confidence	Score (Rank)
[40]	0.9462	0.90	0.9231	0.934	0.9114	93% Valid	0.689 (3)
[41]	0.843	0.751	0.7580	1.0	0.7544	68% Valid	0.196 (4)
[42]	0.928	0.9334	0.9255	0.959	0.9293	99% Valid	0.755 (2)
[43]	0.8119	0.8103	0.7612	0.895	0.7811	86% Valid	0.143 (5)
[44]	0.9790	0.9802	0.9790	0.986	0.9789	99% Valid	0.971 (1)

Table 7, adapted from article [49], which focuses on electrocardiogram heartbeat interpretation using convolutional and recurrent neural networks, indicates that the metric profile attributed to article [48] was classified as inconsistent by the proposed framework. Manual examination of article [48] further indicated that the reported metric values listed in Table 7 could not be located in the original publication.

Table 7: Five-Class Classification Metrics

Article	Accuracy	Precision	Recall	Specificity	F1-Score	Confidence	Score (Rank)
[45]	0.9755	0.965	0.9705	0.988	0.9744	99% Valid	0.649 (2)
[46]	0.9350	0.932	0.9335	0.978	0.9341	99% Valid	0 (4)
[47]	0.9495	0.943	0.9546	0.983	0.9491	99% Valid	0.271 (3)
[48]	0.9902	0.833	0.7632	0.858	0.8097	99% Invalid	–
[49]	0.9961	0.983	0.9933	0.994	0.9942	99% Valid	1.00 (1)

In summary, the proposed framework, combining mask-aware validity screening with entropy-weighted MCDM, was evaluated using results reported for binary, three-class, and five-class classification tasks in the scientific literature. We conducted the analysis using data compiled from Refs. [35], [39], [44], and [49]. Most evaluated performance-metric profiles were classified as valid, and most corresponding cases received confidence scores above 98%. Three potential reporting issues were identified. First, the F1-score attributed to Ref. [32] was inconsistent with the value calculated from its reported precision and recall, and no confusion matrix was available for independent verification. Second, the metric profile reported in Ref. [41] received a comparatively low validity confidence score and was likewise not accompanied by a confusion matrix. Third, closer examination indicated that the values attributed to Ref. [48] could not be located in the original publication. These observations do not by themselves establish that the corresponding studies contain incorrect results. Rather, they indicate metric profiles that warrant further examination. The flagged cases should therefore be checked against the original experimental settings, metric definitions, averaging strategies, and

reported confusion matrices, where available. The results demonstrate the proposed framework's potential as a screening tool for identifying reported classification metrics that may be inconsistent, incomplete, or insufficiently documented.

4.4 Missing-Value Imputation Performance

To evaluate the missing-value reconstruction component independently of the validity-classification results, we introduced controlled missingness into the complete simulated metric profiles. We considered four missingness levels: 10%, 20%, 30%, and 40%. Five repeated missingness realizations were generated at each level, and the same missing-cell patterns were used for all competing methods within each repetition.

Mean imputation, KNN, MICE, an iterative Random-Forest-based (MissForest-style) method, a conventional MLP, and the proposed MaskAwareMLP were compared. We evaluated performance only on artificially masked cells using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2).

Table 8: Cell-Level Missing-Value Imputation Performance

Method	Missingness	MAE	RMSE	R^2
KNN	10%	0.064950	0.110759	0.280273
	20%	0.067163	0.111492	0.262986
	30%	0.069191	0.111674	0.253265
	40%	0.072272	0.113230	0.231587
MICE	10%	0.056491	0.103731	0.368580
	20%	0.061128	0.106435	0.328301
	30%	0.066223	0.110689	0.266342
	40%	0.072824	0.117172	0.177066
MLP	10%	0.060926	0.108767	0.305818
	20%	0.062517	0.108900	0.296853
	30%	0.064248	0.109214	0.285801
	40%	0.066454	0.110541	0.267652
MaskAwareMLP	10%	0.060785	0.108481	0.309512
	20%	0.061340	0.108413	0.303131
	30%	0.062721	0.108922	0.289607
	40%	0.064546	0.110176	0.272478
Mean	10%	0.087508	0.119063	0.168331
	20%	0.087070	0.118474	0.167752
	30%	0.086884	0.117921	0.167404
	40%	0.086807	0.117813	0.168131
MissForest-style	10%	0.056388	0.103765	0.368202
	20%	0.061009	0.106419	0.328501
	30%	0.066268	0.110657	0.266762
	40%	0.072198	0.116738	0.183202

The results show that no single method consistently achieved the lowest reconstruction error across all missingness levels, as shown in Table 8. At 10% missingness, the MissForest-style method achieved the lowest MAE (0.056388), followed closely by MICE (0.056491), representing the two lowest MAE values at this missingness level. At 20% missingness, the MissForest-style method again achieved the lowest MAE (0.061009), while MICE (0.061128) and MaskAwareMLP (0.061340) remained competitive, as shown in Table 8. At 30% and 40% missingness, the performance pattern changed. MaskAwareMLP achieved the lowest MAE among all evaluated methods at both levels, with values of 0.062721 and 0.064546, respectively. At 40% missingness, for example, MaskAwareMLP's MAE was lower than that of KNN (0.072272), MICE (0.072824), and MissForest-style imputation (0.072198). The R^2 results showed a similar pattern, with MaskAwareMLP achieving the highest R^2 at both 30% and 40% missingness.

A direct comparison between the conventional MLP and MaskAwareMLP further demonstrates the effect of explicitly incorporating the missingness pattern. MaskAwareMLP produced lower MAE and RMSE values and higher R^2 values than the conventional MLP at all four missingness levels. The improvement in MAE increased from approximately 0.23% at 10%

missingness to approximately 2.87% at 40% missingness. These results suggest that the observation mask provides useful information for the neural reconstruction model, particularly as missingness increases.

Overall, the cell-level results indicate that conventional iterative methods remain highly effective under low missingness, whereas the proposed mask-aware model becomes increasingly competitive under moderate-to-high missingness. Importantly, these results alone do not establish the universal superiority of MaskAwareMLP; instead, they motivate evaluating whether its reconstruction behavior provides an advantage for the downstream ranking task.

4.5 Ablation Study of the Missingness-Aware Architecture

We conducted an ablation study to assess the contribution of the explicit missingness mask in the proposed MaskAwareMLP. To do so, we compared MaskAwareMLP with a conventional MLP under the same experimental conditions at four missingness levels: 10%, 20%, 30%, and 40%. We analyzed both the cell level, using imputation accuracy metrics, and the decision level, using ranking preservation measures.

4.5.1 Cell-Level Imputation Ablation

The cell-level imputation results are presented in Table 9. The evaluation was performed only on artificially masked cells using MAE, RMSE, and R^2 .

Table 9 Ablation comparison of MLP and MaskAwareMLP for missing-value imputation

Missingness	Model	MAE ↓	RMSE ↓	R^2 ↑
10%	MLP	0.060926	0.108767	0.305818
	MaskAwareMLP	0.060785	0.108481	0.309512
20%	MLP	0.062517	0.108900	0.296853
	MaskAwareMLP	0.061340	0.108413	0.303131
30%	MLP	0.064248	0.109214	0.285801
	MaskAwareMLP	0.062721	0.108922	0.289607
40%	MLP	0.066454	0.110541	0.267652
	MaskAwareMLP	0.064546	0.110176	0.272478

As shown in Table 9, the MaskAwareMLP consistently achieved better performance than the conventional MLP across all missingness levels. The improvement in MAE became more noticeable as missingness increased, with relative reductions of approximately 0.23%, 1.88%, 2.38%, and 2.87% at 10%, 20%, 30%, and 40% missingness, respectively. These results suggest that explicitly representing the missingness pattern provides additional information that benefits the reconstruction process, particularly under higher levels of missingness.

4.5.2 Downstream Ranking Ablation

To determine whether the improvement in reconstruction also affects the final decision process, the reconstructed profiles were subsequently processed using the same entropy-weighted TOPSIS procedure. The resulting rankings were compared with the complete-data reference ranking using Kendall's τ -b and Spearman's ρ . The results are reported in Table 10.

Table 10. Ablation comparison of MLP and MaskAwareMLP for ranking preservation

Missingness	Model	Kendall's τ -b ↑	Spearman's ρ ↑
10%	MLP	0.9074	0.9704
	MaskAwareMLP	0.9213	0.9738
20%	MLP	0.8390	0.9412
	MaskAwareMLP	0.8557	0.9457
30%	MLP	0.7634	0.9017
	MaskAwareMLP	0.7879	0.9106
40%	MLP	0.6982	0.8580
	MaskAwareMLP	0.7165	0.8651

As shown in Table 10, the MaskAwareMLP achieved higher ranking agreement than the conventional MLP at all missingness levels. The advantage became more evident as missingness increased; for example, at 40% missingness, Kendall's τ -b increased from 0.6982 to 0.7165, while Spearman's ρ increased from 0.8580 to 0.8651. Overall, the ablation results indicate that the explicit

missingness mask contributes to both improved metric reconstruction and better preservation of downstream rankings. Importantly, these findings demonstrate the contribution of the mask relative to the conventional MLP, rather than establishing universal superiority over all competing imputation methods.

4.6 Ranking Consistency Analysis

4.6.1 Ranking Preservation under Missingness

Cell-level imputation accuracy does not necessarily preserve the final MCDM ranking. Therefore, we conducted a separate ranking-preservation experiment to evaluate whether the relative ordering of alternatives from incomplete data remained consistent with the ranking from the corresponding complete reference data.

For each missingness level, we used the complete test dataset only to establish the reference ranking. We then introduced the missing values into the corresponding data, reconstructed them using each competing method, and processed them through the same entropy-weighted TOPSIS procedure. We used Kendall's tau-b and Spearman's rank correlation coefficient to quantify agreement between the reconstructed-data ranking and the complete-data reference ranking.

The comparison included mean imputation, KNN, MICE, the iterative Random-Forest-based (MissForest-style) approach, conventional MLP, and MaskAwareMLP. We applied the same missingness realizations to all methods in each repetition to ensure a consistent basis for comparison.

Table 11: Ranking Preservation under Different Missingness Levels

Method	Kendall τ -b (10%)	Kendall τ -b (20%)	Kendall τ -b (30%)	Kendall τ -b (40%)	Spearman ρ (10%)	Spearman ρ (20%)	Spearman ρ (30%)	Spearman ρ (40%)
KNN	0.9190	0.8512	0.7859	0.7147	0.9734	0.9446	0.9104	0.8640
MICE	0.9235	0.8560	0.7867	0.7099	0.9766	0.9473	0.9082	0.8545
MLP	0.9074	0.8390	0.7634	0.6982	0.9704	0.9412	0.9017	0.8580
MaskAwareMLP	0.9213	0.8557	0.7879	0.7165	0.9738	0.9457	0.9106	0.8651
Mean	0.8872	0.8029	0.7305	0.6645	0.9638	0.9272	0.8873	0.8423
MissForest-style	0.9252	0.8579	0.7874	0.7092	0.9773	0.9478	0.9080	0.8529

The ranking-preservation results complement the cell-level imputation results. As shown in Table 11, ranking agreement decreased as the proportion of missing values increased for all methods. Nevertheless, all evaluated approaches maintained substantial positive agreement with the complete-data reference ranking.

At 10% and 20% missingness, the iterative Random-Forest-based method achieved the highest Kendall's tau-b values (0.9252 and 0.8579, respectively). MICE also performed strongly at these lower missingness levels. Thus, the proposed method did not preserve rankings as well under low missingness.

In contrast, MaskAwareMLP achieved the highest Kendall's tau-b at both 30% and 40% missingness, with values of 0.7879 and 0.7165, respectively. At 30% missingness, the difference from the strongest competing methods was small: MaskAwareMLP achieved 0.7879 compared with 0.7874 for the MissForest-style method and 0.7867 for MICE. At 40% missingness, the advantage was more pronounced, with MaskAwareMLP reaching 0.7165 compared with 0.7147 for KNN, 0.7099 for MICE, and 0.7092 for the MissForest-style method.

The Spearman results show a consistent pattern. MaskAwareMLP achieved the highest Spearman correlation at 30% and 40% missingness, with values of 0.9106 and 0.8651, respectively. At 10% and 20% missingness, the MissForest-style approach provided the highest Spearman correlations (0.9773 and 0.9478, respectively). These results reinforce the observation that the proposed approach is particularly competitive under moderate-to-high missingness rather than uniformly superior across all missingness levels.

The comparison between the two neural approaches is also informative. MaskAwareMLP consistently produced higher Kendall's tau-b and Spearman correlations than the conventional MLP. For example, at 40% missingness, Kendall's tau-b increased from 0.6982 for MLP to 0.7165 for MaskAwareMLP, while Spearman's rho increased from 0.8580 to 0.8651. This consistent improvement supports the importance of explicitly representing missingness patterns when using a neural network for metric reconstruction.

Overall, the ranking-preservation experiment indicates that the benefit of the proposed mask-aware formulation is more apparent when missingness is moderate or high. The results also demonstrate that imputation quality and ranking quality are not identical objectives: a method with strong cell-level reconstruction performance at low missingness does not necessarily remain the best method for preserving the final MCDM ranking at higher missingness levels.

4.6.2 Interpretation of Kendall's Tau-b and Spearman's ρ

Kendall's tau-b and Spearman's rank correlation coefficient were used to quantify the agreement between the reference ranking and the ranking obtained after missing-value reconstruction and MCDM processing. Kendall's tau-b was selected because it evaluates pairwise ordinal agreement and explicitly accounts for tied ranks.

Let $X = \{x_1, x_2, \dots, x_n\}$ denote the reference ranking scores and $Y = \{y_1, y_2, \dots, y_n\}$ denote the corresponding scores produced by the proposed framework for n alternatives. For any pair of alternatives (i, j) , where $i < j$, the pair is defined as concordant if the relative order of the two alternatives is the same in both rankings:

$$(x_i - x_j)(y_i - y_j) > 0$$

Conversely, the pair is defined as discordant if the relative order is reversed between the two rankings:

$$(x_i - x_j)(y_i - y_j) < 0$$

Pairs for which $x_i = x_j$ or $y_i = y_j$ are treated as tied pairs. Kendall's tau-b corrects for such ties and is computed as follows:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_X)(C + D + T_Y)}} \quad (24)$$

where C denotes the number of concordant pairs, D denotes the number of discordant pairs, T_X denotes the number of pairs tied only in the reference ranking, and T_Y denotes the number of pairs tied only in the proposed ranking. The coefficient τ_b ranges from -1 to $+1$, where $+1$ indicates perfect agreement, 0 indicates no ordinal association, and -1 indicates a completely reversed ranking order.

Spearman's ρ was additionally used as a complementary rank-based measure to assess the preservation of the relative ordering of alternatives across different missingness conditions.

4.7 Ranking Consistency across Classification Configurations

In addition to the controlled missingness experiments, we evaluated the ranking behavior of the classification configurations across simulated binary and multiclass settings. Table 12 reports Kendall's tau-b values obtained by comparing the ranking generated by the proposed framework with the corresponding reference ranking.

Table 12: Kendall's Tau-b Ranking Consistency Results

Num. Classes	Average Type	Ranking Samples	Kendall's tau-b	p-value
3	macro	420	0.904807	7.158288e-169
3	weighted	408	0.909043	1.381683e-165
4	macro	410	0.917586	1.691672e-169
4	weighted	421	0.899695	2.181767e-167
5	macro	412	0.923015	2.555492e-172
5	weighted	408	0.912415	8.409509e-167
6	macro	407	0.924886	6.439413e-171
6	weighted	404	0.922242	1.050194e-168
7	macro	409	0.908409	9.219617e-166
7	weighted	413	0.924775	2.184134e-173
8	macro	413	0.890265	7.731515e-161
8	weighted	412	0.920535	2.095637e-171
9	macro	420	0.925332	1.587504e-176
9	weighted	409	0.904957	1.608844e-164
10	macro	408	0.893144	6.468476e-160
10	weighted	402	0.916453	8.485176e-166

Across the evaluated classification configurations, Kendall's tau-b values ranged from 0.890265 to 0.925332, indicating strong positive ordinal agreement between the reference and framework-generated rankings. The lowest value was observed for the 8-class macro configuration (0.890265), whereas the highest value was obtained for the 9-class macro configuration (0.925332). All corresponding p-values were substantially below 0.001, indicating that the observed associations were statistically significant. However, interpret statistical significance alongside the magnitude of the rank correlation. Given the relatively large number of ranking pairs, very small p-values are expected even for moderate deviations from the null hypothesis. Therefore, the high Kendall's tau-b coefficients, rather than the p-values alone, provide the primary evidence of ranking agreement.

Overall, the results in Table 12 indicate that the framework can maintain a strong ordinal relationship with the reference ranking across different numbers of classes and averaging strategies. These results complement the controlled missingness experiment reported in Section 4.5 by demonstrating ranking consistency across a broader range of simulated classification configurations.

4.8 Overall Results Summary

The experimental results provide evidence for three complementary aspects of the proposed framework. First, the validity-classification component achieved relatively consistent classification performance across binary and multiclass simulated settings, with accuracy values ranging from 0.8961 to 0.9274 and F1 scores of approximately 0.90–0.93. Second, the scientific-article analysis showed the framework's practical use for screening reported metric profiles and identifying cases requiring further verification, including discrepancies involving reported F1 scores and metric values that could not be located in the corresponding source publication.

Third, and most importantly for evaluating the integration of missing-value handling with MCDM ranking, the controlled experiments demonstrated that no single imputation method was optimal across all missingness levels. MICE and the iterative Random-Forest-based approach performed particularly well at lower missingness levels, whereas MaskAwareMLP became increasingly competitive as missingness increased. MaskAwareMLP achieved the lowest cell-level MAE among the compared methods at 30% and 40% missingness and consistently outperformed the conventional MLP.

The ranking-preservation analysis further showed that the proposed method achieved the highest Kendall's tau-b and Spearman's rho at 30% and 40% missingness. In particular, Kendall's tau-b values of 0.7879 and 0.7165 were obtained at 30% and 40% missingness, respectively. These results support the hypothesis that explicitly incorporating missingness information can improve the robustness of downstream MCDM rankings under moderate-to-high missingness.

At the same time, the results do not support a claim of universal superiority. Conventional iterative methods achieved higher ranking agreement at 10% and 20% missingness, and the differences at 30% missingness were relatively small. Therefore, the main empirical contribution of the proposed approach is not that MaskAwareMLP always outperforms conventional imputation methods, but that explicitly representing missingness patterns provides a consistent advantage over a conventional MLP and becomes particularly useful for preserving MCDM rankings when missingness is moderate to high.

5. DISCUSSION

The results in Section 4 show the proposed framework's potential to support evaluating and ranking classification models when reported performance metrics are incomplete or potentially inconsistent. By integrating mask-aware neural reconstruction, validity screening, and entropy-weighted MCDM, the framework addresses several stages of evaluation in a unified workflow. The validity-screening component provides an automated mechanism to assess the consistency of reported metric profiles, while the reconstruction component estimates metric values not available in the corresponding comparison tables. These capabilities support an automated decision-support framework for organizing, reconstructing, and comparatively analyzing classification-performance results.

A key finding is that missing-value reconstruction and ranking preservation are related but distinct objectives. At the cell level, iterative methods such as MICE and the iterative Random-Forest-based approach were particularly competitive at lower missingness levels. In contrast, MaskAwareMLP achieved the best reconstruction performance among the evaluated methods at 30% and 40% missingness. This pattern suggests that explicitly representing missingness information becomes more useful as the proportion of unavailable metric values increases. At lower missingness levels, established iterative methods can provide highly competitive estimates from the available information, while MaskAwareMLP remains competitive.

The downstream ranking analysis provides an additional perspective on imputation performance. After reconstruction, we processed all methods using the same entropy-weighted MCDM procedure and compared them with the complete-data reference ranking. MaskAwareMLP achieved the highest Kendall's τ -b at 30% and 40% missingness (0.7879 and 0.7165, respectively), whereas the iterative Random-Forest-based approach achieved the highest values at 10% and 20% missingness. Spearman's ρ showed a similar pattern. These findings indicate that MaskAwareMLP's relative advantage becomes more apparent under moderate-to-high missingness, while its performance remains competitive with alternative methods at lower missingness levels. Overall, the results indicate that no single imputation approach consistently performs best across all missingness conditions.

The ablation analysis further supports the contribution of the explicit observation mask. Under identical experimental conditions, MaskAwareMLP consistently achieved lower reconstruction errors and higher R^2 values than the conventional MLP and also produced higher ranking agreement at all missingness levels. The improvement increased with missingness, supporting the hypothesis that explicitly distinguishing observed and missing metric values provides useful information for neural reconstruction and subsequent decision analysis. These results highlight the missingness mask as an informative component of the proposed neural architecture rather than merely an auxiliary preprocessing step.

The ranking-consistency analysis across classification configurations also indicated stable ordinal agreement, with Kendall's τ -b ranging from 0.890265 to 0.925332 across the evaluated class configurations and averaging strategies. Although all corresponding p-values were statistically significant, the magnitude of Kendall's τ -b is more informative than statistical significance alone for assessing practical ranking agreement, particularly when the number of pairwise comparisons is large. The consistently high positive values across the evaluated configurations indicate that the ranking procedure maintains substantial ordinal agreement across different classification settings.

The proposed framework's principal contribution is integrating established methodological components rather than introducing a fundamentally new imputation or MCDM method in isolation. Existing approaches such as MissForest and GAIN primarily address missing-data reconstruction, whereas entropy-weighted TOPSIS and related MCDM methods focus on multi-criteria ranking. The proposed framework connects these areas by incorporating missingness-pattern information into neural reconstruction, assessing metric-profile consistency, and evaluating the effect of reconstruction on downstream rankings. This decision-oriented perspective matters because accurate cell-level reconstruction does not necessarily preserve the final ranking. Evaluating both reconstruction quality and ranking consistency therefore provides a more comprehensive assessment of missing-value reconstruction effectiveness in classification-performance analysis.

Another important consideration is that entropy-based weighting depends on the reconstructed decision matrix. Consequently, imputation can influence both individual metric values and the relative importance assigned to criteria. The resulting ranking differences therefore reflect the combined effects of reconstruction, entropy weighting, and MCDM aggregation. This further justifies evaluating imputation methods using downstream ranking-preservation measures in addition to MAE, RMSE, and R^2 . This evaluation shows whether reconstructed metric profiles remain useful for subsequent comparative decision-making.

From a practical perspective, the framework provides a decision-support approach for organizing, reconstructing, and comparatively analyzing reported classification-performance metrics. The graphical tool also supports data entry and upload, visualization, metric-profile screening, and article ranking, facilitating the proposed workflow in a user-oriented environment. These capabilities may support more systematic and transparent comparison of classification models in domains where reliable performance evaluation is important, including medical diagnosis and industrial process analysis.

Overall, the findings indicate that explicit missingness information can improve neural reconstruction and better preserve downstream MCDM rankings, particularly under moderate-to-high missingness. At lower missingness levels, established iterative methods remain highly competitive, while MaskAwareMLP remains competitive. The combined reconstruction and ranking results therefore show the value of evaluating imputation methods not only by cell-level prediction accuracy but also by their effect on subsequent model ranking. Future work should focus on broader real-world evaluation, more diverse missingness mechanisms, improved calibration of validity-confidence estimates, additional experimental repetitions, and ranking-aware learning objectives.

6. LIMITATIONS

Interpret the results of this study in light of the characteristics of the proposed training data generation framework. The developed MaskAwareMLP model demonstrated the capability to effectively identify potential inconsistencies among commonly reported

classification performance metrics, including accuracy, precision, recall, specificity, and F1-score. Since the training samples were generated through stochastic processes based on general statistical principles rather than predefined error patterns or manually constructed inconsistency structures, the model was not optimized for a specific type of reporting error. This design reduces the likelihood of overfitting to a particular inconsistency pattern and enables the model to learn general relationships among evaluation metrics.

Nevertheless, the use of predefined statistical mechanisms, including Gaussian perturbation, Beta-based sampling, and Dirichlet-based class distribution generation, represents a limitation of the current study. Although these mechanisms provide a broad, randomized, and unbiased simulation environment for model training, real-world reporting errors in scientific studies may occasionally exhibit more complex structures or follow distributions that differ from those considered in the current framework. Consequently, when the model encounters previously unseen inconsistency patterns beyond the range of simulated distributions, its performance may vary to some extent.

Despite this limitation, the proposed framework's robustness to incomplete and perturbed metric profiles shows its ability to preserve meaningful relationships among performance indicators and identify potentially unreliable metric combinations with high confidence. The evaluation results showed that the model achieved an overall classification accuracy of approximately 92% and yielded confidence levels of approximately 99.9% when identifying reliable or unreliable metric profiles in the evaluated scenarios. These findings indicate that the proposed approach can detect potentially inconsistent evaluation reports while minimizing the unnecessary rejection of valid metric profiles, thereby preserving informative inputs for the subsequent multi-criteria decision-making (MCDM) stage.

The imputation results also showed that MaskAwareMLP performed competitively with the other evaluated methods at 10% and 20% missingness, while performing better at 30% and 40% missingness. These findings indicate that the proposed method remains competitive at lower missingness levels while providing greater benefits as the proportion of missing metric values increases. However, the relative performance of the imputation methods varied across missingness levels, indicating that no single method is uniformly optimal under all experimental conditions.

Future studies should further validate the proposed framework using larger, independently verified real-world datasets, develop more diverse noise-generation strategies that incorporate domain-specific and structurally dependent reporting errors, and investigate the model's sensitivity to low-magnitude inconsistencies. These extensions may further improve the framework's generalizability and practical applicability in applied scientific evaluation scenarios.

7. CONCLUSION AND FUTURE WORKS

This study introduced an automated framework to screen, reconstruct, and rank classification-performance metrics reported in machine learning studies. The proposed framework combines the MaskAwareMLP model to screen potentially inconsistent or incomplete metric profiles and reconstruct missing values with an entropy-weighted multi-criteria decision-making (MCDM) procedure to rank models comparatively. The main contribution of the proposed framework is not the introduction of a new imputation or MCDM algorithm in isolation, but the integration of missingness-aware metric reconstruction, validity screening, and multi-criteria ranking within a single workflow, together with explicit evaluation of downstream ranking preservation.

The imputation experiments showed that MaskAwareMLP performed competitively with the evaluated methods at 10% and 20% missingness, while achieving stronger reconstruction performance at 30% and 40% missingness. The ablation analysis further demonstrated that explicitly incorporating the missingness mask improved both reconstruction performance and downstream ranking preservation relative to a conventional MLP under the same experimental conditions. In the ranking-preservation experiments, MaskAwareMLP achieved the highest Kendall's τ -b at 30% and 40% missingness, with values of 0.7879 and 0.7165, respectively. These findings indicate that the mask-aware formulation can provide particular benefits under moderate-to-high missingness while remaining competitive with alternative imputation methods at lower missingness levels. Overall, the results suggest that no single imputation method is uniformly optimal across all missingness conditions.

The graphical tool supports data entry and upload, visualization, metric-profile screening, and article ranking, thereby facilitating a more efficient preliminary assessment of reported classification results. These capabilities may support more systematic and transparent comparisons of machine learning studies in domains where reliable model evaluation matters, including medical diagnosis and industrial process analysis. The framework provides an integrated workflow for handling missingness, ensuring

metric consistency, enabling reconstruction, and supporting downstream multi-criteria ranking, thereby enabling more systematic evaluation of incomplete classification-performance information.

Although this study focuses on classification evaluation metrics, extending the framework to other machine learning tasks is an important direction for future research. One potential extension is clustering evaluation, where metrics such as the Silhouette Score, Davies-Bouldin Index, and Adjusted Rand Index could be incorporated into appropriate validation and reconstruction scenarios. Such an extension would require designing suitable data-generation procedures for clustering outputs, as well as validation rules tailored to unsupervised learning settings. Future studies should also investigate more diverse and realistic missingness and inconsistency mechanisms, including structured and non-random missingness, domain-specific reporting errors, and low-magnitude inconsistencies. Validation using larger, independently verified real-world datasets, calibrated confidence estimates, and more independent experimental repetitions would further strengthen the empirical evidence and generalizability of the framework. Future work may also investigate using large language models to support the automated extraction, structuring, and preliminary verification of reported experimental results. However, such models should be accompanied by rigorous validation, human oversight, and safeguards against hallucinated or unsupported outputs. In addition, ranking-aware learning objectives could be explored to directly optimize the preservation of downstream rankings rather than evaluating ranking preservation only after imputation. These directions may contribute to more intelligent and reliable tools for scientific evaluation and to a more systematic assessment of incomplete and potentially inconsistent machine learning performance reporting.

COMPETING INTERESTS AND FUNDING

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

CODE AVAILABILITY

The source code and datasets are available at <https://github.com/MasoudKargar/mcdm.git>. The repository is currently private; however, the materials are available to the Editor upon request.

FUNDING

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

REFERENCES

- [1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep learning. MIT Press, Cambridge, MA.
- [2] MdKamrul Hasan, Md Ashrafal Alam, Shidhartho Roy, Aishwariya Dutta, Md Tasnim Jawad, and Sunanda Das. 2021. Missing value imputation affects the performance of machine learning: a review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked* 27 (2021), Article 100799. <https://doi.org/10.1016/j.imu.2021.100799>
- [3] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (July 2009), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- [4] Nathalie Japkowicz and Mohak Shah. 2011. Evaluating learning algorithms: a classification perspective. Cambridge University Press, Cambridge.
- [5] Luciano C. Blomberg and Duncan D. A. Ruiz. 2013. Evaluating the influence of missing data on classification algorithms in data mining applications. In *Proceedings of the 14th Simpósio Brasileiro de Sistemas de Informação (SBSI '13)*, May 2013, João Pessoa, Brazil. SBC, 734–743. <https://doi.org/10.5753/sbsi.2013.5736>
- [6] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine* 25, 1 (Jan. 2019), 24–29. <https://doi.org/10.1038/s41591-018-0316-z>
- [7] Michael I. Jordan and Tom M. Mitchell. 2015. Machine learning: trends, perspectives, and prospects. *Science* 349, 6245 (July 2015), 255–260. <https://doi.org/10.1126/science.aaa8415>
- [8] Ethem Alpaydin. 2020. Introduction to machine learning (4th ed.). MIT Press, Cambridge, MA.
- [9] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, and Nan Liu. 2023. Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine* 142 (Aug. 2023), Article 102587. <https://doi.org/10.1016/j.artmed.2023.102587>
- [10] Daniel J. Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (Jan. 2012), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- [11] Ahmad Abdulla and George Baryannis. 2024. a hybrid multi-criteria decision-making and machine learning approach for explainable supplier selection. *Supply Chain Analytics* 7 (2024), Article 100074. <https://doi.org/10.1016/j.sca.2024.100074>

- [12] Serafim Opricovic and Gwo-Hshiung Tzeng. 2004. Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research* 156, 2 (Apr. 2004), 445–455. [https://doi.org/10.1016/S0377-2217\(03\)00020-1](https://doi.org/10.1016/S0377-2217(03)00020-1)
- [13] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, Manuel López-Coello, and María-Dolores Cubiles-de-la-Vega. 2011. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 24(1), 121–129. <https://doi.org/10.1016/j.neunet.2010.09.006>
- [14] Hao Pan, Zhi Ye, Qiang He, Cheng Yan, Jia Yuan, Xin Lai, and Rui Li. 2022. Discrete missing data imputation using multilayer perceptron and momentum gradient descent. *Sensors* 22, 15 (Aug. 2022), Article 5645. <https://doi.org/10.3390/s22155645>
- [15] Avigyan Bhattacharya, Sreeja Bhose, and Suvra Jyoti Choudhury. 2023. Classification of incomplete data using augmented MLP. In *Proceedings of the 2023 4th International Conference for Emerging Technology (INCET)*. IEEE, Belgaum, India, 1–5. <https://doi.org/10.1109/ICONAT57137.2023.10080793>
- [16] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: missing data imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*, July 10–15, 2018, Stockholm, Sweden. PMLR, 5689–5698. <https://proceedings.mlr.press/v80/yoon18a.html>
- [17] Wan, B., Shi, H., Jing, L., Yan, M., & Wang, D. (2026). Tabular data imputation via local and global information: A score-based diffusion approach. *Pattern Recognition*, 114336.
- [18] Yi Peng, Guoxun Wang, and Honggang Wang. 2012. User preferences based software defect detection algorithms selection using MCDM. *Information Sciences* 191 (May 2012), 3–13. <https://doi.org/10.1016/j.ins.2011.12.006>
- [19] Tunç, M., & Solmaz, R. (2026). A Hybrid Fine-Kinney and Entropy-Weighted TOPSIS Approach for Risk Prioritization in Chemistry Laboratories. *ACS Chemical Health & Safety*.
- [20] Parveen Sharma, Kashmir Singh Ghatrha, Lenka Cepova, NillohitMitra Ray, Ajay Kumar, Saneh Lata Yadav, Vladimira Schindlerova, and Rakesh kumar Phanden. 2025. A hybrid FAHP–entropy–TOPSIS model for selecting the facility layout in small-scale manufacturing. *Frontiers in Mechanical Engineering* 11 (2025), Article 1666571. <https://doi.org/10.3389/fmech.2025.1666571>
- [21] Amin Mahmoudi, Xiaopeng Deng, Saad Ahmed Javed, and Jingfeng Yuan. 2021. Large-scale multiple criteria decision-making with missing values: project selection through TOPSIS-OPA. *Journal of Ambient Intelligence and Humanized Computing* 12, 10 (Oct. 2021), 9341–9362. <https://doi.org/10.1007/s12652-020-02649-w>
- [22] Gwo-Hshiung Tzeng and Jih-Jeng Huang. 2011. Multiple attribute decision making: methods and applications. CRC Press.
- [23] Liu, Y., Wang, J., & Shi, B. (2026). Comparing TOPSIS and VIKOR to prioritize international communication interventions across digital platforms. *Scientific Reports*.
- [24] Mohammed Elhassan Alqaysi, Aso Salah Albahri, and Rosli A. Hamid. 2022. Hybrid diagnosis models for autism patients based on medical and sociodemographic features using machine learning and multicriteria decision-making (MCDM) techniques: an evaluation and benchmarking framework. *Computational and Mathematical Methods in Medicine* 2022 (2022), Article 9410222. <https://doi.org/10.1155/2022/9410222>
- [25] Alptekin Ulutaş, Mladen Krstić, Ayşe Topal, Leonardo Agnusdei, Snežana Tadić, and Pier Paolo Miglietta. 2024. a novel hybrid Gray MCDM model for resilient supplier selection problem. *Mathematics* 12, 10 (May 2024), Article 1444. <https://doi.org/10.3390/math12101444>
- [26] Lim, J., An, S., & Jeon, J. J. (2026, March). Impute missing entries with uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 28, pp. 23514–23522).
- [27] Zhao Jinbo, Li Yufu, and Mo Haitao. 2025. Handling missing data using the XGBoost-based multiple imputation by chained equations regression method. *Frontiers in Artificial Intelligence* 8 (2025), Article 1553220. <https://doi.org/10.3389/frai.2025.1553220>
- [28] Cristian Velasco-Gallego and Iraklis Lazakis. 2020. Real-time data-driven missing data imputation for short-term sensor data of marine systems: a comparative study. *Ocean Engineering* 218 (Nov. 2020), Article 108261. <https://doi.org/10.1016/j.oceaneng.2020.108261>
- [29] Masoud Kargar, Ayaz Isazadeh, and Habib Izadkhah. 2020. New internal metric for software clustering algorithms validity. *IET Software* 14, 4 (Aug. 2020), 402–410. <https://doi.org/10.1049/iet-sen.2019.0138>
- [30] Masoud Kargar, Habib Izadkhah, and Ayaz Isazadeh. 2019. Tarimliq: A new internal metric for software clustering analysis. In *Proceedings of the 27th Iranian Conference on Electrical Engineering (ICEE '19)*, April 2019, Yazd, Iran. IEEE, 1879–1883.
- [31] Pooja Sharma, Shail Kumar Dinkar, and Dinesh V. Gupta. 2021. A novel hybrid deep learning method with cuckoo search algorithm for classification of arrhythmia disease using ECG signals. *Neural Computing and Applications* 33, 19 (Oct. 2021), 13123–13143. <https://doi.org/10.1007/s00521-021-06005-7>
- [32] Sanjay Kumar, Abhishek Mallik, Akshi Kumar, Javier Del Ser, and Guang Yang. 2023. Fuzz-ClustNet: coupled fuzzy clustering and deep neural networks for Arrhythmia detection from ECG signals. *Computers in Biology and Medicine* 153 (Feb. 2023), Article 106511. <https://doi.org/10.1016/j.combiomed.2022.106511>
- [33] Taki Hasan Rafi and Young Woong Ko. 2022. HeartNet: self multihead attention mechanism via convolutional network with adversarial data synthesis for ECG-based arrhythmia classification. *IEEE Access* 10 (2022), 100501–100512. <https://doi.org/10.1109/ACCESS.2022.3206401>
- [34] Bach-Tung Pham, Phuong Thi Le, Tzu-Chiang Tai, Yi-Chiung Hsu, Yung-Hui Li, and Jia-Ching Wang. 2023. Electrocardiogram heartbeat classification for arrhythmias and myocardial infarction. *Sensors* 23, 6 (Mar. 2023), Article 2993. <https://doi.org/10.3390/s23062993>
- [35] Ali Bayani and Masoud Kargar. 2024. LDCNN: a new arrhythmia detection technique with ECG signals using a linear deep convolutional neural network. *Physiological Reports* 12, 17 (Sept. 2024), Article e16182. <https://doi.org/10.14814/phy2.16182>
- [36] KittipolWisaeng. 2023. U-Net++ DSM: Improved U-Net++ for Brain Tumor Segmentation with Deep Supervision Mechanism. *IEEE Access* 11 (2023), 132268–132285. <https://doi.org/10.1109/ACCESS.2023.3329194>
- [37] Md Shaharier Alam, Md Mahbubur Rahman, Mohammad Amazad Hossain, Md Khairul Islam, Kazi Mowdud Ahmed, Khandaker Takdir Ahmed, Bikash Chandra Singh, and Md Sipon Miah. 2019. Automatic human brain tumor detection in MRI image using template-based K means and improved fuzzy C means clustering algorithm. *Big Data and Cognitive Computing* 3, 2 (June 2019), Article 27. <https://doi.org/10.3390/bdcc3020027>
- [38] Ayesha Jabbar, Naseem Shahid, Mahmood Tariq, Saba Tanzila, Faten S. Alamri, and Amjad Rehman. 2023. Brain Tumor Detection and Multi-Grade Segmentation Through Hybrid Caps-VGGNet Model. *IEEE Access* 11 (2023), 72518–72536. <https://doi.org/10.1109/ACCESS.2023.3289224>
- [39] Nayer Seyed Hoseini, Masoud Kargar, Ali Bayani, and SondasArdebili. 2025. HNMD-CNN: ahierarchical narrowing Multi-Deep Convolutional Neural Network for Precision Glioma Classification in 3D MRI Images. *Computational and Systems Oncology* 5, 1 (2025), Article e70002. <https://doi.org/10.1002/cso2.70002>
- [40] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, Chiun-Sheng Huang, and Ruey-Feng Chang. 2020. Computer-Aided Diagnosis of Breast Ultrasound Images Using Ensemble Learning from Convolutional Neural Networks. *Computer Methods and Programs in Biomedicine* 190 (Apr. 2020), Article 105361. <https://doi.org/10.1016/j.cmpb.2020.105361>
- [41] Jie Xing, Chen Chao, Lu Qinyang, Cai Xun, Yu Aijun, Xu Yi, Xia Xiaoling, Sun Yue, Xiao Jing, and Huang Lingyun. 2021. Using BI-RADS Stratifications as Auxiliary Information for Breast Masses Classification in Ultrasound Images. *IEEE Journal of Biomedical and Health Informatics* 25, 6 (June 2021), 2058–2070.

- [42] Hossam Magdy Balaha, Mohamed Saif, Ahmad Tamer, and Ehab H.abdelhay. 2022. Hybrid deep learning and genetic algorithms approach (HMB-DLGAHA) for the early ultrasound diagnoses of breast cancer. *Neural Computing and Applications* 34, 11 (June 2022), 8671–8695.
- [43] Jie Du, Kai Guan, Yanhong Zhou, Yuanman Li, and Tianfu Wang. 2023. Parameter-Free Similarity-Aware Attention Module for Medical Image Classification and Segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 3 (June 2023), 845–857.
- [44] Mohsin Furkh Dar and AvatharamGanivada. 2023. EfficientU-Net: A Novel Deep Learning Method for Breast Tumor Segmentation and Classification in Ultrasound Images. *Neural Processing Letters* 55, 8 (Dec. 2023), 10439–10462.
- [45] Tae Joon Jun, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, and Young-Hak Kim. 2018. ECG Arrhythmia Classification Using a 2-D Convolutional Neural Network. *arXiv:1804.06812*. Retrieved from <https://arxiv.org/abs/1804.06812>
- [46] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. 2017. A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine* 89 (Oct. 2017), 389–396. <https://doi.org/10.1016/j.combiomed.2017.08.022>
- [47] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25, 1 (Jan. 2019), 65–69. <https://doi.org/10.1038/s41591-018-0268-3>
- [48] Eko Ihsanto, Kalamullah Ramli, Dodi Sudiana, and Teddy S. Gunawan. 2020. An Efficient Algorithm for Cardiac Arrhythmia Classification Using Ensemble of Depthwise Separable Convolutional Neural Networks. *Applied Sciences* 10, 2 (Jan. 2020), Article 483. <https://doi.org/10.3390/app10020483>
- [49] Guoliang Yao, Xiaobo Mao, Nan Li, Huaxing Xu, Xiangyang Xu, Yi Jiao, and Jinhong Ni. 2021. Interpretation of electrocardiogram heartbeat by CNN and GRU. *Computational and Mathematical Methods in Medicine* 2021 (2021), Article 6534942. <https://doi.org/10.1155/2021/6534942>
- [50] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10), 1087-1091.
- [51] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.
- [52] Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45, 1-67.
- [53] Chowdhury, N. K., Kabir, M. A., Rahman, M. M., & Islam, S. M. S. (2022). Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. *Computers in Biology and Medicine*, 145, 105405.
- [54] Kou, G., Lu, Y., Peng, Y., & Shi, Y. (2012). Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, 11(01), 197-225.
- [55] Ma, W., Xiong, W., & Luo, X. (2013). A model for decision making with missing, imprecise, and uncertain evaluations of multiple criteria. *International Journal of Intelligent Systems*, 28(2), 152–184. <https://doi.org/10.1002/int.21562>
- [56] Labreuche, C., & Destercke, S. (2019, August). How to handle missing values in multi-criteria decision aiding?. In *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)* (pp. 1756-1763).
- [57] Topal, A., Guler Bayazit, N., & Ucan, Y. (2024). A Method to Handle the Missing Values in Multi-Criteria Sorting Problems Based on Dominance Rough Sets. *Mathematics*, 12(18), 2944.
- [58] Kovács, G., & Fazekas, A. (2024). mlscorecheck: Testing the consistency of reported performance scores and experiments in machine learning. *Neurocomputing*, 583, 127556.