To Impute or Not: Recommendations for Multibiometric Fusion

Melissa R Dale*, Elliot Singer†, Bengt J. Borgström†, Arun Ross*
*Dept of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
†Artificial Intelligence Technology and Systems Group
MIT Lincoln Laboratory
Lexington, MA, USA

2023 IEEE International Workshop on Information Forensics and Security (WIFS) | 979-8-3503-2491-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/WIFS58808.2023.10374772

***Abstract*—Combining match scores in multibiometrics via fusion is a well-established approach to improving recognition performances. However, when scores are missing, this can degrade performance, as well as limit the possible fusion techniques that can be applied. Imputation is a technique for estimating reasonable values to replace missing data and an approach that has previously shown promise in addressing missing scores in multibiometrics. In this paper, we evaluate various approaches to imputation methods on three multimodal biometric score datasets: NIST BSSR1, BIOCOP2008, and MIT LL TRIMODAL, and investigate the factors which might influence the effectiveness of imputation. Our studies reveal three key observations: (1) Imputation is preferable to not imputing missing scores, even when the fusion rule does not necessitate complete score data. (2) Balancing the classes in the training data is crucial to mitigate biases in the imputation technique and prevent favoritism towards the overrepresented class, even if it involves dropping a substantial number of score vectors. (3) Multivariate imputation approaches exhibit better estimation for genuine scores, while univariate imputation approaches yield stronger results for imputed imposter scores.**

***Index Terms*—Imputation, Fusion, Multibiometrics**

I. Introduction

Biometric systems are indispensable for recognizing individuals based on the uniqueness of their biological and behavioral traits such as face, fingerprint, iris, voice, and gait [1]. However, in many real-world applications, the reliance on a single biometric modality may not be sufficient to meet the criteria of high recognition accuracy and enhanced security. As a result, the fusion of multiple biometric modalities or algorithms has become a crucial avenue of research and development. In addition to improving performance and increasing security [2]–[5], using multiple biometrics can also improve accessibility. By incorporating different biometric modalities, biometric systems can accommodate individuals with varying physical characteristics or limitations. For example, individuals who may have difficulty providing a fingerprint due to a physical disability can still participate in the system by utilizing their face or voice as an alternative modality [6], [7]. However if the fusion approach is not designed carefully, the benefits of fusion can be stymied by an unnecessarily convoluted system with slow performance [8].

Biometric fusion can be accomplished at multiple levels, including data, feature, score, decision, and rank levels. In score-level fusion, the match scores from the participating modalities or matchers are combined. One design decision when considering score-level fusion is how to handle missing match score values. Missing scores can arise from various factors within a biometric system. This includes failures in acquiring the biometric sample or encountering samples of insufficient quality. Additionally, the integration of new biometric modalities into an existing system may introduce a discrepancy where the input probe data contains more modalities than the corresponding gallery identities, resulting in a missing data scenario. While some fusion techniques can be applied to data with missing values, such as the simple sum rule, [1] many fusion techniques, however, are not designed to implicitly account for missing scores. In these instances, a choice must be made: either ignore the score vectors that contain missing values or replace missing values with an estimated value (a process known as imputation). If the proportion of missing data is small, simply ignoring those samples may not influence overall aggregate performance. However, if there is a large proportion of missing data, ignoring incomplete data may be harmful to the performance. Imputation can help address these situations. Additionally, it has been shown that implementing imputation techniques can improve biometric recognition performance even when not required by the fusion rule [9]. The authors in [9] show that applying the simple sum rule with the imputed score data frequently improved both verification and identification performance in biometric recognition tasks even when up to 90% of data was incomplete. However, imputation can also add undesirable computational and time complexity to a multibiometric system.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

[1] A score vector consists of scores from multiple matchers.

II. Background

Fusion in multibiometrics encompasses various levels of integration, including data, feature, score, decision, or rank [1]. Among these, score level fusion has received significant attention due to its applicability when working with biometric systems that provide match scores rather than raw data or features.

The simple sum rule is a popular choice of fusion thanks to its straight forward approach that often produces strong recognition accuracies. Simple sum fusion is a transformation-based approach, i.e, since all scores must share a common range, a transformation is required (e.g., normalizing scores into the range of [0-1]). Other approaches to score-level fusion include classifier-based techniques [10], [11] and density-based techniques [12]. These techniques often require the estimation of a number of parameters and, hence, depend on the availability and representativeness of the training data.

When a fusion technique requires score data to be complete, careful consideration must be paid to ***how*** the data is missing. For this analysis, we define a score vector as the vector of match scores between identities i and j, where $\mathbf{s_{ij}} = [s_1, ..., s_m]$ for the m modalities present.

Rubin defines the following patterns of missing values [13]: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Accurately identifying the reason for missing data is vital, as the suitability of imputation methods depends on whether the missingness follows the MCAR or MAR assumptions. It is important to note that MNAR can introduce biases and lead to erroneous conclusions if not properly addressed. Assuming the missingness is either MCAR or MAR, the literature provides various approaches for handling missing data, which are outlined in the following paragraphs.

One option for missing data is to simply ignore vectors with missing scores. This approach is referred to as *Listwise Deletion* [14] and works if there is only a small proportion of missing data and that missing data is truly MCAR. If data is missing because of a failing sensor, for instance, the missing values are due to the sensor and, thus, a bias would be introduced into the analysis. A drawback of this approach is that otherwise usable score vectors are lost.

Imputation is an alternative approach to listwise deletion. A univariate approach to imputation is to simply replace missing values with the corresponding modality's mean or median score observed in the training data. This approach is referred to as univariate because imputation is only dependent on one modality and is unaffected by other modalities. For example, consider a scenario presented in Table I. Median imputation, for example, would replace the face modality's missing score with 0.41 (the median of the available face modality scores from the face scores in the training data) and the missing fingerprint score would be replaced with 0.74.

Another imputation approach leverages potential relationships between modality scores to better estimate the missing value. One such multivariate approach is *Multivariate Imputation by Chained Equations* (MICE), where missing values are temporarily filled with a placeholder value and then iteratively updated using a trained machine learning model [15].

TABLE I: A simple example of a score dataset with missing values, denoted as ?.

Subject	Face	Fingerprint	Iris
Subject 1	?	0.74	1.00
Subject 2	0.41	0.89	0.47
Subject 3	0.27	?	0.03
Subject 4	0.85	0.00	0.31

In the given example, shown in Table I, both the face and iris missing values are initialized with each modality's mean or median. The scores of individual modalities are sequentially and iteratively updated with a specified machine learning classifier. Once the classifier has been trained, the missing values are updated from the initial placeholder value to the value predicted by the trained classifier, and then the next modality's scores are fixed and the classifier is trained again to update the placeholder values. This process is repeated for a specified number of iterations, or until the imputed values stop changing between iterations.

Previous studies have shown that using imputation methods can reliably improve recognition performance in multibiometric systems when scores are missing [9], [16]. Alternatively, fusion approaches that can adapt to missing data points, such as a likelihood ratio scheme that incorporates both rank and score, can also be used [17], [18].

III. Approach

The experiments in this paper are conducted on 3 multimodal biometric datasets (described in detail below). It should be noted that only match scores are available in these datasets, and no additional information about the samples or the sample quality is known.

For the MIT LL TRIMODAL dataset, scores are distributed in development (dev) and test sets. In this dataset, the training phase is conducted on the dev set, which comprises less than 10% of the total MIT LL TRIMODAL dataset. The remaining two datasets are randomly divided into train (80%) and test (20%) sets. All dataset partitions are subject disjoint, ensuring that the data used for training and testing do not overlap, and all imputation approaches applied to the test set are exclusively derived from calculations performed on the training set.

We next simulate up to 90% incomplete score vectors on 3 versions of each dataset: randomly missing across all score vectors, randomly missing from genuine score vectors, and randomly missing from imposter score vectors. To ensure the robustness of our findings, we repeat this simulation process five times on the complete training and testing partitions of the datasets.

We finally compare the performance of applying the simple sum fusion on the non-imputed version to the imputed versions, noting the mean Pearson's pairwise and Spearman's Rank correlation coefficients between scores of all possible modality pairs in the dataset. This experimental setup is summarized in Table II.

TABLE II: Summary of settings used in the experiments.

Experimental Parameter	Settings
Multibiometric Datasets	NIST BSSR1 [19] BIOCOP2008 MIT LL TRIMODAL [20]
Training, Testing Split	80%, 20% (NIST BSSR1, BIOCOP2008) 7%, 93% (MIT LL TRIMODAL)
# of Missing Score Simulations	5
% Missing	[0, 10, 20, 30, 40, 50, 60, 70, 80, 90]
Univariate Imputations	Mean Median
Multivariate Imputations	Bayesian Regression [21] Decision Tree [22] K Nearest Neighbors [23]
Fusion Applied	Simple Sum Fusion
Modality Relationship Metrics	Pearson Pairwise Correlation Spearman Rank Correlation

A. Datasets

This paper uses 3 multimodal datasets: NIST BSSR1 [19][2], BIOCOP2008 , and MIT LL TRIMODAL [20]. In all these datasets, only the similarity scores are available, and only complete score vectors are analyzed. A brief description of these datasets is outlined below, and a summary is provided in Table III.

The BIOCOP2008 multimodal score dataset is produced by a multichannel CNN that performs cross-modal matching of face images with iris images. The CNN is trained on data collected from the BioCop 2008 ocular dataset, consisting of RGB face and NIR ocular images. From the RGB face images, the left and right eyes are cropped and aligned with the corresponding NIR images. Three copies of the NIR ocular image are stacked on top of the RGB image, creating a 6 channel patch which is then fed into a multichannel CNN to produce similarity scores. In addition to the ocular images for left and right eyes, the irises for each are cropped out to obtain a new type of score belonging to the iris (diagrammed in Figure 1). This process produces 4 types of scores that form the score vector between 2 subjects: left ocular, right ocular, left iris, and right iris. In this dataset, approximately 30% of the score vectors contain naturally occurring missing score values.

The NIST BSSR1 multimodal dataset comprises similarity scores obtained from four modalities: left thumbprint, right thumbprint, face matcher C, and face matcher G. These scores are derived by comparing a subject's sample against the samples of identities in a gallery consisting of 517 subjects. Note, the provided face matcher scores come from matching algorithms NIST generically refers to as C and G.

The MIT LL TRIMODAL multimodal dataset is created by collecting scores from high quality face and speech modalities present in the VoxCeleb-H dataset (referred to as the hard-set) [24]. In addition to the face and speech modalities, a text modality is pulled from a subset of the PAN Celebrity Profiling Twitter dataset [25]. Note, as opposed to the previous datasets where the amount of imposter score vectors vastly outnumber the genuine score vectors, the majority of the score vectors in the MIT LL TRIMODAL dataset are genuine.

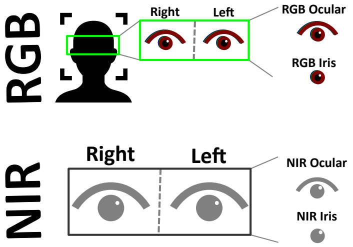


Fig. 1: Diagram of iris and ocular images cropped from ocular region in the BIOCOP2008 dataset images (RGB facial images top, NIR ocular images bottom).

TABLE III: Summary of Multibiometric datasets analyzed.

	Modalities	Score Vectors	Genuine
BIOCOP2008	4	139,230	435 (0.31%)
NIST BSSR1	4	133,903	517 (0.39%)
MIT LL TRIMODAL	3	107,471	77,789 (72.38%)

B. Missing Score Simulations

We simulate missing score data from 0% missing up to 90% missing. To simulate these missing scores, we first randomly select the score vectors to be corrupted. From each of these selected vectors, a random number between 1 and the number of modalities-1 of the vector's scores is dropped. This ensures that at least one score will be dropped and at most all-but-one score will be dropped. The pseudocode for this process is summarized in Algorithm 1. This process is repeated 5 times on the train and test sets. For the BIOCOP2008 dataset, which contains naturally occurring missing score values, we first drop the incomplete score vectors before simulation. Note, we provide the performance of the simulated missing values to the naturally occurring missing values in the Results section.

Algorithm 1 Simulation of Missing Data

```
for proportion $= 0, 10, 20, \ldots, 90$ do
    $n = \text{Integer}\left(\frac{\text{proportion}}{100} \times \text{length(score data)}\right)$
    $corrupted = \text{random.sample}(n, \text{score data})$
    for vector $\in$ corrupted do
        $\text{amount2drop} = \text{random}(1, \text{length(modalities)} - 1)$
        $dropped = \text{random.sample}(\text{amount2drop}, \text{vector})$
        $\text{score data}[dropped] = \text{NaN}$
    end for
end for
```

We apply this process to multiple versions of each dataset: Genuine Missing, Imposter Missing, and Any Missing. For the *Genuine Missing*, only the vectors belonging to the genuine label are altered to contain missing values. Likewise, the *Imposter Missing* indicates only imposter vectors have been altered. *Any Missing* contains randomly simulated missing values regardless of the label.

[2] NIST BSSR1 dataset is available at https://www.nist.gov/itl/iad/image-group/nist-biometric-scores-set-bssr1

TABLE IV: BIOCOP2008 Comparison of original and balanced dataset versions

	Original	Balanced
Score Vectors	139,230	870
Genuine Score Vectors	435	435
Imposter Score Vectors	138,795	435

TABLE V: NIST BSSR1 Comparison of original and balanced dataset versions

	Original	Balanced
Score Vectors	133,903	1,034
Genuine Score Vectors	517	517
Imposter Score Vectors	133,386	517

Because the genuine and imposter classes are often unbalanced, we also consider a balanced training set for the above versions. For example, in the NIST BSSR1 dataset each subject id contains a score vector for every subject in the gallery. That is, for every 1 genuine score vector, there are 516 imposter score vectors, leading to the potential for over-fitting. We consider a balanced version of the training dataset, where we randomly down sample from the larger class to be the same size as the smaller class. Note that the results presented in Section IV have been generated on the same test set. Tables IV, V, and VI show the differences between the original dataset and the balanced dataset versions.

C. Imputation

For our analysis, we consider the verification performance of simple sum fusion on the simulated missing dataset as the baseline performance. We then apply univariate mean and median imputations, as well as multivariate imputation through MICE with Bayesian, Decision Tree, and KNN regressor models. We additionally examine how reducing training data impacts imputation outcomes. Note that none of the imputation techniques use vector labels (i.e., genuine or imposter), but rather are only calculated on scores within the modality (univariate) or scores across modalities (multivariate). These models are defined below.

Univariate imputation approaches utilize only the scores of the missing score's modality from the training set. **Mean** imputation replaces missing scores within a modality with the mean score of the available scores for that modality in the training set. Similarly, **Median** imputation replaces missing scores within a modality with the median score of the available scores for that modality in the training set.

TABLE VI: MIT LL TRIMODAL Comparison of original and balanced dataset versions

	Original	Balanced
Score Vectors	107,471	59,364
Genuine Score Vectors	77,789	29,682
Imposter Score Vectors	29,682	29,682

In addition to the above univariate approaches, we apply multivariate imputation methods using the MICE method (described in Section II) with the following supervised models. It should be noted that emphasis was not placed on the parameter tuning in these models, and it is possible performance could be further improved with parameter optimization strategies.

MICE with **Bayesian Ridge Regression** is a probabilistic model of regression $p(y \mid X, w, \alpha) = \mathcal{N}(y \mid Xw, \alpha)$, where parameters are estimated by maximizing the log marginal likelihood.

MICE with **Decision Tree Regression** is a non-parametric model of regression. Decision Trees aim to learn a hierarchy of decision rules inferred from the training data's features. This approach can potentially be more resilient to violations in underlying model assumptions; however it can also be prone to over-fitting.

MICE with **KNN Regression** imputation applies a K Nearest Neighbor (KNN) approach. Score vectors in the training data are sorted by distance, and the scores of the k-nearest neighbors (i.e. those with the smallest distances) are averaged. The experiments presented here set k to 5 neighbors and use the Euclidean distance to measure the distance of 2 points in the m dimensional space.

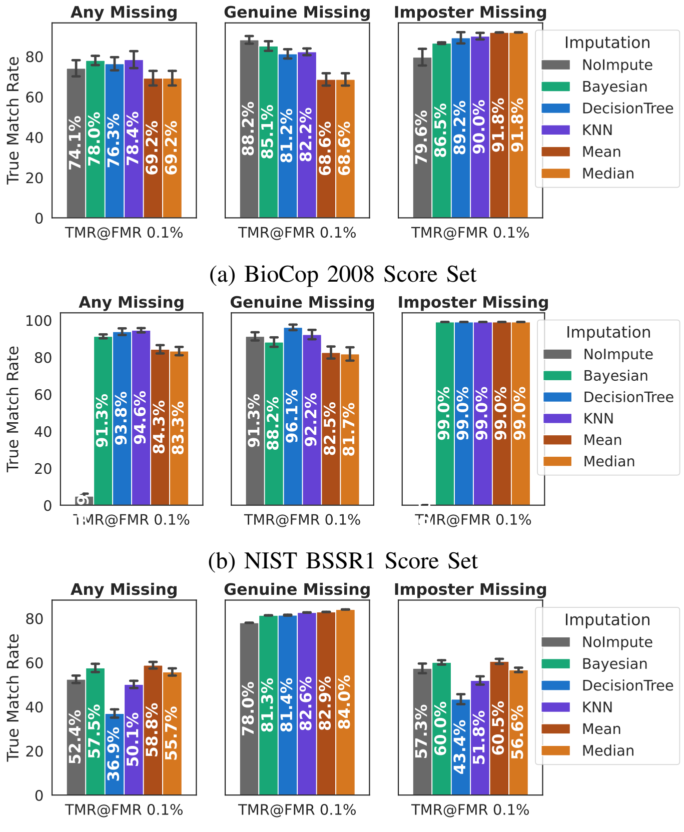
IV. Results

In this section we highlight a small subset of the results from the experimental approach described above. Here we highlight the verification performance at False Match Rate (FMR) = 0.1% on the 50% missing levels in Figures 2 and 3 [3].

In both the unbalanced and balanced versions, we see that fusion performance for the *Any Missing* version the MICE with Bayesian Ridge Regression consistently improves performance over applying no imputation at all (even if it is not *the best* imputation technique for each dataset). However we note that the verification performance of the test set varies between the approach trained on balanced data and the approach trained with the unbalanced data. In the unbalanced training approaches, the imputed genuine scores do not show such a clear gain for the BIOCOP2008 and NIST BSSR1 datasets, where there are substantially fewer genuine scores to train imputation methods on. Conversely the MIT LL TRIMODAL dataset contains a larger proportion of genuine scores in which to train imputation methods (Figure 2). This observation highlights the biases introduced by the overrepresented class in the training set. However, we also emphasize that despite the exclusion of a significant number of valid score vectors to achieve class balance, the overall verification performances are not severely compromised. While more training samples are generally preferred in machine learning, our observations remain consistent regardless of training data size.

Additionally, our analysis indicates that multivariate approaches outperform univariate approaches when imputing missing genuine scores, while mean or median imputation

[3] Complete results may be viewed at https://melissadale.github.io/WIFS2023/



(a) BioCop 2008 Score Set

(b) NIST BSSR1 Score Set

(c) MIT LL TRIMODAL Score Set

Fig. 2: Estimated TMR at FMR=0.1% for 50% incomplete score vectors found in the unbalanced version of the training data.

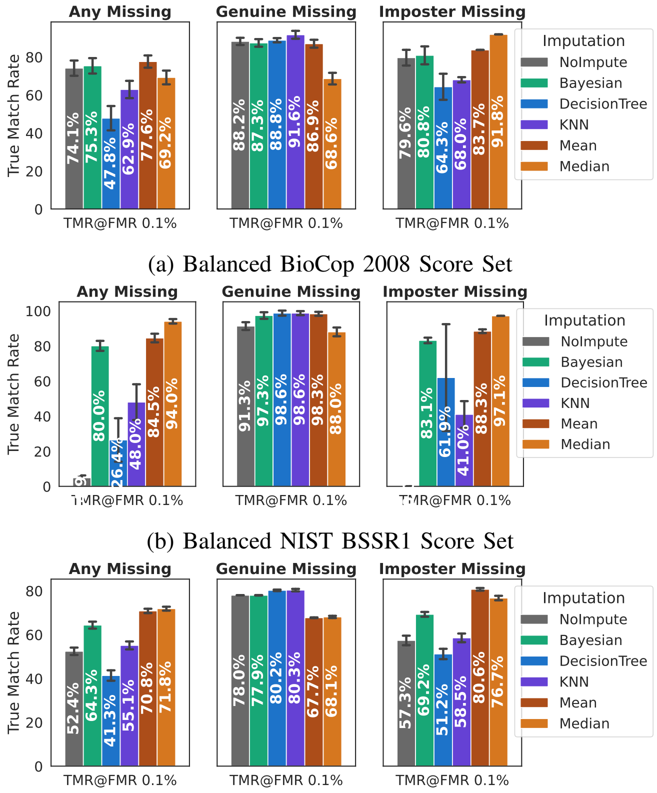
achieves better results for imposter scores. We note that the average pairwise correlation between modalities is higher for genuine scores than for imposter scores. However, the overall mean correlation remains the highest, suggesting that correlation alone does not account for differences in the observed performances. These findings highlight the need for further research to develop an imputation approach that can deliver robust performance during test time, where the label classification is unknown in advance.

A. *Missing Simulation versus Naturally Occurring Missing*

In our experiments, we simulated random missing scores from complete versions of each dataset to meet the Missing at Random (MAR) requirements. However, it is important to highlight that the verification performance of naturally occurring missing scores within the BIOCOP2008 dataset is comparable to the performance reported in the simulated missing data, as presented in Table VII. This demonstrates the robustness and relevance of our findings, indicating that our results hold promise for real-world scenarios.

V. Conclusion and Discussion

In this study, we investigated the influence of imputation techniques on verification performance in multibiometric score



(a) Balanced BioCop 2008 Score Set

(b) Balanced NIST BSSR1 Score Set

(c) Balanced MIT LL TRIMODAL Score Set

Fig. 3: Estimated TMR at FMR=0.1% for 50% incomplete score vectors trained on the balanced training set.

TABLE VII: Comparison of performance observed in naturally occurring missing value and simulated missing values observed in the BIOCOP2008 dataset.

	Amount Missing	TMR@FMR=0.1%
Naturally Missing	30.39%	81.68
Simulated Missing	30%	80.20 (+/- 5.0)

datasets. Our findings have important implications for improving the accuracy and reliability of multibiometric systems with missing scores. We summarize key observations and the resulting recommendations below.

Firstly, our results consistently demonstrate the benefits of imputation over not imputing missing scores, regardless of the type of scores being imputed. By incorporating imputation into the data preprocessing stage, multibiometric systems can benefit from more complete score data, thereby improving overall system performance. **Recommendation 1:** *Enhance multibiometric system design by integrating imputation techniques.* Invest time in finding the most appropriate approach for your data.

Secondly, we observed that imputation methods tend to favor the overrepresented class, introducing biases in the imputed scores. To mitigate this issue, we emphasize the importance of balancing the classes within the training dataset. Despite the potential need to drop a substantial number of data points from the overrepresented class, our analysis demonstrates that this step does not severely harm the overall verification performance. Balancing the training data helps alleviate biases and ensures fair representation of both genuine and imposter scores, leading to more reliable and unbiased performance evaluation. **Recommendation 2:** *Balance the imposter and genuine score vectors in the training set.* Although balancing the training data may involve disregarding a significant number of score vectors from the overrepresented class, it is important to note that the performance of the balanced versions in the test set is not significantly compromised for the datasets analyzed.

Furthermore, our study highlights the effectiveness of different imputation approaches based on the classification of the missing scores. Specifically, multivariate imputation approaches excel in estimating missing genuine scores, while mean or median imputation methods outperform multivariate approaches for imputed imposter scores. **Recommendation 3:** *When designing an imputation approach, consider the nature of missing scores.* Understanding which types of scores are more prone to being missing can inform the development of targeted and effective imputation strategies. By tailoring the imputation process to the specific characteristics of the missing scores, we can enhance the accuracy and reliability of the overall biometric system. This observation suggests future research efforts are required to develop a novel approach to imputation in multibiometrics. By creating methods to manage missing scores without prior label knowledge, we can boost recognition accuracy and strengthen practical biometric applications.

In conclusion, our study provides insights into the role of imputation techniques in multibiometric score datasets. By leveraging imputation, multibiometric systems can enhance their recognition accuracy and reliability. Balancing the training data and employing appropriate imputation methods based on score type are essential considerations for achieving optimal performance. Our findings contribute to the understanding of imputation in the context of multibiometric systems and pave the way for future research in this area.

VI. Future Works

While this study has shed light on the role of imputation techniques in multibiometric score datasets, there are several avenues for future research in this area. One important direction for future work is the development of innovative imputation methods that can effectively handle missing scores without relying on prior knowledge of the label. This includes investigating hybrid approaches that combine multiple imputation techniques or incorporate other data preprocessing methods. Hybrid methods have the potential to leverage the strengths of different imputation techniques and improve overall system performance. Additionally, although balancing the training data was found to mitigate biases introduced by imputation methods, further investigation is needed to explore more advanced techniques for addressing class imbalance. Future research can explore methods such as oversampling, undersampling, or generating synthetic data to ensure fair representation of both genuine and imposter scores in the training dataset.

References

[1] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. Springer Science & Business Media, 2006.

[2] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, 1995.

[3] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, no. 4, pp. 861–874, 2002.

[4] K.-A. Toh, X. Jiang, and W.-Y. Yau, "Exploiting global and local decisions for multimodal biometrics verification," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 3059–3072, 2004.

[5] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern recognition letters*, vol. 24, no. 13, pp. 2115–2125, 2003.

[6] N. V. Boulgouris, K. N. Plataniotis, and E. Micheli-Tzanakou, *Biometrics: Theory, Methods, and Applications*. John Wiley & Sons, 2009.

[7] A. Mishra, "Multimodal biometrics it is: Need for future systems," *International Journal of Computer Applications*, vol. 3, no. 4, pp. 28–33, 2010.

[8] L. Allano, B. Dorizzi, and S. Garcia-Salicetti, "Tuning cost and performance in multi-biometric systems: A novel and consistent view of fusion strategies based on the sequential probability ratio test (sprt)," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 884–890, 2010.

[9] M. R. Dale, A. Jain, and A. Ross, "On missing scores in evolving multibiometric systems," in *26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 982–988.

[10] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures," *Pattern recognition*, vol. 38, no. 5, pp. 777–779, 2005.

[11] Y. Ma, B. Cukic, and H. Singh, "A classification approach to multibiometric score fusion," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2005, pp. 484–493.

[12] K. Nandakumar, Y. Chen, S. C. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 342–347, 2007.

[13] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[14] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[15] S. Van Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 1, pp. 1–67, 2011.

[16] Y. Ding and A. Ross, "A comparison of imputation methods for handling missing scores in biometric fusion," *Pattern Recognition*, vol. 45, no. 3, pp. 919–933, 2012.

[17] K. Nandakumar, A. K. Jain, and A. Ross, "Fusion in multibiometric identification systems: What about the missing data?" in *International Conference on Biometrics*. Springer, 2009, pp. 743–752.

[18] O. Fatukasi, J. Kittler, and N. Poh, "Estimation of missing values in multimodal biometric fusion," in *IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, 2008.

[19] "NIST biometric scores set," https://www.nist.gov/itl/iad/ig/biometricscores, 2004.

[20] E. Singer, B. J. Börgstrom, K. Alperin, T. Nguyen, C. Dagli, M. R. Dale, and A. Ross, "On the design of the MIT LL trimodal dataset for identity verification," in *11th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2023.

[21] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[23] O. Kramer, "K-nearest neighbors," in *Dimensionality reduction with unsupervised nearest neighbors*. Springer, 2013, pp. 13–23.

[24] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[25] "Pan celebrity profiling 2019," https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html.