**DOI 10.54596/2958-0048-2024-4-195-203**
**UDK 004.421.4**
**IRSTI 67.01.77**

# MISSING VALUES IMPUTATION TOOL USING IMPUTEX ALGORITHM

**Fatimah Sidi[1*], Lili Nurliyana Abdullah[1], Mustafa Alabadla[1], Iskandar Ishak[1]**
*[1*]Universiti Putra Malaysia, Serdang, Selangor D. E., Malaysia*
*[*]Corresponding author: fatimah@upm.edu.my*

**Abstract**

Missing data is a prevalent issue affecting data quality across numerous fields. One frequent challenge arises when data is lost during the input stage. Numerous studies have proposed methods to impute missing values for data across multiple fields. However, certain domains present unique challenges due to the involvement of attributes from multiple scientific disciplines, such as biology, chemistry, and medical which complicates the imputation process. The purpose of this study is to design an application that addresses missing values and maintains accuracy in large datasets, with a focus on minimizing processing time. The application's performance is evaluated based on classification accuracy using various imputation methods. The proposed application outperforms performance compared to current software tools such as against R package, Statistical Package for the Social Sciences (SPSS), Stata, and Microsoft Excel. This study helps to improve data quality and contributes to data science by improving the data cleaning procedure, which is a step in the data pre-processing stage.

**Keywords**: Missing Values, Imputation, Web Application, Data Quality.

# IMPUTEX АЛГОРИТМІН ПАЙДАЛАНУ АРҚЫЛЫ ЖЕТІСПЕЙТІН МӘЛІМЕТТЕРДІ ТОЛТЫРУ ҚҰРАЛЫ

**Фатимах Сиди[1*], Лили Нурлияна Абдулла[1], Мустафа Алабада[1], Искандар Ишак[1]**
*[1*]Путра университеті, Серданг, Селангор, Малайзия*
*[*]Хат-хабар үшін автор: fatimah@upm.edu.my*

**Андатпа**

Мәліметтердің жетіспеуі көптеген салаларда деректердің сапасына кері әсерін тигізетін кең таралған мәселе. Жиі кездесетін себептердің бірі – ақпараттың енгізу кезінде жоғалуы. Түрлі зерттеулер жетіспейтін мәліметтерді толтыру әдістерін ұсынады, бірақ биология, химия және медицина сияқты салаларда көпсалалы сипаттамаларға байланысты қосымша қиындықтар туындайды. Бұл зерттеудің мақсаты – үлкен мәліметтер жиынтығында жетіспейтін деректерді тиімді түрде толтыратын қосымшаны жасап шығару, өңдеу уақытын азайтуға бағытталған. Қосымшаның тиімділігі әртүрлі толтыру әдістерінің классификациялық дәлдігіне негізделе отырып бағаланды. Ұсынылған қосымша R, SPSS, Stata және Microsoft Excel сияқты құралдармен салыстырғанда артықшылықтарды көрсетіп, деректер сапасын және оларды тазалау процесін жақсартады.

**Кілт сөздер:** жетіспейтін мәліметтер, толтыру, веб-қосымша, деректер сапасы.

# ИНСТРУМЕНТ ДЛЯ ВОССТАНОВЛЕНИЯ ОТСУТСТВУЮЩИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА IMPUTEX

**Фатимах Сиди[1*], Лили Нурлияна Абдулла[1], Мустафа Алабада[1], Искандар Ишак[1]**
*[1*]Университет Путра Малайзия, Серданг, Селангор, Малайзия*
*[*]Автор для корреспонденции: fatimah@upm.edu.my*

**Аннотация**

Отсутствие данных представляет распространенную проблему, негативно влияющую на качество данных во многих областях. Одной из частых причин является утрата информации на этапе ввода. Различные исследования предлагают методы восстановления отсутствующих данных, однако в некоторых

сферах, таких как биология, химия и медицина, возникают дополнительные сложности из-за многодисциплинарности атрибутов. Целью данного исследования является разработка приложения, способного эффективно восстанавливать отсутствующие данные в крупных наборах данных с минимизацией времени обработки. Производительность приложения оценивалась на основе классификационной точности различных методов восстановления. Предложенное приложение продемонстрировало превосходство по сравнению с существующими инструментами, такими как R, SPSS, Stata и Microsoft Excel, улучшая качество данных и процесс их очистки.

**Ключевые слова:** отсутствие данных, восстановление, веб-приложение, качество данных.

## I. INTRODUCTION

The absence of data represents a serious issue undermining the integrity and quality of information, which ultimately negatively impacts analytical outcomes. This deficiency can significantly reduce the accuracy of analysis and increase bias caused by the disparity between available and missing values [1]. The primary reasons for such phenomena include respondents' refusal to provide information, typographical errors, or equipment failures [2][3][4]. These gaps inevitably arise during the data collection stage and must be addressed before initiating preprocessing. Working with a complete dataset is critically important as data quality directly affects decision-making in organizations [5]. For example, low-quality data can result in inaccurate analysis outcomes, leading to poor decision-making. Thus, maintaining data quality is a crucial factor that greatly affects decision-making in an organization. In other words, the quality of the decision-making process relies on the accuracy of the data and the varying abilities to interpret that data. For instance, low-quality data can lead to inaccurate data analysis which in return results in providing wrong decisions. Therefore, these wrong decisions may result in catastrophic effects on society and produce undesirable outcomes.

The purpose of this study is to develop the ImputeX application, which efficiently imputes missing data in large datasets, minimizes processing time, and improves classification accuracy. This application is designed to address the limitations of existing tools, such as limited performance, complex configurations, and reliance on programming skills.

Research Objectives:

1. Develop the ImputeX algorithm using an ensemble machine learning approach to restore missing values.
2. Conduct a comparative performance analysis of ImputeX with existing tools, including R, SPSS, Stata, and Excel.
3. Apply the developed application to solve real-world problems in medicine and bioinformatics.

This study contributes to the field of data imputation by presenting a user-friendly web application, ImputeX, which efficiently handles missing data with high accuracy and reduced processing time. The key contributions of this study include the comparative evaluation of ImputeX with widely-used tools such as R, SPSS, Stata, and Microsoft Excel, and the demonstration of its superior performance across various missing data scenarios. The remainder of this paper is organized as follows: Section II provides a review of related works, Section III details the methodology, Section IV presents the results and discussion, and Section V concludes the study with future directions

This paper is organized as follow: Section II presents the related works that emphasized on the cause and effect of missing value problems, as well as the types of mechanism and methods to address them. This is then followed by Section III which discusses the research methodology that is employed by this study. The result and discussion section is presented in Section IV and the conclusion is in Section V.

## II. RELATED WORKS

This section reviews the causes and effects of the missing values issue, highlighting its effect on decision-making outcomes. It outlines the tools available for imputing missing data and addresses the limitations of current methods. Missing values are a common issue that can greatly impact the accuracy and reliability of data analysis. [3] examined the factors contributing to missing values during the execution of an experiment. These missing values may occur due to mistakes in manual data entry, equipment or mechanical failures, and errors in data transmission. These issues can arise randomly and are often difficult to manage. The main causes of missing values can be divided into three key categories: data collection practices, data entry errors, and technical issues.

Data collection practices frequently result in missing values, which happen when specific data points are either not measured or not recorded. For example, healthcare providers may fail to record certain information due to time constraints, oversight, or a lack of standardized protocols [1]. Moreover, patient non-compliance, such as missing follow-up appointments or declining to provide specific information, can lead to incomplete records [6]. Data entry errors are another common cause of missing values with factors such as typographical mistakes, incorrect coding, or incomplete data entry contribute to the occurrence of missing values [7]. In terms of technical issues related to data management systems can also lead to missing values. These issues encompass system crashes, software bugs, and data transmission failures, all of which can result in data loss or corruption. Furthermore, discrepancies in data integration from multiple sources, such as merging data from different departments or external databases, can cause missing values [8].

There are numerous techniques for imputing missing data, but identifying the most effective approach generally requires testing and comparing several methods before deciding. Many public software tools have been developed to impute missing values through a graphical user interface (GUI) A software tool called WIMP, developed using .NET technology by [9], enabling users to create accounts, log into the system, and apply various imputation techniques to address missing values. MIDA, a web-based imputation tool introduced by [10], is designed to address the missing value problem specifically for data missing at random (MAR). MIDA offers a user-friendly interface accessible to the public, requiring no software installation and accommodating users of all programming skill levels. [11] intruduced ImputeEHR, a Python-based imputation tool that enables the use of multiple machine learning techniques for data imputation. Another imputation tool available to the public is BIMAM, developed by [12]. It is designed to impute both continuous and binary types of missing data. However, the aforementioned software tools have their limitations. For instance, WIMP's inability to export results instantly can leave users waiting for a long time without any indication of the imputation process's outcome. Regarding MIDA, the user can choose from five machine learning approaches based on the type of variable that needs to be imputed. It also requires the user's email to send back the results of the data imputation. Though ImputeEHR focused more on reducing the execution time, it has outperformed in certain cases and is not viewed as an optimal solution for missing values. In the case of BIMAM, users must specify the clustering variable, output, covariate, initial iteration, and the number of updates. Additionally, there are numerous tools and software available to the public that are commonly utilized by researchers and practitioners to tackle the issue of missing values, each providing distinct features and functionalities. Some of the most commonly used tools for missing data imputation include R Package, IBM SPSS (Statistical Package for the Social Sciences), Stata, and Microsoft Excel, when paired with the XLSTAT add-on.

## III. METHODOLOGY

The imputation tool has been developed as a web application known as the ImputeX Application, driven by the ImputeX algorithm [13], utilizes modern web technologies to provide a robust platform that enables users to impute missing values without requiring any programming or machine learning expertise. The ImputeX Application has been developed using the latest React technology for the frontend and the Flask web framework for the backend [14]. The proposed application is a public website that allows users to impute missing values using the ImputeX algorithm without requiring any registration process.

The ImputeX algorithm is based on decision trees and ensemble techniques. The main stages of the algorithm are as follows:

1. Data preparation involves identifying missing values and determining their types, whether numerical or categorical. The dataset is then split into training and testing subsets.
2. Selection of the imputation method is carried out. For numerical data, regression based on an ensemble of decision trees is applied. For categorical data, classification is performed using probabilistic models.
3. Prediction is executed using the "Extremely Randomized Trees" algorithm, which predicts the missing values.
4. Post-processing includes validation of the imputed data using cross-validation techniques and an evaluation of the imputation accuracy.

To further clarify the workflow of the ImputeX web application, Figure 1 presents a UML Sequence Diagram illustrating the interaction between the key components of the platform:

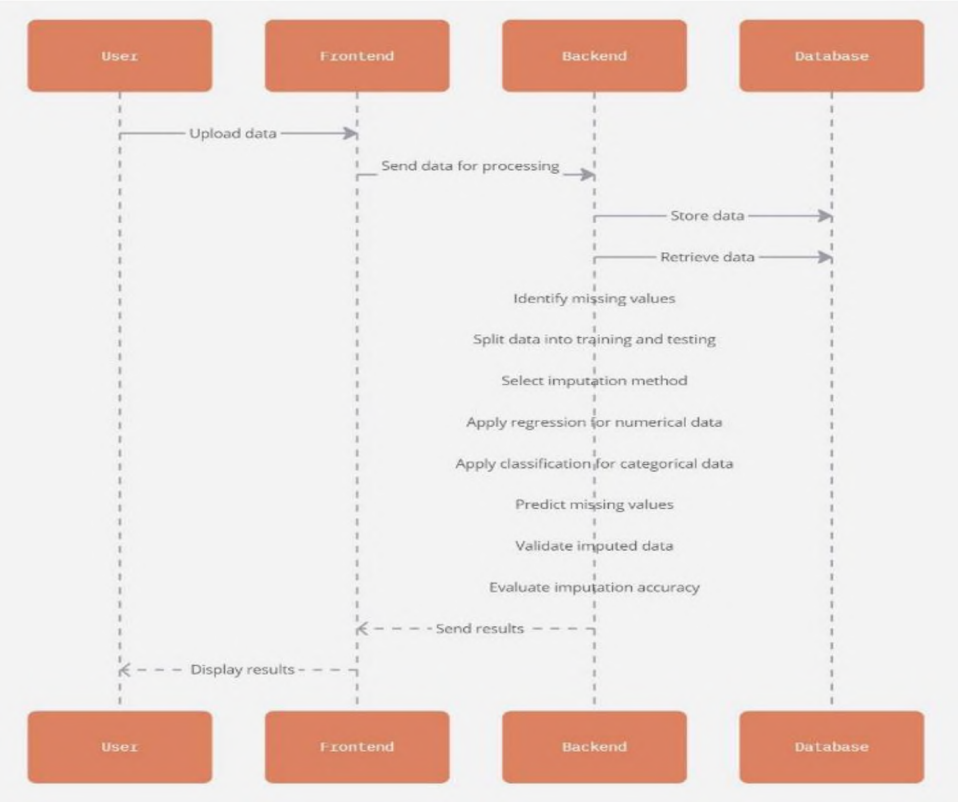


Figure 1. UML Sequence Diagram of ImputeX Platform Workflow

The diagram outlines the interactions between the user, frontend, backend, and the database, showing the steps involved in uploading data, initiating the imputation process, and retrieving the results.

The application is designed with modern web technologies to ensure ease of use. The frontend is built with ReactJS for user interaction, while the backend relies on Flask and Python libraries for data processing. The infrastructure utilizes a cloud-based server to enable fast and efficient operations.

The TADPOLE dataset was used to validate the algorithm. This dataset contains 13,915 records with 99 attributes, including both numerical and categorical variables. Missing data ratios ranged from 10% to 90%, providing a robust test environment for the algorithm.

To evaluate performance, two key metrics were used: accuracy, which measures the proportion of correctly imputed values, and root mean squared error (RMSE), which quantifies the deviation of predicted values from true values. Together, these metrics provide a comprehensive assessment of the algorithm's effectiveness.

## IV. RESULTS AND DISCUSSION

Here is an example of processing a dataset with an illustration of the input data, algorithm steps and results. To test the performance of the algorithm, 10 runs were conducted on datasets with missing values in the range of 10-90%. The following metrics were used for evaluation: classification accuracy, RMSE error and execution time.

Table 1. Presents the performance of ImputeX compared to other tools:

| Missing Ratio | ImputeX | R | SPSS | Stata | Excel |
|---|---|---|---|---|---|
| 10% | 98.4% | 98.2% | 95.8% | 97.2% | 96.2% |
| 50% | 90.1% | 88.6% | 76.2% | 72.4% | 83.0% |
| 90% | 78.2% | 73.5% | — | 72.2% | 76.2% |

The results demonstrate that ImputeX achieves higher classification accuracy, especially with high missing value ratios (70% and above), outperforming R, SPSS, Stata, and Excel. This can be attributed to its adaptive and ensemble-based approach.

Table 2. Illustrates an example of how ImputeX imputes missing data:

| ID | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|
| 1 | 45 | 78 | ? |
| 2 | 50 | ? | 20 |
| 3 | ? | 85 | 25 |

Table 3. Imputed data:

| ID | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|
| 1 | 45 | 78 | 22 |
| 2 | 50 | 80 | 20 |
| 3 | 47 | 85 | 25 |

This example showcases the algorithm’s ability to restore missing values based on correlations between variables.

Table 4 below provides a comparative analysis of existing data imputation tools, highlighting their key features and the advantages offered by ImputeX:

Table 4. Comparative analysis of existing data:

| Tool | Imputation Method | Accuracy (%) | Processing Time | User Interface |
|---|---|---|---|---|
| R | Multiple Imputation | 88.6 | Moderate | Command-Line |
| SPSS | Multiple Imputation | 76.2 | Slow | GUI |
| Stata | Statistical Models | 72.4 | Slow | GUI |
| MS Excel | Basic Statistical Tools | 83.0 | Fast | GUI |
| ImputeX | Ensemble ML Techniques | **90.1** | **Fast** | **Web-Based** |

This table clearly demonstrates that ImputeX excels in both accuracy and processing efficiency.

The main objective of the proposed application is to assist users by allowing them to focus on conducting the experiments instead of devoting their resources on searching, analysing, interpreting, and implementing different imputation methods. Furthermore, to accelerate the imputation process, the ImputeX algorithm is implemented in the backend of the proposed application, significantly reducing the execution time for imputation. Whenever a user uploads a dataset and starts a new imputation through the application environment, it will send the dataset to the backend to execute the imputation process separately from the frontend. Once the imputation process is complete, the backend will send the imputed dataset back to the frontend as a result. Then, the user will be able to download the complete dataset from the user interface of the application. The application can be accessed by any web browser and the way it manages the imputation request is shown in Fig 2. The application is accessible at the following link as shown in Fig. 3: https://autoimputex.upm.edu.my.

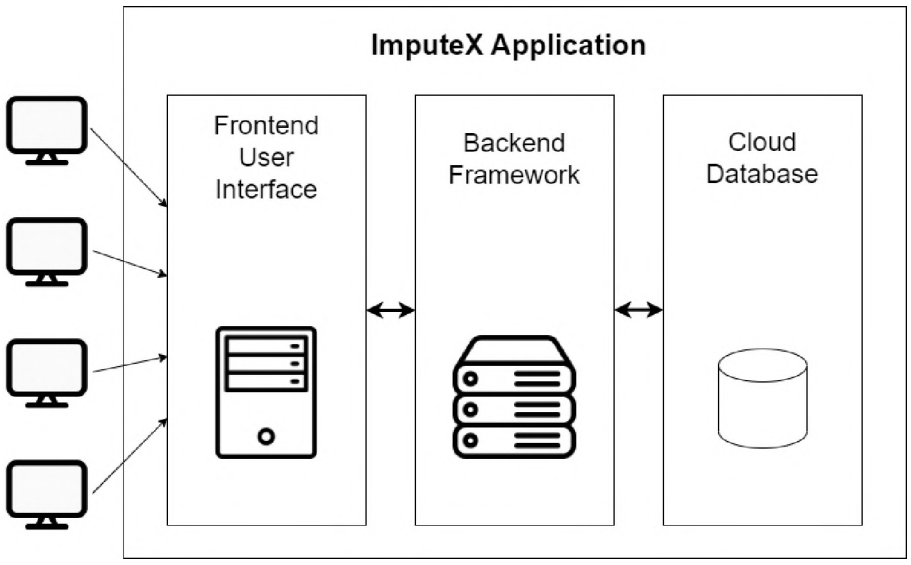


Fig. 2. System Architecture of the ImputeX Application

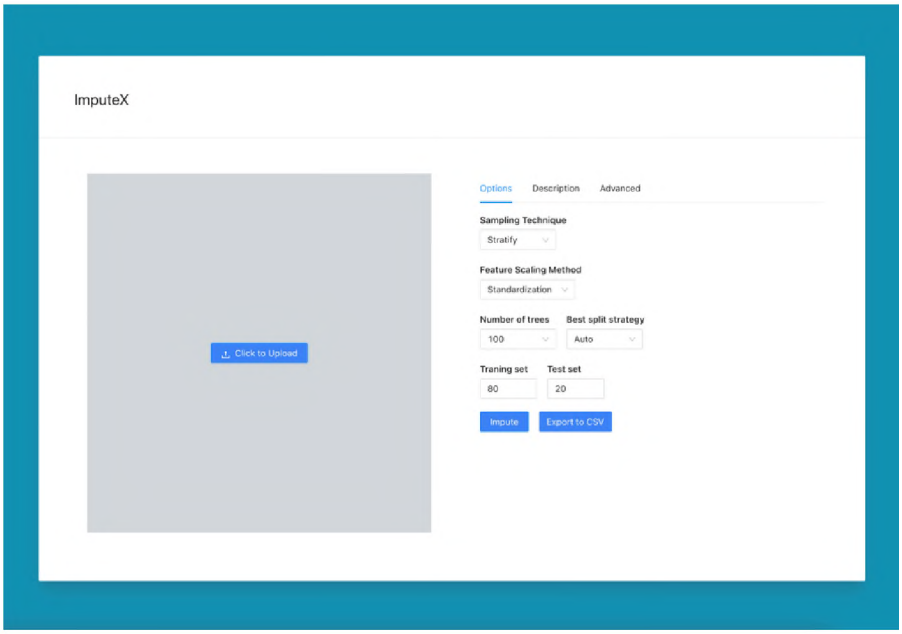


Fig. 3. Screenshot of the advanced tab in the ImputeX Application Homepage

The efficiency of ImputeX Application is evaluated against R, SPSS, Stata and MS Excel using multiple imputation in each software tool by imputing missing values. The main dataset used in this experiment is the TADPOLE (The Alzheimer's Disease Prediction of Longitudinal Evolution) dataset collected from the University of Southern California. The original dataset contains 13,915 instances and 99 attributes [15]. Nonetheless, a sample of 15 variables was selected from the TADPOLE dataset. The imputation tools are applied to the TADPOLE dataset multiple times, each time with a different missing ratio ranging from 10% to 90% for a total of 10 runs in each scenario. Table 1 shows the average accuracy for the ImputeX Application against existing imputation tools under different missing ratios for the TADPOLE dataset.

Table 5. Average accuracy for the ImputeX Application against existing imputation tools under different missing ratios

| Missing Ratio | ImputeX | R | SPSS | Stata | MS Excel |
|---|---|---|---|---|---|
| 10% | **0.984** | 0.982 | 0.958 | 0.972 | 0.962 |
| 20% | **0.967** | 0.964 | 0.921 | 0.943 | 0.927 |
| 30% | **0.945** | 0.934 | 0.877 | 0.746 | 0.884 |
| 40% | **0.928** | 0.917 | 0.806 | 0.892 | 0.868 |
| 50% | **0.901** | 0.886 | 0.762 | 0.724 | 0.830 |
| 60% | **0.873** | 0.858 | 0.696 | 0.821 | 0.763 |
| 70% | **0.842** | 0.825 | NA | 0.786 | 0.856 |
| 80% | **0.620** | 0.598 | NA | 0.571 | 0.616 |
| 90% | **0.782** | 0.735 | NA | 0.722 | 0.762 |

As can be seen, the ImputeX Application has the highest classification accuracy amongst all imputation tools. It is also obvious that the accuracy decreases as the number of missing values increases reaching approximately 0.782 at 90% missing percentage. On the other hand,

R programming language used in RStudio as the software tool for data imputation has achieved a good performance compared to other existing imputation tools and falling behind ImputeX with a slight difference at lower missing value percentages. However, the gap in accuracy between ImputeX and R escalates when dealing with high numbers of missing values. Following R in terms of classification accuracy, Stata and Excel were exchanging the roles for who is better and more accurate in estimating categorical missing values. For instance, Stata has exceeded both SPSS and MS Excel at 10%, 20%, 40%, and 60% missing ratio. While MS Excel has achieved better accuracy than Stata and SPSS at 5 out of 9 different missing ratios. Making it slightly better than Stata when dealing with different missing scenarios. It is also noticed that the difference in accuracy between MS Excel and R is more than the difference between ImputeX and R. Furthermore, SPSS software has different methods of imputing missing values and the most commonly used one is the multiple imputation. Nevertheless, according to the average accuracy results for the imputation, it seems that SPSS software is not reliable enough to deal with high missing values proportions especially at 70% and above. The reason for that is because SPSS is unable to impute a complete missing instance as it was observed that some rows were completely missing in the dataset. Thus, the evaluation methods were unable to read missing values and calculate any performance metrics.

## 5. CONCLUSION

The ImputeX application demonstrates significant advantages in imputing missing data, particularly for large datasets with high missing value ratios. The proposed algorithm outperforms existing tools in accuracy and efficiency, while the application's design ensures user-friendliness. Future work will focus on integrating real-time data processing capabilities through cloud technologies to further enhance its utility.

The whole imputation process can be tedious for some researchers and analysts due to the long steps that needs to be done in some of these tools. Also, most of these tools have multiple imputation method as the best imputation method and the most accurate one. The ImputeX Application has addressed all these drawbacks by conducting the imputation without requiring any additional steps or configurations or any codes/commands from the user. This study helps to improve data quality. Additionally, it contributes to data science by improving the data cleaning procedure, which is a step in the data pre-processing stage.

While the ImputeX algorithm itself is not newly developed, this study contributes significantly by integrating it into an accessible and efficient web-based platform. The key contributions include reducing technical barriers for end-users, improving processing efficiency through cloud infrastructure, and offering a robust user interface that simplifies data imputation tasks. This practical implementation bridges the gap between complex machine learning techniques and non-expert users, ensuring broader adoption in real-world scenarios

To improve it even further beyond, the ImputeX Application need to be provided with a real-time listener on a cloud dataset to impute missing values autonomously without any user intervention.

**References:**

1. Phung, S., Kumar, A., & Kim, J. (2019). A deep learning technique for imputing missing healthcare data. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 6513–6516. https://doi.org/10.1109/EMBC.2019.8856760

2. Deb, R., & Liew, A.W.C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences, 339, 274–289. https://doi.org/10.1016/j.ins.2016.01.018

3. Dhindsa, K., Bhandari, M., & Sonnadara, R.R. (2018). What's holding up the big data revolution in healthcare? BMJ (Online), 363, 1–2. https://doi.org/10.1136/bmj.k5357

4. Tsai, C.F., & Chang, F.Y. (2016). Combining instance selection for better missing value imputation. Journal of Systems and Software, 122, 63–71. https://doi.org/10.1016/j.jss.2016.08.093
5. Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. Journal of Business Study, 70, 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007
6. Batra, S., Khurana, R., Khan, M.Z., Boulila, W., Koubaa, A., & Srivastava, P. (2022). A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records. Entropy, 24(4), 1–20. https://doi.org/10.3390/e24040533
7. Chen, Z., Tan, S., Chajewska, U., Rudin, C., & Caruana, R. (2023). Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help? Proceedings of Machine Learning Research, 209, 86–99.
8. Feng, S., Hategeka, C., & Grépin, K.A. (2021). Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. Population Health Metrics, 19(1), 1–28. https://doi.org/10.1186/s12963-021-00274-z
9. Urda, D., Subirats, J.L., García-Laencina, P.J., Franco, L., Sancho-Gómez, J.L., & Jerez, J.M. (2012). WIMP: Web server tool for missing data imputation. Computer Methods and Programs in Biomedicine, 108(3), 1247–1254. https://doi.org/10.1016/j.cmpb.2012.08.006
10. Acampora, G., Vitiello, A., & Siciliano, R. (2020). MIDA: A web tool for missing data imputation based on a boosted and incremental learning algorithm. IEEE International Conference on Fuzzy Systems, 1–6. https://doi.org/10.1109/FUZZ48607.2020.9177644
11. Zhou, Y.H., & Saghapour, E. (2021). ImputEHR: A Visualization Tool of Imputation for the Prediction of Biomedical Data. Frontiers in Genetics, 12(July), 1–9. https://doi.org/10.3389/fgene.2021.691274
12. Elfadaly, F.G., Adamson, A., Patel, J., Potts, L., Potts, J., Blangiardo, M., Thompson, J., & Minelli, C. (2021). BIMAM - A tool for imputing variables missing across datasets using a Bayesian imputation and analysis model. International Journal of Epidemiology, 50(5), 1419–1425. https://doi.org/10.1093/ije/dyab177
13. Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., & Hamdan, H. (2022). ExtraImpute: A Novel Machine Learning Method for Missing Data Imputation. Journal of Advances in Information Technology, 13(5). https://doi.org/10.12720/jait.13.5.470-476
14. Alabadla, M., Sidi, F., Ishak, I., Ibrahim, H., Hamdan, H., Amir, S. I., Nurlankyzy, A.Y. (2023). AutoImpute: An Autonomous Web Tool for Data Imputation Based on Extremely Randomized Trees. In Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA2023), (Italy, Rome), 11-13 July 2023. Volume 1, pp 598-605.
15. Jabason, E., Ahmad, M.O., & Swamy, M.N.S. (2018). Missing Structural and Clinical Features Imputation for Semi-supervised Alzheimer's Disease Classification using Stacked Sparse Autoencoder. 2018 IEEE Biomedical Circuits and Systems Conference, BioCAS 2018 - Proceedings, 1–4. https://doi.org/10.1109/BIOCAS.2018.8584844

**Information about the authors:**

**Fatimah Sidi** – corresponding author, PhD, Associate Professor, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia; e-mail: fatimah@upm.edu.my;
**Lili Nurliyana Abdullah** – PhD, Associate Professor, Department of Mulitimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia; e-mail liyana@upm.edu.my;
**Mustafa Alabadla** – PhD Candidate, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia; e-mail gs59711@student.upm.edu.my;
**Iskandar Ishak** – PhD, Associate Professor, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia; e-mail iskandar_i@upm.edu.my.