Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 4, August 2025, pp. 3091~3100
ISSN: 2302-9285, DOI: 10.11591/eei.v14i4.8785

Solving missing categorical data in questionnaire responses for automated classification

Saifon Aekwarangkoon[1], Thanatep Namponwatthanakul[2], Adisorn Amonwet[3], Siranuch Hemtanon[4]
[1]School of Nursing, Excellence Center of Community Health Promotion, Walailak University, Nakhon Si Thammarat, Thailand
[2]Faculty of Education, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat, Thailand
[3]Huasai Bumrungrat School, Nakhon Si Thammarat, Thailand
[4]Management Program, Faculty of Business Administration, Rajamangala University of Technology Srivijaya, Songkhla, Thailand

Article Info

Article history:

Received May 31, 2024
Revised Jun 11, 2025
Accepted Jul 5, 2025

Keywords:

Classification performance
Data imputation
Mental health screening
Missing categorical data
Questionnaire responses

ABSTRACT

Handling missing categorical data is critical for maintaining the accuracy and reliability of automatic classification tasks, particularly in mental health screening based on questionnaire responses. This study investigates several imputation methods, including last observation carried forward (LOCF), k-nearest neighbor (KNN) imputation, hot-deck imputation, and multivariate imputation by chained equations (MICE). Results show that KNN imputation achieves the lowest root mean square error (RMSE), indicating the most faithful reconstruction of the original data. However, for classification performance, MICE-imputed datasets produced models that outperformed those generated by other methods and even surpassed models trained on the original incomplete data. Interestingly, we also found that using observed data over multiple iterations of imputation tuning can introduce greater deviation from original missing values, but this process can help form datasets with clearer class boundaries, ultimately improving classification accuracy. These findings emphasize the need to balance data fidelity and model performance when selecting imputation strategies.

Corresponding Author:

Siranuch Hemtanon
Management Program, Faculty of Business Administration
Rajamangala University of Technology Srivijaya
2/3 63 Tower, Ratchadamnoennok road, Bo Yang subdistrict, Mueng district, Songkhla, Thailand
Email: Siranuch.h@rmutsv.ac.th

1. INTRODUCTION

Questionnaire is an efficient tool for collecting large amounts of data from a diverse population in a relatively short period. It allows for standardized data collection and ensuring consistency in the questions asked and the response options provided to all participants. This standardization minimizes bias and facilitates comparability across different respondents and studies. In addition, questionnaire data can inform decision-making processes in various domains, including healthcare, education, marketing, public policy, and organizational management. Insights derived from questionnaire responses can guide strategic planning, program development, resource allocation, and policy formulation [1], [2]. In healthcare, questionnaires are often used as invaluable tools for gathering information about patients' medical history, symptoms, lifestyle, and other relevant factors. Questionnaire responses can help to classify patients into different risk groups or disease categories, especially in the field of mental health for screening and diagnosis. There are many standardized questionnaires designed to assess symptoms associated with different mental disorders, such as depression [3], [4], anxiety [5], bipolar disorder [6], and posttraumatic stress disorder (PTSD) [7] for screening purposes. The responses to these questionnaires help classifying individuals into risk categories and indicating

whether further assessment or intervention is necessary. Generally, interpretation of the questionnaire responses follows the scoring and interpretation guidelines provided by the questionnaire's developers. The guidelines typically outline how to score individual items, calculate total scores or subscale scores, and interpret the results in terms of severity, risk levels, or clinical significance of a patient based on the scores.

With the growth of machine learning usage, automated classification of questionnaire responses is a common practice in various fields, including psychology and mental health. Several research studies [8]-[10] applied machine learning algorithms to classify individuals into different diagnostic or identify risk factors for specific mental health outcomes. One of the advantages of using machine learning algorithms is to identify subtle patterns and associations in questionnaire data that may not be apparent through manual analysis. Namely, automated classification techniques leverage advanced statistical methods to optimize accuracy and minimize errors in classifying respondents as well as minimizing the risk of human error or bias in the manual analysis process [11], [12]. Furthermore, automated classification is scalable and adaptable to diverse settings and populations as they can accommodate large datasets with varying sample sizes suitable for either small-scale studies or population-level surveys.

However, one of the challenges in applying machine learning is missing data. Missing data in questionnaire responses pose problems impacting the validity, reliability, and interpretability. Missing data bring bias into the analysis as the characteristics of respondents with missing data may differ from those with complete data leading to misclassification of variables, especially if missingness is related to the values of the variables themselves. This can distort the relationships between variables and lead to erroneous conclusions about the associations and patterns observed in the data. Despite the increasing use of machine learning for mental health screening, missing data remains a persistent obstacle. Studies show that missing values can occur in over 10–20% of questionnaire-based datasets, particularly when participants skip sensitive or complex questions. This compromises the ability of models to generalize effectively and increases the risk of biased predictions. Thus, addressing missing data effectively is essential to mitigate the problems and ensure the validity and reliability of questionnaire-based classification.

While missing data imputation has been explored in broader healthcare datasets, relatively few studies have focused on categorical questionnaire data in mental health contexts. Existing approaches often rely on basic strategies such as mode imputation or listwise deletion, with limited comparisons of more advanced imputation methods tailored to this data type. This paper aims to address the challenge of missing categorical data in mental health screening questionnaires, where the accuracy of automated classification is crucial. We explore how different imputation methods affect both the fidelity of reconstructed data and the downstream performance of classification models. We seek to answer the following questions: i) which imputation method most accurately reconstructs missing categorical questionnaire data? and ii) how do these methods impact the performance of machine learning classifiers trained on the imputed data? The questionnaire in this work refers to the questionaire designed to screen for mental health. Several data imputation techniques are applied to impute the missing data in questionnaire responses.

2. BACKGROUND

2.1. Missing data and data imputation

Missing data refers to the absence of values for variables in a dataset for training in machine learning processes. It occurs when respondents do not provide a response to certain questions or when data are not recorded. Missing data can be categorized into 3 forms [13] as follows.

- Missing completely at random (MCAR): missingness occurs randomly and is unrelated to the values of the variables or any other factors in the dataset. For example, missing data due to data entry errors or technical issues may be considered MCAR. The most common approach for handling MCAR data is to use complete case analysis, where observations with missing values are simply excluded from the analysis.
- Missing at random (MAR): missingness depends on the observed values of other variables in the dataset but not on the missing values themselves. For instance, if respondents are less likely to provide certain sensitive information, missing data on those variables may be considered MAR. Multiple imputation is often used to handle MAR data. This technique involves creating multiple imputed datasets, where missing values are replaced with estimated values based on observed data and relationships between variables.
- Missing not at random (MNAR): missingness is related to the values of the missing variables themselves. In other words, the probability of missingness depends on unobserved or unmeasured factors. For example, if individuals with high levels of depression are less likely to complete a questionnaire about mental health, missing data on depression scores may be considered MNAR. Tackling MNAR data poses additional challenges, as MNAR occurs when the probability of missingness depends on unobserved or unmeasured variables, making it difficult to model and handle effectively. Collecting additional auxiliary information related to the missingness mechanism can help mitigate the impact. For example, if data are missing due to

non-response to mental-sensitive questions, auxiliary data on respondents' characteristics or behaviors that may be associated with missingness should be collected and included in the analysis model.

Missing data is a prevalent challenge in machine learning, often leading to biased models, reduced predictive accuracy, and compromised generalizability. Studies have demonstrated that the presence of missing values can significantly degrade the performance of classification algorithms, particularly when the missingness is not completely at random [14], [15]. For instance, Buczak *et al.* [14] analyzed the effects of various imputation methods under different missing data mechanisms and found that strategies such as random forest–based imputation can help mitigate performance loss, while others may introduce bias depending on the missingness pattern. Similarly, Choudhury and Kosorok [15] proposed a class-weighted k-nearest neighbor (KNN) imputation method that incorporates class labels during imputation and showed improved classification outcomes over traditional techniques. These findings highlight the importance of choosing imputation methods that are well-suited to both the data structure and the classification task at hand. To mitigate missing data issue, data imputation techniques are used to fill in missing values in a dataset for the analysis of incomplete data. There are several data imputation techniques frequently used for completing training data for machine learning process. They can be categorized into 2 groups as simple imputation group, and statistic-based group as follows.

a. Simple imputation methods

- Mean/median/mode imputation: replace missing values with the mean, median, or mode of observed values for the variable.
- Last observation carried forward (LOCF): fill missing values with the most recent observed value in time-series or longitudinal data.
- Hot-deck imputation: assign missing values to observed values from similar cases based on a defined similarity criterion.

b. Statistic-based methods

- KNN imputation: KNN replaces missing values based on the responses of the most similar observations in the feature space. It is particularly suitable for categorical data, as similarity can be defined using distance metrics adapted for discrete variables. Its non-parametric nature makes it flexible for complex datasets where the underlying data distribution is unknown [16].
- Local least squares (LLS) imputation: LLS estimates missing values through local regression using the most similar instances identified by L2-norm or Pearson correlation. Though originally designed for continuous data, it is included for comparison due to its ability to model local variable relationships—a potentially useful trait in questionnaires with structured, scale-based categorical responses [17], [18].
- Singular value decomposition (SVD) imputation: SVD is rooted in low-rank matrix approximation and is widely used for dimensionality reduction and missing value estimation. While primarily applied to continuous data, it was selected here to examine whether underlying latent structures in questionnaire responses (e.g., symptom clusters) can be leveraged for imputation despite categorical formats [19].
- Optimizer-based imputation: techniques such as genetic algorithms (GA) and bee algorithm use stochastic optimization to infer missing values. These are included to test whether global optimization strategies can outperform traditional statistical methods in capturing complex dependencies. However, most optimizer-based methods are better suited for numerical data, so their inclusion also helps explore their limits on categorical data [20], [21].

In this work, the missing data can be any of the missing forms. However, data imputation for this work is focused based on MCAR and MAR, as MNAR requires auxiliary information from observing behavior of questionnaire responders to impute data effectively. Furthermore, since this work aims to tackle categorical missing data in a questionnaire, some methods including bee-based imputation and LLS may not be able to apply for imputing the missing data in this work.

2.2. Related work

There are many publications on using data imputation techniques. In this section, we synthesize their information and give a brief summary of the recent works. Song and Shepperd [22] studied on KNN Imputation to improve performance of a classification model. In their work, they randomly deleted numerical values in a dataset for 10, 20, 30, 40, and 50% as five incomplete datasets for testing the performance of the imputation. The KNN was applied to impute the missing data to be trained for decision tree classification model. They compared the classification of dataset with missing value and imputed value against the model from the original dataset (complete dataset). The results of this study indicated that the imputation of missing numerical data using KNN helps to improve the accuracy of the decision tree model.

Malarvizhi and Thanamani [23] propose a comparative study on single imputation techniques such as mean, median, and standard deviation combined with KNN algorithm. Training set with their corresponding

class groups the data of different sizes. The above techniques are applied in each group and the results are compared. Median/standard deviation shown better result than mean substitution.

Troyanskaya *et al.* [24] published a comparative study of three imputation methods including KNN imputation, SVD imputation, and Row average imputation on the DNA microarray dataset. The dataset is a time-series of yeast saccharomyces cerevisiae [25]. The data were processed to remove instances with missing value from a complete dataset for 1 to 20%. The experimental results signified that KNN imputation method outperformed and was more robust than the rest for a time-series data. The results were also concluded that KNN yielded the best imputation performance with 10% missing value.

AI-Helali *et al.* [26] proposes a new imputation method for symbolic regression with incomplete data. This method uses genetic programming (GP) and weighted KNN. It constructs GP-based models using other available features to predict the missing values of incomplete features. The instances used for constructing such models are selected using weighted KNN. The experimental results on real-world data sets show that the proposed method outperforms a number of state-of-the-art methods with respect to the imputation accuracy, the symbolic regression performance, and the imputation time.

Patil and Bichkar [27] published multiple imputation of missing data with GA based techniques to show how to apply GA for multiple imputation. In the research, they experimented by comparing the classification model from J48 and CART with 3 numerical datasets as breast cancer dataset, weather dataset, and lymphography dataset. The result showed that the GA imputation gained significantly higher performance in terms of imputation accuracy. However, the experiments lacked comparison to other algorithms and did not study on performance regarding amount of missing value.

These studies demonstrate that data imputation plays a vital role in addressing missing data for machine learning–based applications, with numerous techniques showing effectiveness—particularly for numerical data. However, in the context of categorical data derived from questionnaire responses, especially those used for mental health screening, far fewer studies have been conducted. Despite the widespread acceptance and use of mental health questionnaires in clinical and research settings, most existing imputation approaches either overlook categorical variables or rely on simplistic methods such as mode imputation or listwise deletion. Unlike numerical data, where imputation can rely on mathematical operations such as means, medians, or regression-based predictions, categorical data imputation presents unique challenges due to the discrete and non-ordinal nature of values. Operations like averaging are not meaningful for nominal categories, and the imputation process must preserve the integrity of class labels without introducing artificial patterns. Furthermore, while numerical imputation often minimizes error based on distance or distribution assumptions, categorical imputation requires techniques that respect frequency distributions and class balance, which are critical for tasks like classification. As a result, methods effective for numerical data, such as mean imputation or low-rank matrix factorization, may perform poorly or produce biased results when applied directly to categorical data. This underscores the need for imputation techniques specifically adapted to handle the characteristics of categorical variables, particularly in domains like mental health where questionnaire responses often consist of ordinal or nominal scales. Consequently, there remains a gap in systematically evaluating advanced imputation techniques for categorical questionnaire data. This work aims to address that gap by investigating the impact of various imputation strategies on classification performance in mental health screening tasks.

3. METHOD

This work focuses on solving missing data in questionnaire for automated classification. The overview of the methods is illustrated in Figure 1.

3.1. Dataset and data preparation

The dataset in this work is a collection of questionnaire responses for mental health. The questionnaire is to screen for mental health problems including depression, stress, and anxiety disorder under the research license ID WUEC-24-386-01. The questionnaire consists of two parts of questions regarding general information and mood/emotion information. The former has 17 questions about age, gender, academic grade, household financial situation, current guardian, and parental marriage status. The latter composes of 35 questions including query about concentration/focus, thoughts and perceptions, sleep pattern, appetite, for example. All survey questions are choice questions which responders can choose only one answer. The responders are 1,089 Thai high school students in Norther Region of Thailand by random sampling method. For labeling, mental health experts apply a guideline to score their answers and categorizes the result into four labels as no risk, risk of depression disorder, risk of stress disorder, and risk of anxiety disorder. To simulate missing data for the experiment, synthetic missing values were introduced by randomly removing 2%, 5%,

10%, and 20% of the values from the dataset. The missing values were introduced independently across five separate batches.

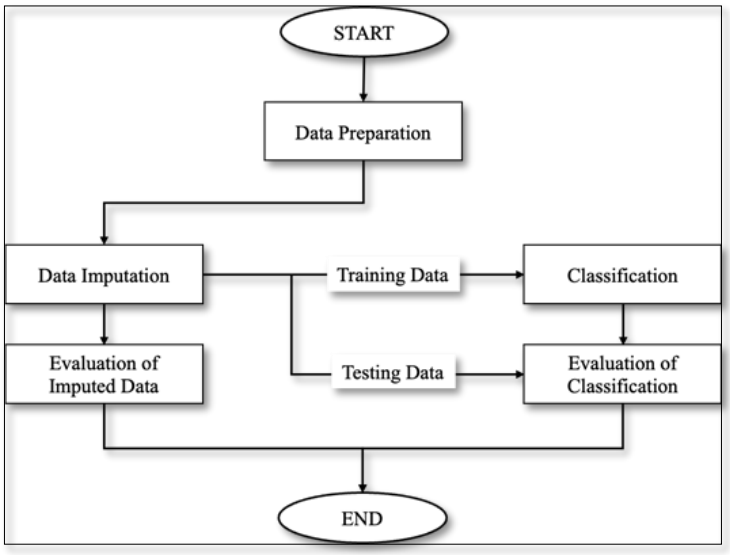

Figure 1. An overview of the framework

3.2. Data imputation

To solve missing data, data imputation techniques are applied to fill in missing values in a dataset. In this paper, we choose 4 commonly used imputation techniques including LOCF, KNN imputation, hot-deck imputation, and multivariate imputation by chained equation (MICE). The details of each imputation methods are as follows.

- LOCF: for each missing value, look back in time to find the most recent observed value before the missing point. The observed value then is used to fill in the missing value.
- KNN imputation: the method focuses on finding a set of the closely related neighbor (values) by calculating the distance of the attribute values to predict the missing value. For each data point with a missing value, calculate its distance to all other data points in the dataset (excluding other missing values). Then, select the KNN based on the calculated distances. These are the data points with the most similar feature values to the data point with the missing value. For each missing value, take the mode of the corresponding feature values from the KNN and assign the mode value to the missing value. For this work, 5 and 10 neighbors are chosen.
- Hot-deck imputation: the mothod begins with measuring similarity between units based on categorical variables using Dice coefficient. For each data point with a missing categorical value, find the nearest neighbors with observed values based on the defined similarity measure for categorical variables. Use the one random observed value from the nearest neighbors to impute the missing value in the original data point. For this work, 5 and 10 neighbors are chosen.
- MICE: this method imputes missing values by iteratively modeling each variable with missing data as a function of the other variables in the dataset. In this work, we use multinomial logistic regression for variables since the answers of the questionnaire are more than two categories. For each categorical variable with missing values, it builds a separate imputation model. These models predict the missing categories based on the observed categories and other variables in the dataset. In the chained equations step, it iteratively updates the imputed values for each variable while considering the imputed values of other variables. This process continues for multiple iterations until achieving convergence of reaching a consistent state (where further iterations do not significantly change the imputed values).

3.3. Classification

In this study, the impact of data imputation methods is evaluated based on downstream classification performance. To isolate the effects of imputation quality from those of complex learning algorithms, we employ the ID3 decision tree algorithm, implemented via the scikit-learn library. ID3 is a classic supervised learning method that uses information gain to iteratively select the most informative attribute for splitting at

each node, building a tree structure that is both interpretable and aligned with human decision logic. The algorithm is well-suited for categorical features, as it handles discrete-valued input efficiently and creates splits that are intuitive to understand. As with most decision tree models, ID3 provides a high level of explainability by producing interpretable decision paths that show how input features influence outcomes—a critical aspect for applications in mental health screening.

A multiclass classification setting is used, reflecting the categorical nature of the target variable derived from questionnaire responses. For consistency and to emphasize the effect of imputation rather than model tuning, we apply the default hyperparameters of the DecisionTreeClassifier in scikit-learn, which are conceptually aligned with ID3 behavior (criterion="entropy" and max_depth=None).

3.4. Evaluation

In this work, we evaluate two key aspects: imputation performance and classification performance. First, for imputation performance, we use root mean squared error (RMSE) to compare the imputed values to the original values. RMSE is a widely used metric to quantify the average magnitude of errors between imputed and actual values. It is calculated by taking the square root of the average of the squared differences between imputed and original values, as shown in (1). A lower RMSE indicates that the imputed values are closer to the original values.

For classification performance, we calculate accuracy to assess the effectiveness of the classification model based on the imputed data. The classification accuracy is computed using (2), where a higher accuracy signifies better classification performance:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} \quad (1)$$

where N is the number of observations. $y(i)$ is the actual value of the i^{th} observation, while $\hat{y}(i)$ is the predicted value of the i^{th} observation:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (2)$$

where $TP, FP, TN,$ and FN represent the number of true positives, false positives, true negatives and false negatives, respectively.

4. RESULTS

4.1. Evaluation of imputation

In this section, experiment results of data imputation are presented. The settings include a variation of missing data as 2, 5, 10, and 20% missing values, and different imputation methods as LOCF, KNN-5, KNN-10, Hotdeck-5, Hotdeck-10, and MICE. The evaluation metric is RMSE where the lower score represents the higher performance. The RMSE results are given in Table 1.

Table 1. RMSE results of data imputation methods

Missing percentage (%)	Average RMSE					
	LOCF	KNN-5	KNN-10	Hotdeck-5	Hotdeck-10	MICE
2	1.04	0.53	0.55	0.61	0.64	0.63
5	1.02	0.53	0.55	0.62	0.64	0.64
10	1.05	0.54	0.56	0.62	0.66	0.63
20	1.07	0.54	0.55	0.64	0.65	0.66

From the results, we found that KNN imputation with a set of 5 neighbors produced the least errors among all applied methods, while imputed data from LOCF contained the most errors in average. Since RMSE represents the resemblance of imputed data and actual data, we can conclude that both KNN imputations generate the most closed values for categorical data of questionnaire answers. The difference missing percentages unfortunately does not show the significant difference of errors among all applied methods.

4.2. Evaluation of classification accuracy

To evaluate the imputed data in classification, we apply 5-fold validation to split training and testing data. The evaluation metric in this part is accuracy score. We also include the classification of the intact original dataset (no value removed) for comparison. The accuracy results are given in Table 2.

Table 2. Accuracy results of classifier trained from imputed datasets

Missing percentage (%)	Average RMSE					
	LOCF	KNN-5	KNN-10	Hotdeck-5	Hotdeck-10	MICE
2	69.64	73.39	73.38	71.49	71.87	77.63
5	70.85	72.82	73.03	71.73	72.06	77.49
10	71.08	73.40	73.13	71.95	72.23	78.26
20	69.65	72.97	73.05	71.92	71.98	78.30
Intact	75.58					

In terms of accuracy of the classifier using imputed data, the baseline from the original dataset is 75.58. The imputed dataset using MICE however generated a classification model that can yield higher accuracy score than the original dataset. On the other hands, datasets from other imputation methods obtained less accuracy score.

4.3. Discussion

Existing studies on missing data imputation have primarily focused on datasets with missing attribute values and have seldom explored questionnaire data. This paper investigates how imputation techniques can address missing values in questionnaire responses, evaluating four commonly used methods: LOCF, KNN imputation, hot-deck imputation, and MICE.

From the results, we observe that KNN achieved the best imputation performance, producing values closest to the original data according to RMSE. In contrast, datasets imputed using MICE yielded the highest classification performance, even surpassing the classifier trained on the original complete dataset. This finding is particularly noteworthy and requires deeper interpretation.

In KNN imputation, missing values are imputed based on a majority vote from the KNN. Given that our dataset is categorical with fewer than 10 distinct values per question, KNN performs effectively by preserving the original structure and local patterns of the data.

MICE imputation, while not producing the lowest RMSE, iteratively models each variable based on others through chained equations. This iterative process strengthens the internal relationships among variables, producing imputed data that may not exactly match the original missing values but enhance the overall separability of classes. Essentially, MICE imputes values that conform more tightly to the underlying multivariable patterns, leading to a form of data smoothing or implicit regularization. This effect results in better-defined decision boundaries for classification, which explains why MICE-imputed datasets outperform the original data in classification tasks. This phenomenon can be interpreted as a form of data augmentation, where the imputed data enhances the signal-to-noise ratio for the classifier.

However, this advantage comes with a trade-off. While MICE improves classification, it may not be suitable for applications that require faithful recovery of the original data (e.g., descriptive analysis or profiling), since the imputed values may diverge from the true values.

Regarding degradation of performance, datasets imputed by LOCF showed the worst results in both imputation (highest RMSE) and classification accuracy. LOCF's simple strategy of carrying forward the last observed value does not preserve inter-variable dependencies, especially in complex questionnaire data, leading to both inaccurate imputations and poor model performance.

In addition to imputation and classification performance, computational efficiency and scalability are important considerations when choosing an imputation method. Among the four methods tested, LOCF is the most computationally efficient due to its simplicity and lack of dependency on other variables. Hot-deck and KNN imputation are moderately demanding. KNN requires distance computations for every missing value, which can be time-consuming as dataset size grows. The most computationally intensive method is MICE, as it involves iterative modeling of each variable through chained equations and multiple rounds of convergence. While MICE yielded the best classification results, its runtime increases significantly with the number of variables and missing entries, making it less scalable to large datasets. Therefore, practitioners should weigh accuracy improvements against resource availability when choosing an imputation strategy.

5. CONCLUSION

This study explored the use of data imputation techniques to address missing categorical responses in questionnaires, with a focus on enhancing automatic classification performance. Four imputation methods were evaluated: LOCF, KNN, hot-deck imputation, and MICE. Experimental results show a dual outcome: KNN yielded the most accurate imputations in terms of similarity to the original values, while MICE significantly improved classification performance—surpassing even the performance of models trained on the complete original dataset. These findings suggest that the choice of imputation method should align with the intended downstream task—whether preserving data fidelity or optimizing classification accuracy.

In practical terms, this work highlights the potential of imputation-driven preprocessing to improve automated mental health screening tools, particularly when handling incomplete self-reported questionnaire data. Future research can extend this work along several directions. First, expanding the study to include datasets with different characteristics—such as varying sample sizes, more diverse question formats, or datasets from different domains beyond mental health—would help validate the generalizability of the findings. Second, while this study assumes that missing data occur randomly, real-world datasets often exhibit MNAR patterns, where the likelihood of missingness depends on unobserved data. Addressing MNAR scenarios would involve developing more sophisticated imputation models that can account for underlying missingness mechanisms, potentially integrating domain knowledge into the imputation process. Lastly, designing hybrid imputation frameworks—combining the strengths of methods like KNN (for fidelity) and MICE (for predictive modeling)—could optimize both the accuracy of imputation and the performance of machine learning classifiers. Such hybrid approaches may dynamically adapt the imputation strategy based on the ultimate application goal, whether it is data recovery, feature engineering, or classification.

ACKNOWLEDGMENTS

The Path 2 Health Foundation, under the Thai Health Promotion Foundation, financially supported this research. The authors would like to express their sincere gratitude for the support provided, which played a crucial role in the successful completion of this study.

FUNDING INFORMATION

This work was supported by Path 2 Health Foundation under the Thai Health Promotion Foundation, Thailand [Grant No. 1050/THP19-2021].

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Saifon Aekwarangkoon	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Thanatep Namponwatthanakul				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Adisorn Amonwet				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Siranuch Hemtanon	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all participants and their parents included in this study.

ETHICAL APPROVAL

The dataset in this work is a collection of questionnaire responses for mental health. The questionnaire is used to screen for mental health problems including depression, stress, and anxiety disorder. This study obtained approval from the Human Research Ethics Committee of Walailak University, Thailand, under Research License ID WUEC-24-386-01.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [initials: SH]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] S. Shamim, J. Zeng, Z. Khan, and N. U. Zia, "Big data, analytics capability and decision making performance in emerging market firm: The role of contractual and relational governance mechanisms," *Technological Forecasting and Social Change*, vol. 161, pp.1-10, Sep. 2020, doi: 10.1016/j.techfore.2020.120315.

[2] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 215–225, Apr. 2016, doi: 10.1016/j.rser.2015.11.050.

[3] L. Manea, S. Gilbody, and D. McMillan, "A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression," *General Hospital Psychiatry*, vol. 37, no. 1, pp. 67–75, Sep. 2015, doi: 10.1016/j.genhosppsych.2014.09.009.

[4] T. Q. Nguyen, K. Bandeen-Roche, J. K. Bass, D. German, N. T. T. Nguyen, and A. R. Knowlton, "A tool for sexual minority mental health research: The Patient Health Questionnaire (PHQ-9) as a depressive symptom severity measure for sexual minority women in Vietnam," *Journal of Gay & Lesbian Mental Health*, vol. 20, no. 2, pp. 173–191, 2016, doi: 10.1080/19359705.2015.1080204.

[5] M. Balsamo, F. Cataldi, L. Carlucci, and B. Fairfield, "Assessment of anxiety in older adults: A review of self-report measures," *Clinical Interventions in Aging*, vol. 13, pp. 573–593, Apr. 2018, doi: 10.2147/CIA.S114100.

[6] C. F. Balassano, "Assessment tools for screening and monitoring bipolar disorder," *Bipolar Disorders: An International Journal of Psychiatry and Neurosciences*, vol. 7, no. 1, pp. 8–15, Mar. 2005, doi: 10.1111/j.1399-5618.2005.00189.x.

[7] P. Rodriguez, D. W. Holowka, and B. P. Marx, "Assessment of posttraumatic stress disorder-related functional impairment: A review," *Journal of Rehabilitation Research and Development*, vol. 49, no. 5, pp. 649–666, Jan. 2012, doi: 10.1682/JRRD.2011.09.0162.

[8] J. Chung and J. Teo, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," *Applied Computational Intelligence and Soft Computing*, pp. 1–19, Jan. 2022, doi: 10.1155/2022/9970363.

[9] S. Aleem, N. U. Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine learning algorithms for depression: Diagnosis, insights, and research directions," *Electronics*, vol. 11, no. 7, pp. 1–20, Mar. 2022, doi: 10.3390/electronics11071111.

[10] M. Arif *et al.*, "Classification of anxiety disorders using machine learning methods: A literature review," *Insights of Biomedical Research*, vol. 4, no. 1, pp. 95–110, Nov. 2020, doi: 10.36959/584/455.

[11] C. H. Brown, E. W. Holman, S. Wichmann, and V. Velupillai, "Automated classification of the world's languages: A description of the method and preliminary results," *Language Typology and Universals*, vol. 61, no. 4, pp. 285–308, Nov. 2008, doi: 10.1524/stuf.2008.0026.

[12] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., New York, NY, USA: John Wiley & Sons, 2002.

[13] P. D. Allison, "Missing Data," in *The SAGE Handbook of Quantitative Methods in Psychology*, R. E. Millsap and A. Maydeu-Olivares, Eds. London, UK: Sage Publications Ltd., 2009.

[14] E. M. Buczak, R. W. Carroll, P. A. Sattigeri, and J. van Wingerden, "Evaluation of machine learning imputation methods in clinical data under different missingness mechanisms," *PLOS ONE*, vol. 18, no. 3, Mar. 2023, doi: 10.1371/journal.pone.0282094.

[15] M. Choudhury and M. R. Kosorok, "Missing data imputation for classification problems," *arXiv preprint*, Feb. 2020, doi: 10.48550/arXiv.2002.10709.

[16] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, "A novel kNN algorithm with data-driven k parameter computation," *Pattern Recognition Letters*, vol. 109, pp. 44–54, Oct. 2017, doi: 10.1016/j.patrec.2017.09.036.

[17] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: Local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, Jan. 2005, doi: 10.1093/bioinformatics/bth499.

[18] X. Zhang, X. Song, H. Wang, and H. Zhang, "Sequential local least squares imputation estimating missing value of microarray data," *Computers in Biology and Medicine*, vol. 38, no. 10, pp. 112–120, Oct. 2008, doi: 10.1016/j.compbiomed.2008.08.006.

[19] R. Wei *et al.*, "Missing value imputation approach for mass spectrometry-based metabolomics data," *Scientific Reports*, vol. 8, no. 1, p. 663, Jan. 2018, doi: 10.1038/s41598-017-19120-0.

[20] W. Shahzad, Q. Rehman, and E. Ahmed, "Missing data imputation using genetic algorithm for supervised learning," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, pp. 438–445, Mar. 2017, doi: 10.14569/IJACSA.2017.080360.

[21] K. H. Olsson, N. Weiss, S. Shalev, and J. Schackermann, "Spread of the invasive dwarf honey bee Apis florea facilitates winter presence of oriental honey buzzard Pernis ptilorhynchus in Eilat, Israel," *Acta Ornithologica*, vol. 56, no. 2, pp. 189–198, Mar. 2022, doi: 10.3161/00016454AO2021.56.2.005.

[22] Q. Song and M. Shepperd, "A new imputation method for small software project data sets," *Journal of Systems and Software*, vol. 80, no. 1, pp. 51–62, Jan. 2007, doi: 10.1016/j.jss.2006.05.003.

[23] R. Malarvizhi and A. S. Thanamani, "K-nearest neighbor in missing data imputation," *International Journal of Engineering Research and Development*, vol. 5, no. 1, pp. 5–7, Nov. 2012.

[24] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Feb. 2001, doi: 10.1093/bioinformatics/17.6.520.

[25] A. P. Gasch *et al.*, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, no. 12, pp. 4241–4257, Dec. 2000, doi: 10.1091/mbc.11.12.4241.

[26] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data," *Soft Computing*, vol. 25, no. 8, pp. 5993–6012, 2021, doi: 10.1007/s00500-021-05590-y.

[27] D. V. Patil and R. S. Bichkar, "Multiple imputation of missing data with genetic algorithm-based techniques," *International Journal of Computer Applications, Special Issue on Evolutionary Computation for Optimization Techniques (ECOT)*, pp. 74–78, 2010.

BIOGRAPHIES OF AUTHORS

Assoc. Prof. Dr. Saifon Aekwarangkoon is an instructor at the School of Nursing, Walailak University, Thailand. She graduated with a Ph.D. from Prince of Songkla University and specialized in psychiatric and mental health nursing. She works to help patients, families, schools, communities, and organizations with mental health and psychiatric issues through research, teaching, and academic service in collaboration with networks since 2000. She has opened a smile clinic for people facing mental health and psychiatric problems and published books and researches. She can be contacted at email: saifon.aekwarangkoon@gmail.com.

Dr. Thanatep Namponwatthanakul holds a Ph.D. in Education (Educational Administration) from Nakhon Si Thammarat Rajabhat University, Thailand. He is currently instructor of Faculty of Education, Nakhon Si Thammarat Rajabhat University, no. 1 Tha Ngio Subdistrict, Mueang District, Nakhon Si Thammarat 80280, Thailand. His research interests include education administration, student support system, innovation management for development, research for learning develoment, project measurement, and evaluation. He can be contacted at email: thanatep_nam@nstru.ac.th.

Mr. Adisorn Amonwet is a guidance counselor, Huasai Bumrungrat School, No. 330 Moo 1, Hua Sai Subdistrict, Hua Sai District, Nakhon Si Thammarat Province 80170, Thailand. He graduated with a bachelor's degree in psychology from Phra Nakhon Rajabhat University, has experience in individual counseling, group counseling, and taking care of students in school. He can be contacted at email: adisorn@hbr.ac.th.

Dr. Siranuch Hemtanon holds a Ph.D. in Management Information Technology from Walailak University, Thailand. She is currently instructor of Management Program, Faculty of Business Administration, Rajamangala University of Technology Srivijaya, no.1 Ratchadamnoen Nok Road, Bo Yang Subdistrict, Mueang District, Songkhla 90000, Thailand. Her research interests include data mining, natural language processing, machine learning, and question answering systems. She can be contacted at email: Siranuch.h@rmutsv.ac.th or Mitnuchie@gmail.com.