Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 18(1) 2025: 61-72
http://ejournal.undip.ac.id/index.php/media_statistika

COMPARISON OF MISSING VALUE IMPUTATION USING MEAN, BAYESIAN KNN, AND NON-BAYESIAN KNN ON TEP GENE EXPRESSION DATA

Mastika, Titin Siswantining, Alhadi Bustamam
Master's Program in Mathematics, Universitas Indonesia, Depok, Indonesia

e-mail: titin@sci.ui.ac.id

DOI: 10.14710/medstat.18.1.61-72

Article Info:
Received: 16 March 2025
Accepted: 13 October 2025
Available Online: 16 October 2025

Keywords:
Mean Absolute Error; Mean Squared Error; Normalized Root Mean Squared Error; Gaussian Process; Optimization
.

Abstract: Analysis of gene expression data, particularly in cancer data, often faces challenges due to the presence of missing values. One approach to overcome this is data imputation. This study evaluates the performance of three imputation methods, namely mean imputation, K-Nearest Neighbors (KNN), and KNN with Bayesian optimization using Gaussian Process modeling, on Tumor Educated Platelets (TEP) gene expression data. Missing values were introduced using Missing Completely at Random (MCAR) gradually at levels of 5%, 10%, 15%, and up to 60%, and performance was evaluated using three metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Normalized Root Mean Squared Error (NRMSE). The results show that the three methods produce relatively similar performance, with differences in MAE, MSE, and NRMSE values only at a small decimal scale. Although Bayesian Optimization is expected to improve the accuracy of KNN, the resulting improvement on this dataset is not significant. These findings indicate that simple imputation such as the average and KNN-based methods still provide competitive results on TEP data with data characteristics that have 14,020,496 zeros out of a total of 16,512,496 existing values, which is approximately 84.91% of the total data.

1. INTRODUCTION

Bioinformatics is a field that uses computational techniques to analyze biological data, such as gene expression data. This data is often used to help classify cancer types and improve diagnostic accuracy (Ravindran & Gunavathi, 2023). Cancer is the second leading cause of death worldwide after cardiovascular disease (Miller et al., 2021). Early detection of cancer can be done through non-invasive biomarkers, one of which is Tumor Educated Platelets (TEP). TEP carries RNA from cancer cells and plays an important role in liquid biopsy methods, which is the examination of cancer through blood samples without surgery. However, RNA-seq data is complex, so it requires a robust and accurate analysis method to process it (Liu et al., 2020).

The presence of missing values in gene expression data poses a major obstacle in the analysis process, as it can affect the results of normalization, feature selection, and even the final biological interpretation (Brown et al., 2018). Therefore, the proper handling of missing values is essential to ensure the reliability of further analysis of gene expression data.

Imputation aims to detect and correct errors in data, which has long been a topic of debate in health and biological data literature, especially in relation to large data sets such as gene expression data or medical records (Ismail et al., 2022). The presence of missing values poses a major challenge in various fields, especially in health, as it can hinder accurate decision-making (Ayilara et al., 2019). Machine learning approaches such as Random Forest have been proven to outperform traditional methods in the imputation process (Mostafa, 2019).

In addition, Latief et al. (2020) evaluated the performance of XGBoost in handling missing values in liver cancer (hepatocellular) gene expression data and found that this model continued to perform very well even without imputation. However, when using KNN-based imputation, model performance varied depending on the percentage of missing values, with the best results obtained when missing values were 10%. These findings emphasize the importance of selecting the appropriate imputation technique to optimize classification performance. In line with this, Farswan et al. (2020) investigated the effectiveness of the Deep Sparse Neural Network (DSNN) method in handling missing values in blood cancer gene expression data. The results showed that DSNN outperformed the KNN, SVM, and PCA methods, even under conditions of missing values varying from 10% to 90%.

In a different approach, Siswantining et al. (2021) optimized missing data imputation using the K-Harmonic Means (KHM) method, so that the imputed values approximated the actual data distribution. Then, recent research by Jafrasteh et al. (2023) introduced the Missing Gaussian Process (MGP) approach, a hierarchical composition of variational sparsity Gaussian Processes inspired by deep GP and recurrent GP, which showed superior performance compared to other imputation methods such as KNN, multiple imputations using chained equations (MICE), generative adversarial network (GAIN), deep belief networks (DBN), variational Auto-encoders (VAE), deep gaussian process (DGP), and sparse variational gaussian process (SVGP) in terms of RMSE and classification accuracy. MGP performs very well when the proportion of missing data is not too high, making it one of the models that can be used for gene expression data imputation.

Research on data imputation continues, and Chungnoy et al. (2024) introduced a new bee-based imputation method, namely Bees-based KNN Linear regression (BKL), which integrates KNN and Linear Regression. This method demonstrates improved accuracy compared to conventional methods such as KNN, Probabilistic PCA, LLS, SVD, NLPCA, and MIDASpy across various cancer datasets.

Based on previous studies, various models have been developed for imputing missing values, ranging from the simplest approach, mean imputation, to more complex methods such as Deep Sparse Neural Network (DSNN), optimized K-Nearest Neighbors (KNN), and Gaussian Process. Previous research has explored methods including Random Forest (Mostafa, 2019), XGBoost (Latief et al., 2020), DSNN (Farswan et al., 2020), K-Harmonic Means (Siswantining et al., 2021), and Missing Gaussian Process (Jafrasteh et al., 2023). Although these methods have demonstrated good performance, most studies focus on applying a single model without adaptive parameter optimization tailored to the data characteristics. In addition, many studies use gene expression data from common cancers, such as liver or blood cancers, which do not contain many non-missing zero values, leaving the challenge of imputing data with high zero distributions less explored.

Most missing value imputation studies focus on general gene expression data, with little attention to Tumor-Educated Platelets (TEP), which have unique biological characteristics such as a high proportion of valid zero values. These zeros can influence

distance calculations in KNN or bias mean estimates in simple imputation methods. Existing methods, including mean imputation, KNN, and Bayesian KNN, have not been systematically compared on TEP data, despite its growing use in cancer diagnostics. Previous research on genomic data has shown that ensemble approaches, which integrate multiple imputation methods, can achieve higher accuracy, robustness, and generalization than single-method imputations (Zhu et al., 2021). This study addresses the gap by evaluating and comparing mean, KNN, and Bayesian KNN specifically for TEP gene expression data, aiming to improve imputation accuracy while preserving the original biological patterns.

2. LITERATURE REVIEW

2.1. Tumor-Educated Platelets (TEP)

Tumor-Educated Platelets (TEP) are platelets that undergo changes due to interaction with tumors in the body. TEP function as an important component in the body's response to tumor growth. They are influenced by the tumor environment and can absorb genetic information, such as messenger RNA (mRNA), which can be used to detect the presence and type of cancer. TEP have become a major focus in cancer diagnostic research due to their role in detecting and monitoring cancer progression and response to therapy (In 't Veld & Wurdinger, 2019).

2.2. Missing Values

Missing values are incomplete or partially missing data (Little & Rubin, 2019). The existence of missing values is common and can have a significant impact on the conclusions of research results. Problems in analysis that can be caused by the existence of missing values include biased parameter estimation, reduced effectiveness, low accuracy of conclusions, and the inability to continue the analysis process (Salleh & Samat, 2017).

According to Little & Rubin (2019), the mechanisms for missing values are divided into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). This study uses MCAR. Missing Completely at Random (MCAR) is a high level of randomness in missing data, indicating that the pattern of missing data is completely random and independent of any variables (Ramanathan et al., 2019). In other words, the missing data is independent of the variables being studied and other parameters in the dataset. When missing values are evenly distributed across measurements, the data can be categorised as MCAR. To test this, a comparison can be made between two datasets, one with missing data and one without missing data. If the t-test results show no significant difference in the means between the two datasets, it can be concluded that the data is MCAR (Hameed & Ali, 2023). Mathematically, MCAR can be expressed as:

$$P\left(P_1 \middle| X, Y_{0,l}, Y_{m,l}\right) = f(l, X)$$

where $P\left(P_1 \middle| X, Y_{0,l}, Y_{m,l}\right)$ represents the probability of missingness in the l variable, conditional on the covariates X, the observed part of the data ($Y_{(0,l)}$), and the missing part ($Y_{(m,l)}$); X denotes the set of covariate variables or explanatory variables in the dataset; $Y_{(0,l)}$ refers to the observed portion of the data in the l^{th} variable; $Y_{(m,l)}$ refers to the missing portion of the data in the l^{th}variable; $f(l, X)$ is a function indicating that the pattern of missingness depends only on the covariates X and is not influenced by the observed or unobserved values of Y, where f is a function, that is, the missing data patterns are determined only by the covariate variables X.

2.3. Missing Value Imputation Technique

A. Mean Imputation

The mean imputation technique calculates the mean of the non-missing values for a given attribute to replace missing data. It is simple, quick, and widely available in most statistical software packages. This method is effective for small datasets and produces accurate results in such cases. However, it may lead to inaccuracies in large datasets. Mean imputation is suitable for data missing at random (MAR) but is not recommended for data missing completely at random (MCAR). Mathematically, the formula for mean imputation is:

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in c_k} \frac{x_{ij}}{n_k}$$

where n_k represents the number of non-missing values in feature j of class k (C_k) (Puri & Gupta, 2017; Hameed & Ali, 2023).

B. K-Nearest Neighbor (KNN) Imputation

The KNN imputation method identifies the similarity between data points and replaces missing values with similar ones using Euclidean distance. This technique is advantageous for datasets containing both qualitative and quantitative attributes, as it does not require creating a predictive model for each missing attribute. Additionally, it is effective in handling multiple missing values. However, a notable drawback is that the algorithm searches through the entire dataset to find similar instances, which can be computationally intensive (Hameed & Ali, 2023). In KNN, we indeed use the Euclidean distance formula, which is expressed as:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{m} (x_{aj} - x_{bj})^2}$$

with $d(x_a, x_b)$ is the Euclidean distance between gen x_a (containing missing values) and gene x_b (with complete data); x_{aj} is the expression value of gene $x_a$in sample j; x_{bj} is the expression value of gene x_b in sample j; m is the number of observations used in the distance calculation, i.e., samples that contain complete data for both genes x_a and x_b (Foud *et al*., 2021).

C. Bayesian Optimization

Missing Not at Random (MNAR or NMAR) occurs when the probability of missingness is related to unobserved information, meaning that the likelihood of data being missing depends on the actual unobserved value itself or on variables outside the dataset (Little & Rubin, 2019). In this case, the missing data pattern cannot be fully explained or predicted using other observed variables. For example, in a depression study, data may be considered MNAR if participants with more severe depression are more likely to refuse completing a survey on depression severity.

In MNAR settings, the missingness mechanism is systematically linked to the unobserved data, which makes handling this type of missingness particularly challenging. As in the MAR case, complete-case analysis may or may not produce bias. However, when bias occurs under MNAR, it generally cannot be resolved analytically because the cause of missingness is itself unmeasured. A common misconception is that complete-case analysis always produces unbiased estimates in MCAR and always biased estimates in MNAR.

In fact, whether bias arises depends on the causal structure of the missingness process. As shown in Daniel et al. (2012) and Westreich (2012), complete case analysis can remain

unbiased if missingness is independent of the outcome variable, a situation that can occur under both MAR and MNAR. If missingness is not independent of the outcome, bias can only be addressed analytically when the missingness is MAR, but not under MNAR. Mathematically, MNAR can be expressed as:

$$P(p_1|X, y_{0,l}, Y_{m,l}) = f(l, X, Y_{0,l}, Y_{m,l})$$

where f is a function indicating that the patterns of missing data are influenced by all three types of variables (Hameed & Ali, 2023).

2.3. Error Evaluation Metrics

A. Normalized Root Mean Squared Error (NRMSE)

NRMSE is defined as a parameter of the average error in analytical methods, measuring the difference between the estimated and observed original values. The formula for NRMSE is

$$\sqrt{\frac{\sum_{i=1}^{n_{mv}}(x_i - \hat{x}_{ij})^2 / n_{mv}}{\sigma_{x_j}}}$$

where x_{ij} is the $i-th$ value from the complete observation data; $\hat{x}_{ij}$ is the imputed value of the $i-th$ missing value; σ_{x_j} is the standard deviation of the observed data; n_{mv} is the total number of missing values in the complete observation.

The imputation result is considered accurate when the NRMSE value is relatively small or approaches zero, indicating that the imputed values closely match the original data (Al Janabi & Alkaim, 2020).

B. Mean Squared Error (MSE)

Mean Squared Error is a metric used to measure the average squared difference between the expected values and the predicted output values. It calculates the error magnitude by squaring each prediction error, making it sensitive to large deviations. A smaller MSE value indicates better prediction accuracy, as the errors are minimal. The formula for MSE is as follows (Khan, 2024):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ represents the predicted value and y_i is the true value

C. Mean Absolute Error (MAE)

The Mean Absolute Error represents the average magnitude of the errors in a set of predictions without considering their direction. It is calculated in Equation (Khan, 2024):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i are the actual values, $\hat{y}_i$ are the predicted values, and n is the number of observations.

3. MATERIAL AND METHOD

3.1. Dataset

The dataset used in this study originates from an RNA-sequencing study on platelets collected from patients with tumors, referred to as Tumor-Educated Platelets (TEPs). This

dataset is publicly available under accession number GSE68086 in the Gene Expression Omnibus (GEO) database. The RNA-seq data were gathered from 283 blood platelet samples, including 228 samples from patients with six different types of cancer (non-small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer, and hepatobiliary carcinoma) and 55 samples from healthy individuals.

3.2. Data Characteristics

The GSE68086 dataset, used for cancer diagnostics through Tumor-Educated Platelets (TEPs), provides 57,736 rows representing gene IDs and 285 columns corresponding to blood platelet samples. The dataset spans multiple cancer types, including non–small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer, hepatobiliary carcinomas, and healthy donor samples. It consists of 283 labeled samples distributed across 15 categories, with the largest groups being Healthy Donor (HD) samples (45) and lung cancer samples (39).

This dataset contains no missing values; however, it includes 27,239 outliers distributed across all columns, which requires careful preprocessing. The range of expression values varies significantly between samples, with a minimum value of 0 and a maximum value of 455,636. The range of values for each variable is between 4.97 and 6.01 units, with minimum values ranging from 1.60 to 3.25 and maximum values between 7.05 and 8.31. This consistency is evident from the relatively minor differences in range between columns, with the lowest variation being 4.97 units and the highest 6.01 units.

These stable characteristics are very beneficial for the process of imputing missing data. The uniformity of the value range indicates that all variables are on a comparable scale, so that similarity-based methods such as KNN can work optimally without requiring complex standardisation. Furthermore, this consistency ensures that the relationship patterns between variables are sufficiently stable, so that the imputed values will be more accurate and will not disrupt the basic structure of the dataset. Such data conditions facilitate the selection of imputation methods because there are no significant scale imbalances between columns that could affect the analysis results.

3.3. Method

This study aims to evaluate the performance of mean imputation, K-Nearest Neighbours (KNN), and KNN optimised using Gaussian Process–based Bayesian optimisation in handling missing values introduced under the MCAR (Missing Completely at Random) mechanism at rates from 5% to 60% (in 5% increments). Evaluation metrics are Normalised Root Mean Square Error (NRMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). The step-by-step analysis procedure is described below:

(1) Import Required Libraries (Import Libraries): All Python libraries required for data manipulation, imputation, statistical analysis, and performance evaluation are imported at an early stage.
(2) Load the Dataset (Read Dataset): The dataset is loaded into the Python environment, and the structure and initial content of the data are examined.
(3) Descriptive Statistics (Check Descriptive Statistics): Calculate descriptive statistics such as mean, median, standard deviation, and range.
(4) Check for Missing Values and Outliers (Check Dataset for Missing Values and Outliers: Missing values are identified and outliers are detected using visualizations (e.g., boxplots).

(5) Handle Outliers (Identification and Handling of Outliers with IQR and Z-score): Outliers are treated using methods such as the interquartile range (IQR) or Z-score, either by capping extreme values or applying suitable transformations.
(6) Standardization and Log Transformation: A log transformation ($\log(x+1)$) is applied to skewed features, followed by standardization to ensure that all features are on a comparable scale.
(7) Save Cleaned Data (Data Backup – Preservation): The cleaned dataset is saved to provide a reliable version for subsequent missing value simulations.
(8) Simulate Missing Values Using MCAR (Insertion of Missing Values): Missing values are randomly introduced into the dataset under the MCAR assumption at levels ranging from 5% to 60%. This ensures that the missingness occurs completely at random, meaning that the probability of a value being missing is the same for all entries and is independent of both observed and unobserved data.
(9) Check for MCAR Suitability Using t-test (t-Test for MCAR Validation): A t-test is performed to validate whether the missing values are truly MCAR by comparing observed and missing data distributions.
(10) Data Imputation Stage: Missing values are imputed using three methods:
   (a) Mean imputation
   (b) K-Nearest Neighbors (KNN) imputation
   (c) KNN Bayesian imputation (KNN combined with Gaussian Process Bayesian optimization)
(11) Evaluate Imputation Performance: The performance of each imputation method is evaluated using the following metrics:
   (a) MAE (Mean Absolute Error)
   (b) NRMSE (Normalized Root Mean Square Error)
   (c) MSE (Mean Squared Error)
(12) Summarize and Interpret Results: Comparative results for each missing value level are recorded and interpreted.
(13) Conclusion (End): A conclusion is drawn to determine which imputation method is the most effective under different missing value conditions.

The flowchart of the comparative study of missing value imputation methods using Mean, Bayesian KNN, and Non-Bayesian KNN on TEP gene expression data is shown in Figure 1.

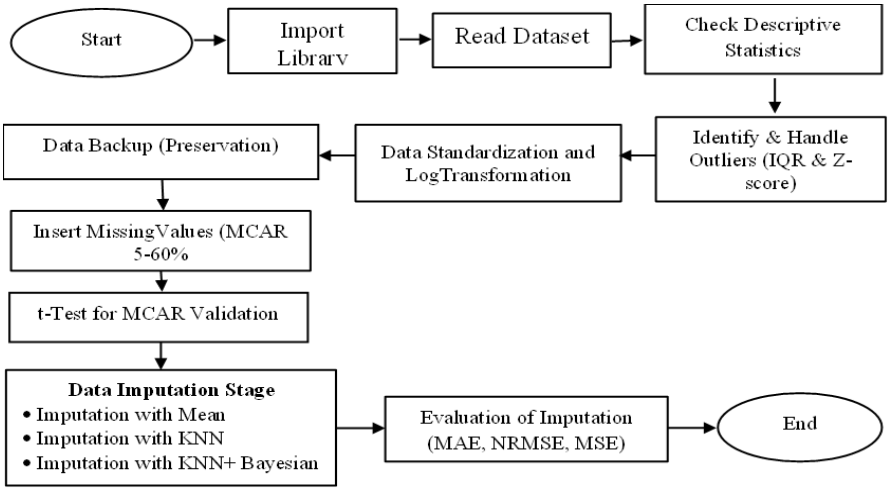

Figure 1. Workflow for Comparing Missing Value Imputation Methods: Mean, Bayesian KNN, and Non-Bayesian KNN

4. RESULTS AND DISCUSSION

The data set used in this study is the Tumor-Educated Platelet (TEP) gene expression data obtained from Gene Expression Omnibus (GEO) with access number GSE68086. This data set consists of 57,736 rows representing gene expression and 286 columns corresponding to individual sample identifications.

Preliminary analysis showed that the dataset did not contain missing values, but it did contain outliers that needed to be addressed, and the data was not normally distributed. Gene expression values also showed a wide range, with minimum values of 0 and maximum values reaching up to 455,636 in certain samples. The outlier detection process highlighted samples with expression values exceeding the interquartile range thresholds. Table 1 presents the number of outliers detected in each sample, illustrating the variability of gene expression across different cancer conditions.

Information on the distribution of cancer types is crucial for mapping biological variations in the dataset and for anticipating the potential influence of cancer types on gene expression patterns in subsequent analyses. By understanding the proportion and diversity of samples based on cancer type, the developed imputation models and methods can be adjusted to be more accurate and relevant to existing biological characteristics. Table 1 presents the number of samples based on the cancer types identified in the dataset.

Table 1. Detected Outliers in TEP Gene Expression Dataset

Sample ID	Outlier Count
3-Breast-Her2-ampl	71
8-Breast-WT	89
10-Breast-Her2-ampl	89
Breast-100	68
15-Breast-Her2-ampl	78
…	…
MGH-NSCLC-L40-TR520	92
MGH-NSCLC-L51-TR521	100
MGH-NSCLC-L58-TR525	138
MGH-NSCLC-L59-TR522	97
MGH-NSCLC-L65-TR523	112

The t-test results showed no significant difference between the missing data and the observed data ($p > 0.05$), supporting the assumption that the missing values follow the MCAR (Missing Completely at Random) mechanism. The evaluation metrics, including NRMSE, MAE, and MSE, for each imputation method across different missing value rates are summarized in Table 2. These metrics provide a comparative assessment of the performance of Mean Imputation, KNN, and Bayesian KNN under increasing levels of missing data.

From the application of the mean imputation model, KNN and Bayesian KNN to TEP data that still contains 0 values (not missing values), the results of the comparison of imputation methods are shown in Table 2 above. Performance is calculated from three metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and Normalised Mean Squared Error (NRMSE). The performance of each method was analysed at various levels of missing data, ranging from 5%, 10%, 15% to 60%.

The results of the imputation evaluation based on the MAE metric show that the average absolute error value ranges from 38.4 to 42.1. In general, the MAE value tends to

decrease as the percentage of missing values increases from 5% to 60%, indicating that all three imputation methods are able to maintain stability even though more data is missing. The differences between the methods are very small, only about 0.01–0.05 at each percentage level, so no single method is truly dominant in terms of MAE.

Table 2. Comparison of Imputation Performance Using MAE, MSE, and NRMSE for Mean, KNN, and Bayesian KNN

Rate of Missing Value	Imputation Method	MAE	NRMSE	MSE
5%	MEAN	42.08	0.0057	1501979.50
5%	KNN	42.124	0.0057	1501469.36
5%	KNN Bayesian	42.119	0.0057	1501474.88
10%	MEAN	39.55	0.0074	987727.52
10%	KNN	39.603	0.0074	987263.15
10%	KNN Bayesian	39.597	0.0074	987265.03
15%	MEAN	39.58	0.0060	1064261.69
15%	KNN	39.63	0.0060	1063791.57
15%	KNN Bayesian	39.62	0.0060	1063793.30
20%	MEAN	39.01	0.0050	1031574.53
20%	KNN	39.061	0.0050	1031114.61
20%	KNN Bayesian	39.056	0.0050	1031116.22
25 %	MEAN	38.51	0.0047	904602.87
25%	KNN	38.560	0.0047	904151.89
25%	KNN Bayesian	38.555	0.0047	904154. 23
30%	MEAN	38.90	0.0038	1004476.19
30%	KNN	38.95	0.0038	1004019.97
30%	KNN Bayesian	38.94	0.0038	1004022.28
35%	MEAN	39.29	0.0041	1129157.61
35%	KNN	39.342	0.0041	1128695.93
35%	KNN Bayesian	39.337	0.00405	1128697.79
40%	MEAN	38.86	0.0047	962406.16
40%	KNN	38.910	0.0047	961951.20
40 %	KNN Bayesian	38.905	0.0047	961953.98
45%	MEAN	39.67	0.0030	1187522.44
45%	KNN	39.72	0.0030	1187057.08
45%	KNN Bayesian	39.71	0.0030	1187060.35
50%	MEAN	38.46	0.0037	961439.84
50%	KNN	38.515	0.0037	960989.54
50%	KNN Bayesian	38.509	0.0037	960991.85
55%	MEAN	39.07	0.0039	1053131.25
55%	KNN	39.13	0.0039	1052673.06
55%	KNN Bayesian	39.12	0.0039	1052675.91
60%	MEAN	38.89	0.0050	1013082.62
60%	KNN	38.939	0.0050	1012630.64
60%	KNN Bayesian	38.934	0.00498	1012633.42

For the NRMSE metric, the values obtained are relatively very small, ranging from 0.0030 to 0.0074. The highest value was at 10% missing values (0.0074) and the lowest at 45% missing values (0.0030). The difference in values between methods was again almost identical, differing only by three to four decimal places. This shows that MEAN, KNN, and Bayesian KNN produce almost the same level of error in the data after the imputation process.

When viewed from the MSE, the values are on a fairly large scale, around 9.0×10^5 to 1.5×10^6. The pattern formed is also unclear, with no consistent upward or downward trend in the percentage of missing data. However, the differences between the methods remain very small. For example, at 25% missing values, the difference in MSE between MEAN, KNN, and Bayesian KNN is only about 450 points out of a total of more than 900 thousand, so it can be said that the three methods produce almost identical performance.

Overall, the three imputation methods used, namely MEAN, KNN, and Bayesian KNN, show very similar performance at all levels of missing values. Although in theory MEAN is usually considered simpler and prone to bias, in this TEP dataset the results are not inferior to KNN-based methods. Meanwhile, Bayesian KNN, which is expected to provide additional optimisation, does not provide a significant improvement over regular KNN. Thus, the selection of imputation methods in this TEP data can consider computational efficiency factors, as the accuracy of the three methods is relatively comparable.

Most studies on missing value imputation have focused on general gene expression datasets, whereas Tumor-Educated Platelets (TEP) data possess distinct biological and statistical characteristics that may influence imputation behavior. Despite the increasing use of TEP in cancer diagnostics and biomarker discovery, there has been limited investigation into how common imputation techniques perform on such data. Moreover, comparative evaluations between conventional approaches (such as mean and KNN) and their optimized variants (like Bayesian KNN) on TEP datasets remain scarce. This study addresses this gap by systematically evaluating and comparing these methods on TEP gene expression data, aiming to determine whether method complexity provides a tangible benefit or if simpler approaches are sufficient for accurate imputation in TEP-based analyses.

5. CONCLUSION

The results of this study indicate that the three imputation methods, namely Mean, KNN, and Bayesian KNN with optimisation using the Gaussian process, produce relatively similar performance in filling in missing values. The MAE, MSE, and NRMSE values obtained from the three methods are within a very close range, with differences only in small decimal places, so that no method is consistently superior. These findings show that although Bayesian Optimisation is expected to improve the accuracy of KNN, the improvement is not significant in this dataset. To strengthen the generalisation, future research could use TEP data with more in-depth pre-processing, for example by removing zero values in the original dataset, handling outliers correctly, and performing standardisation with log transformation correctly so that the MAE value is not too high and the MSE and NRMSE values have a visible distance to clearly compare the results. Additionally, if these steps do not produce the expected results, the Bayesian KNN method with Gaussian Process optimisation can be applied to datasets with other characteristics, which may highlight the benefits of optimisation in handling missing values.

ACKNOWLEDGMENT

This research was supported by BrainAI Lab (Bioinformatics Research, Data Intelligence, and AI Innovation Laboratory), Universitas Indonesia. The study was funded through a research grant provided by Universitas Indonesia under the Hibah Autis scheme, Contract No. PKS-507/UN2.RST/HKP.05.00/2025, issued by the Directorate of Research and Community Engagement, Ministry of Education, Science, and Technology. The availability

of the laboratory's computational resources was essential for data processing and analysis, which significantly contributed to the completion of this study.

REFERENCES

Al-Janabi, S., & Alkaim, A.F. (2020). A Nifty Collaborative Analysis to Predicting a Novel Tool (DRFLLS) for Missing Values Estimation. *Soft Computing*, *24*(1),555–569. https://doi.org/10.1007/s00500-019-03972-x

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of Missing Data on Bias and Precision When Estimating Change in Patient-Reported Outcomes from a Clinical Registry. *Health and Quality of Life Outcomes*, *17*(1), 1–9.

Brown, S. et al. (2018). Technical Variability and Missing Data in Gene Expression Studies. *Bioinformatics*, *34*(22), 3808–3815.

Chungnoy, K., Tanantong, T., & Songmuang, P. (2024). Missing Value Imputation on Gene Expression Data Using Bee-Based Algorithm to Improve Classification Performance. *PLoS ONE*, *19*(8), e0305492. https://doi.org/10.1371/journal.pone.0305492

Farswan, A., Gupta, A., Gupta, R., & Kaur, G. (2020). Imputation of Gene Expression Data in Blood Cancer and its Significance in Inferring Biological Pathways. *Frontiers in Oncology*, *9*, 1442. https://doi.org/10.3389/fonc.2019.01442

Hameed, W. M., & Ali, N. A. (2023). Missing Value Imputation Techniques: A Survey. *UHD Journal of Science and Technology*, *7*(1), 72–81. https://doi.org/10.21928/uhdjst.v7n1y2023.pp72-81

Hong, S., & Lynn, H. S. (2020). Accuracy of Random-Forest-Based Imputation of Missing Data in The Presence of Non-Normality, Non-Linearity, and Interaction. *BMC Medical Research Methodology*, *20*(1), 1–12.

Injadat, M. N., Salo, F., Bou Nassif, A., Essex, A., & Shami, A. (2020). Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection. *arXiv preprint* arXiv:2008.02327v1. https://arxiv.org/abs/2008.02327

Ismail, A. R., Zainal Abidin, N., & Maen, M. K. (2022). Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare. *Journal of Robotics and Control (JRC)*, *3*(2), 143-150. https://doi.org/10.18196/jrc.v3i2.13133

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, *33*(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Jafrasteh, B., Hernández-Lobato, D., Lubián-López, S. P., & Benavente-Fernández, I. (2023). Gaussian Processes for Missing Value Imputation. *Knowledge-Based Systems*, *273*, 110603. https://doi.org/10.1016/j.knosys.2023.110603

Keerin, P., & Boongoen, T. (2022). Improved KNN Imputation for Missing Values in Gene Expression Data. *Computers, Materials & Continua*, *70*(2), 4009-4025. https://doi.org/10.32604/cmc.2022.020261

Khan, M. A. (2024). A Comparative Study on Imputation Techniques: Introducing A Transformer Model for Robust and Efficient Handling of Missing EEG Amplitude Data. *Bioengineering*, *11*(8), 740. https://doi.org/10.3390/bioengineering11080740

Latief, M. A., Bustamam, A., & Siswantining, T. (2020). Performance Evaluation XGBoost in Handling Missing Value on Classification of Hepatocellular Carcinoma Gene Expression Data. *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, 1–6. IEEE. https://doi.org/10.1109/ICICoS51170.2020.9299009

Lee, K. & Tung, C. (2019). Missing Value Imputation for Microarray Data: A Comprehensive Review. *Journal of Bioinformatics and Computational Biology*, *17*(3), 1950023

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data (3rd ed.)*. Wiley Series in Probability and Statistics. Print ISBN: 9780470526798 | Online ISBN: 9781119482260 | https://doi.org/10.1002/9781119482260

Liu, L., Lin, F., Ma, X., Chen, Z., & Yu, J. (2020). Tumor-educated Platelet as Liquid Biopsy in Lung Cancer Patients. *Cancer Letters*, *146*, Article 102863. https://doi.org/10.1016/j.canlet.2020.102863

Lo, A. W., Siah, K. W., & Wong, C. H. (2019). Machine Learning with Statistical Imputation for Predicting Drug Approval. *Harvard Data Science Review*, *2019.*

Miller, K. D., Ortiz, A. P., Pinheiro, P. S., Bandi, P., Minihan, A., Fuchs, H. E., Martinez Tyson, D., Tortolero-Luna, G., Fedewa, S. A., & Jemal, A. M. (2021). Cancer Statistics for the US Hispanic/Latino Population. *CA: A Cancer Journal for Clinicians*, *71*(6), 466–487. https://doi.org/10.3322/caac.21695

Mostafa, S. M. (2019). Imputing Missing Values Using Cumulative Linear Regression. *CAAI Transactions on Intelligent Technology*, *4*(3), 182–200.

Ravindran, U., & Gunavathi, C. (2023). A Survey on Gene Expression Data Analysis Using Deep Learning Methods for Cancer Diagnosis. *Progress in Biophysics and Molecular Biology*, *177*, 1–13. https://doi.org/10.1016/j.pbiomolbio.2022.08.004

Siswantining, T., Anwar, T., Sarwinda, D., & Al-Ash, H. S. (2021). A Novel Centroid Initialization in Missing Value Imputation Towards Mixed Datasets. *Communications in Mathematical Biology and Neuroscience*, *2021*(11). https://doi.org/10.28919/cmbn/5344

Siswantining, T., Vivaldi, K. G., Sarwinda, D., Soemartojo, S. M., Sari, I. M., & Al-Ash, H. S. (2022). Implementation of Ensemble Self-Organizing Maps for Missing Values Imputation. *Indonesian Journal of Statistics and Its Applications*, *6*(1), 1–12. https://doi.org/10.29244/ijsa.v6i1p1-12

Zhu, X., Wang, J., Sun, B., Ren, C., Yang, T., & Ding, J. (2021). An Efficient Ensemble Method for Missing Value Imputation in Microarray Gene Expression Data. *BMC Bioinformatics*, *22*, 188. https://doi.org/10.1186/s12859-021-04109-4