International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I9P229

Volume 69 Issue 9, 242-250, September, 2021
©2021 Seventh Sense Research Group®

Review Article

Data Mining Based Imputation Techniques to Handle Missing Values in Gene Expressed Dataset

Amarjeet Yadav[1], Aditya Dubey[2], Akhtar Rasool[3], Nilay Khare[4]

Department of Computer Science & Engineering, Maulana Azad National Institute of Technology, Bhopal, India

[1] yamarjeet175@gmail.com, [2] dubeyaditya65@gmail.com , [3] akki262@gmail.com,[4] nilay.khare@gmail.com

Abstract *- The microarray analysis results in datasets with massive expression levels of genes as rows and following the various laboratory conditions as columns. Due to experimental errors, these datasets frequently have some content dropping. The presence of missing values in data sets significantly reduces efficiency and accuracy. It can influence the outcome of the visualization study of gene representation. Therefore, how to predict missing records indeed becomes significant to examine the elementary arrangement. Missing data imputation has received numerous attractions from researchers. This paper summarizes most of the techniques proposed for the imputation of missing data. It contains a thorough discussion about various advantages and disadvantages of global, local, and hybrid approaches and knowledge-assisted approaches. This paper has described MCAR, MNAR, MAR techniques to identify the type of missing data. Precisely this article compares all the methods and puts forward a better understanding of these techniques.*

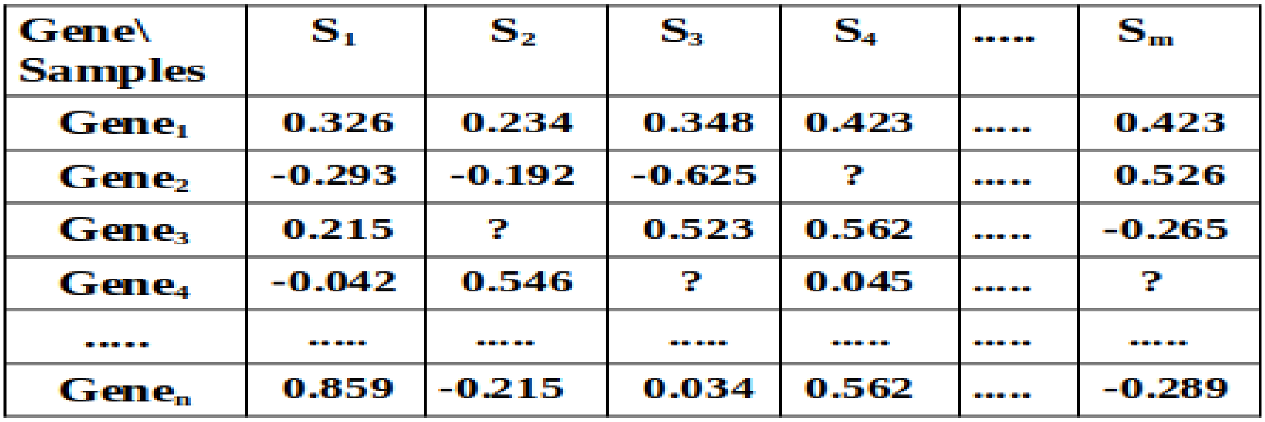
Keywords — *Correlation Structure, Gene Expression Data, Imputation, Missing Value.*

I. INTRODUCTION

In the real-world application dataset, Missing data is a common problem [1]. A more comprehensive datasets class will tolerate the problem that many data entries in the dataset are missing. The problem of missing data happens when no information is stored for the variable in the observation. This condition primarily arises in the manual data entry procedures, apparatus errors, operator failure, and erroneous measurements [2]. Due to various reasons, gene expression data might contain missing values, such as inadequate resolution, image degradation, or dirt or scratches on the slide. Since it is very time-consuming and expensive to iterate data collection, researchers are now working on the missing data imputation technique as a solution. No data means that not sufficient information was available for that field to assign it a value. In datasets, missing data could be represented as '?', 'nan,'' N/A,' blank cell, or sometimes '-9999',' inf,' '-info. It is a prerequisite to understanding the concept of missing values to manage missing data successfully. If the researcher does not correctly control the missing values, they can draw wrong conclusions about the data. Due to improper handling, the results obtained by the researcher will differ from those where missing values exist. Missing data creates different problems.

Missing data scale down the statistical power, which indicates the possibility that the null hypothesis will be rejected in the test when it is false. Also, missing data can cause unfairness in the computation of parameters, which may reduce the representativeness of samples, altering the analysis of studies. Each of these misjudgments can hazard the effectiveness of the test and lead to a worthless outcome. The best way to handle missing values is to avoid the problem by planning the study appropriately and correctly accumulating them. Moreover, this section discuss some convolution techniques to handle the missing value [3]. List-wise deletion is the most common method for missing data to discard those missing value cases and analyze the remaining data. If the dataset comes under the category of missing completely at random (MCAR), then a listed deletion is acknowledged to assemble unbiased conclusions and moderate outcomes.

When the data does not meet the hypothesis of MCAR, list-wise deletion can cause bias in estimates of parameters. Moreover, in Pairwise deletion, missing observations are ignored and analyzed on present variables. Furthermore, Pairwise deletion maintains more information than list-wise deletion, which can delete the case with any missing value. And Mean, Median, and mode is the most common imputed approach. This imputation technique aims to substitute missing values with the arithmetical valuation of missing data for the same variable. Moreover, by making a histogram, considering the dataset distribution and a conclusion can be made regarding mean median or mode. Depending on the nature of knowledge utilized in the methods and classify existent methodology into four distinct categories: (i) Global technique, (ii) Local technique, (iii) Hybrid technique, and (iv) Knowledge assisted technique [4]. And figure 1 represents the gene expression dataset with missing values. Here, the '?' symbol signifies missing entries in the dataset due to various experimental reasons. In this figure, S_1, S_2,…, S_m represents samples of gene expression. The causes of the missing data are present in the gene expression data set can be loss of information and various experimental reasons.



Gene\ Samples	S_1	S_2	S_3	S_4	…..	S_m
$Gene_1$	0.326	0.234	0.348	0.423	…..	0.423
$Gene_2$	-0.293	-0.192	-0.625	?	…..	0.526
$Gene_3$	0.215	?	0.523	0.562	…..	-0.265
$Gene_4$	-0.042	0.546	?	0.045	…..	?
…..	…..	…..	…..	…..	…..	…..
$Gene_n$	0.859	-0.215	0.034	0.562	…..	-0.289

Fig. 1: Example of the gene expressed dataset having missing values.

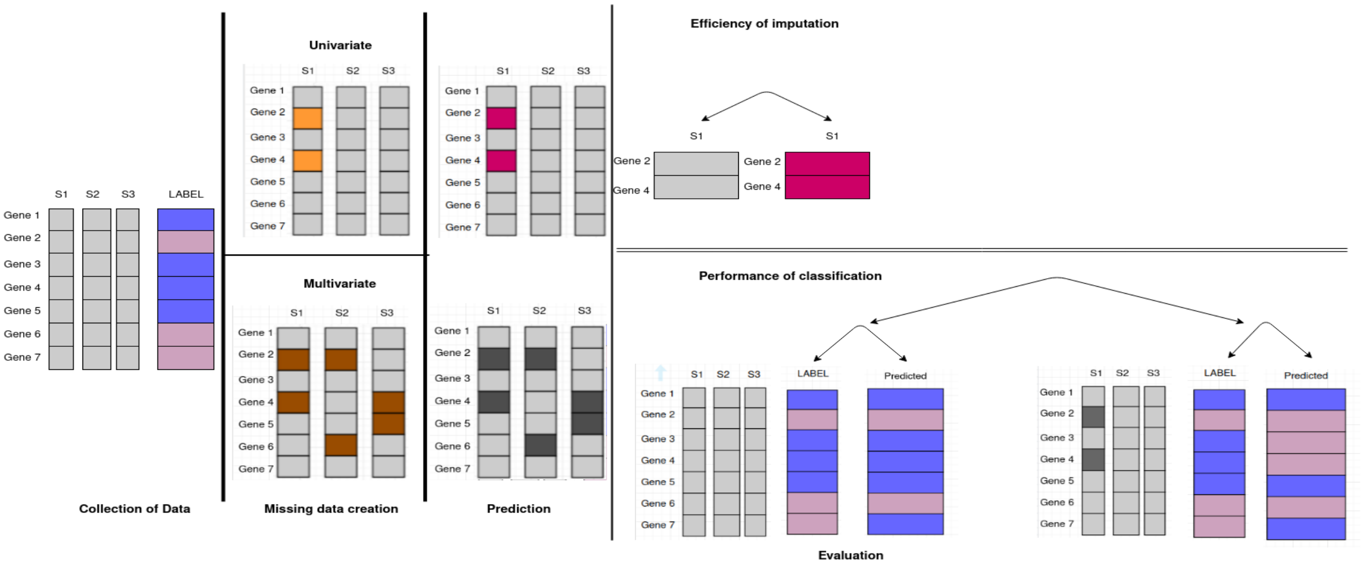


Fig. 2. Missing Data Imputation Procedure.

Figure 2 represents the procedure followed in Missing Data Imputation. S1, S2, S3 symbols are used to represent three samples of gene expressions. Univariate here indicates that there is one sample associated with the missing data model, whereas Multivariate indicates the presence of more than one sample linked with a missing data pattern. The prediction column shows predicted missing values using different techniques. The evaluation part shows the accuracy of imputation in a Univariate and Performance classification in Multivariate. Ensemble learning combines several single imputation approaches into a single imputation technique [5]. Each component method's estimates of missing values are weighted and averaged to form the final forecast in the ensemble approach, which uses bootstrap sampling. The best weights are determined by minimizing a cost function associated with the imputation error using known gene data. The optimal weights are also expressed in closed form. In addition, the ensemble method's performance is evaluated analytically in terms of the sum of squared regression errors. This method is best suitable for gene expression data in terms of efficiency and robustness.

In this paper, a survey of various imputation techniques has been done. The significant contributions of this research are enumerated as follows: Section (II) describes the different missing data mechanisms and explains each mechanism with examples. Section (III) elaborates on

Global approaches of missing data imputation. Global methods utilize the global correlation structure of data. This section discusses various global information-based techniques such as SVDimpute and BPCA, including their advantages and limitations. Section (IV) describes the local information-based methods. These approaches deal with the local structure of the data. KNNimpute, GMCimpute, LLSimpute, etc., are some of the local techniques. This section widely compares the efficiency and accuracy of all these methods. Further, in section (V), Hybrid based approaches have been compared. These approaches use the strength of both local and global based information. Some of these approaches are RMI, HPM_MI, etc. Analysis of all techniques has been done, including their shortcomings. Knowledge-assisted approaches utilize domain knowledge for data imputations. POCSimpute, GOimpute, and HAIimpute are some approaches addressed in section (VI).

II. BACKGROUND

Microarray technology has been one of the most valuable means for analyzing gene expression data. Researchers have used gene expression datasets extensively for various biological studies, such as examining the mechanism of drug response, cancer analysis, and classifying genes associated with an appropriate diagnosis. Before reviewing several different missing value handling techniques, it is crucial to look at the missing data mechanism briefly. The missing data mechanism establishes an interrelation between missing data and the mutable values in the data matrix. Rubin (1976) and cohort (Little & Rubin, 2002) classified missing data problems into three different categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). These classes are essential because missing data cause problems, and the solutions to these problems are distinct for the three classes. While these words have an absolute probabilistic and analytical definition, this section gives a theoretical explanation of each mechanism.

MCAR is defined as the case when the likelihood of missing data on some variable taken as A is irrelative to the other evaluated value of the variable B and the value of itself [6]. The various examples include a survey that may be lost in the post, a measurement scale that ran out of battery, an IQ score that cannot predict the age, or blood specimens in the laboratory which may be contaminated. Data may be termed as MCAR When choice becomes transit or technically inadequate and equipment's disappearance is due to machinery failure. MCAR is a perfect but improper presumption. The significant advantage of MCAR data is that the imputation remains fair. If missingness is irrelative to the missing value but linked to other variables' value, data are considered as MAR. And MCAR is less extensive than MAR [7].

For instance, the weight scale can generate higher missing costs when situated on a soft shallow than on a hardcover. This information which is in the visible form, does not come under the category of MCAR. If the Observer is familiar with classifying surface or shallow and speculates MCAR within the body type, the data are MAR. One more case of MAR is when they take a sample from a population. The prospect of joining relies on several familiar properties. For example, suppose the child does not appear in the experiment because it suffers from pain. In that case, it may be expressed as an outcome of another attribute like a child's biological data. But it would not be linked to what the Observer would have questioned had the kid not been suffering from pain. Some may consider that MAR does not exhibit a problem, although it does not imply that MAR neglects the missing data. MAR is more prevalent and more realistic than MCAR. Current missing data imputation approaches typically begin with the MAR presumption.

MNAR is also known as a Not Missing at Random (NMAR). If MCAR and MAR both do not hold, it will be categorized as MNAR. For example, a player does not appear in an anti-doping test because they took dope before the game. Another example could be a person who does not take a Mathematics skill test because he lacks mathematical skills. Another case of MNAR in a social point of view analysis is when people with weak knowledge do not react as much. MNAR is the more complicated case. The procedure to operate MNAR is to obtain more data about missing or see how sensitive the results are under different products [7]. It analyzes what-if to see. Case studies of MNAR data are questionable. The unique option of obtaining an unbiased result of the parameters is replicating the missing value in such a situation. Nevertheless, for this, a precise judgment of the missing variables and domain knowledge is required.

III. GLOBAL APPROACH

In these classes, algorithms demonstrate missing data Imputation based on global correlation information, and missing value is acquired from the entire data matrix. Moreover, it reflects a global covariance composition amidst every Genes or sample in the representation matrix. When this assumption is inappropriate, it means genes Demonstrate major local similarity structures, modeling them less accurately. In this category, the widely used methods include Singular Value Decomposition Imputation (SVDimpute) and Bayesian Principal Component Analysis (BPCA). The SVDimpute algorithm aims to predict the missing data as a linear sequence of the k-most important eigengenes [8]. An ideal linear combination is established by regressing the incomplete variable against the k-most similar eigengenes. When determining the regression coefficients, if the value at position i is missing, the eigengenes' i^{th} value is a linear combination of some significant axis vectors, where the parameters were labeled by the Bayesian prediction technique [9]. And one more example of the global approach is two gene expression cloning techniques that support the regeneration of a high-dimensional molecular characteristic essential for disease biology and drug target research [10]. Pseudo-Mask Imputer (PMI) and Generative Adversarial Imputation Nets (GAIN)- Genotype-Tissue Expression (GTEx) were fitted to impute absent expression values and estimate gene expression manifolds from incomplete gene expression values and associated covariates (latent global determinants of expression).

IV. LOCAL APPROACH

This local approach methodology satisfies the local correlation arrangement in the data sets to conduct missing information prediction. The genes that are part of a more extensive collection demonstrate an immense correlation with the missing data of genes to estimate the missing value. Examples of local approaches like K Nearest Neighbor Imputation (KNNimpute) and Local Least Squares Imputation (LLSimpute) are primitive and popular methods [11, 12]. The KNNimpute process predicts the missing values using similarity between the actual gene with lost data and the *k* nearest reference genes. Analysis has demonstrated that KNNimpute operates adequately whenever a strong relationship is found within the data in the gene's dataset. Local least squares imputation algorithms are a bit more ambitious than KNNimpute and more complicated than BPCA. The LLSimpute methodology employs a compound regression model to predict the missing data. The significant difference between LLSimpute and BPCA is that BPCA is a boosting algorithm, i.e., based on PCs, while LLSimpute is an improved method based on a similar local structure. The LLSimpute accomplishes advancement over KNNimpute by integrating the least squares. And BPCA attains a renovation over SVDimpute by assimilating Bayesian rectify.

Research has proved that KNNimpute is stronger and more precise than SVDimpute. SVDimpute owns some vulnerabilities and depends on complete genes and examination in the dataset. In addition, it does not acknowledge local structure. Other than this, an understandable interpretation model may not exist for non-time series data. The articulation model for the subset of genes cannot be expressed better by major eigengenes for noise data. It is attainable to predict that LLSimpute can determine immensely complex efficiency based on scrutiny, Accommodating in the experiments [12]. Sequential Local Least Square Imputation (SLLsimpute) technique is the extended version of the LLSimpute. This methodology gives better performance consecutively by initiating from the gene with less missing rate. The predicted genes are reprocessed in SLLsimpute; because of SLLsimpute techniques' efficiency, it is better to compare LLSimpute. And one more algorithm comes under the category of local approach, i.e., MICE-CART, and it stands for Multiple Imputations by Chained Equations (MICE) and Classification and Regression Tree (CART) [13].

Various imputations are a better technique to shorten the missing value problem in the broader data analysis range. A few of the prediction analyses may require complicated modeling along with interactions and unpredictable correlation. MICE-CART is used for parameter custom depletion, and this approach can perform to secure optimum performance while carrying off intricate interconnection of data. Gaussian Mixture Clustering imputation (GMCimpute) approach can use the more global similarity knowledge. However, this is a local approach; this methodology clustered the value into S elements Gaussian mixture applied Expectation-Maximization (EM) method, where S depicts a measure of missing data. Each component's single S value is calculated and then equalized to get the concluding figured missing data [14]. One more example of the local approach is Collateral Missing Value Imputation (CMVE). This methodology uses the process of calculating various analogies of missing data to enhance the outcomes [15]. CMVE yields a better efficacy in Normalized Root Mean Square Error (NRMSE) in comparing KNN, LSimpute, and BPCA on different datasets like ovarian cancer and time-series data like yeast sporulation. Another technique, Ameliorative Missing Value Imputation (AMVI), is more robust in comparison to CMVE. For the most optimistic estimation of reference genes, this method uses the Monte Carlo simulation approach, making AMVI more vital than CMVE. In AMVI, the time-series gene expression outlines have been utilized to express the powerful dependency on outcomes [16].

Another way to handle the missing value, the Doubly Sparse DCT domain with Nuclear Norm Minimization (DSNN) method, uses the dual sparsity concept based on discrete cosine transform and nuclear norm minimization to predict missing values in the data [17]. This model has two stages. In the first stage, to impute missing values, row and column sparsity is used, whereas in the second stage, for removing noise, a low-rank kind of pattern is utilized. The suggested DSNN method outperforms the other matrix completion techniques at each sampling proportion. For optimal representation, various imputation methods required more parameter tuning. In contrast, the DSNN method did not expect a wide scale of parameter tuning. Local information-based approaches have proven to give better outcomes in missing value imputation compared to global information-based processes. The reason behind the more reliable performance is that this method uses the local information of the data. Most of the local information-based methods proposed till now have experienced diminished performance due to overfitting. This paper utilized a technique that uses regularization procedures to resolve to overfit [18]. It employs to build the connections within a target gene and its acquaintances using a regularized sparse structure for imputing missing data. It introduces the RLLSimpute_EN technique based on local least square estimation. According to the missing instance frequencies, this approach analyses the objective genes. Then manages them in order from the least to the maximum missing rate to utilize earlier calculated values. This paper introduces four more techniques with unique regularization specifications, namely fLLSimpute, fLLSimpute_L1, fLLSimpute_L2, and fLLSimpute_EN. This approach proved to be robust and gives higher accuracies in comparison to other proposed techniques.

Another approach for solving missing data imputation is based on Least Absolute Shrinkage and Selection Operator, abbreviated as (LASSO). This technique, also known as SampleLASSO, dynamically trains a machine learning model using a scattered regression method on each expression set [19]. This shows that SampleLASSO is a high precision method based on a comprehensive evaluation of three distinct prediction tasks (intra-technology and cross-technology), two imputation setups, and a multi-gene expression platform to predict unmeasured genes. The advantage of SampleLASSO is that it can efficiently use sample information from the same biological environment. Moreover, helping to measure the performance of the prediction technique, evaluation in distinct prediction environments focuses on some data standardization structures. For the estimation of microarray data, there is a requirement of normalization of training and testing data to a similar distribution.

The small capture value of RNAs expressed by single-cell sequencing methodology is a significant complication to downstream functional genomics research [20]. Currently, several prediction techniques have come out for single-cell transcriptome value; still, working on sizable sparse expression matrices becomes a significant challenge while imputing missing data. The Weighted Decomposition of Gene Expression (WEDGE) technique is utilized as a biased low-rank matrix disintegration method to predict gene expression metrics. The WEDGE method favorably retrieved the expression matrix, recovered cell-wise and gene-wise relationships, and executed more accurate clustering of cells, with exceptional performance pertinence with scattered datasets. Another method known as Variational AutoEncoder (VAE) is a type of probabilistic encoder which is basically used as a generative tool in the domain of images and texts on a large scale [21].VAE is a deep learning framework which is used the distribution of latent space variables that implements a model, i.e., produced an outcome similar to the input data. VAE has the capability to identify the fake contents in images, texts, or sound signals, and it produces a better accuracy in the comparison of other most widely used technology of missing data imputation.

V. HYBRID APPROACH

Local relationship between Genes preponderates and algorithms like KNNimpute or the like LLSimpute carry out better outcomes in comparing BPCA or SVDimpute for heterogeneous data sets. This demonstrates that the interrelation structure in the Data imputation altered the efficacy of prediction methods. In the sense of global correlation, a technique like BPCA or SVDimpute is the most favorable method. LinCmb is a technique that comes under the category of a hybrid approach. It implements both global and local association structure knowledge in the datasets [25]. By utilizing this technique, the missing data are predicted by the consolidation of five distinct imputation techniques, particularly row average, SVDimpute, GMCimpute, KNNimpute, and BPCA. LinCmb produces false missing entries in places where correct Values are recognized and applies component methods to predict missing cell location values. It implements both global and local association structure knowledge in the datasets. This approach is flexible to the data matrix's association structure when higher missing records exist. There will be global methods, the center of attraction on setting missing values. Some hybrid approaches use the combination of some best imputation techniques to improve the data quality significantly.

Hybrid Prediction Model with Missing value Imputation (HPM-MI) is one of the hybrid approaches proposed after qualitative analysis of eleven imputation techniques [26]. This approach blends the clustering technique with Multilayer Perceptron. K-means clustering approach is applied to cluster the results. Adding to its Genetic Algorithm (GA)+ Support Vector Regression (SVR) is another unique approach. This approach uses a genetic algorithm to optimize properties choices and predict the decision attributes using SVR [27]. Further, it relevantly decreases the error between the model prediction and the given input. Another critical approach is Fuzzy C-means+SVR+GA. This approach uses no theoretical motivation for choosing clusters. KNN + Neural Network (NN) is promoted to subdue incompetent input value [28]. Earlier approaches considered data as inputs and based on that predicted output class. This approach does not work when one or more data are missing, making it inappropriate for choice-making when data variables are not present. Appending more approaches such as Fuzzy C-means + GA. These approaches are based on inductance loop indicator results. Another approach could be representing vector data into the matrix-based data. Further, to optimize the centroids in the Fuzzy C-means model, a genetic algorithm is used.

The K-Nearest Neighbor Graph (KNN-G) is commonly utilized to infer the relationship between cell IDs and cells and is the emphasis on widely used dimension reduction and projection techniques [29]. KNN-G is also the basis for the alternative technique that uses adjacent averaging and graph diffusion, for example. Connect each cell to a nearby k-cell based on the distance between the gene profiles and the KNN-G. Denoising Expression data with a Weighted Affinity Kernel and Self-Supervision (DEWAKSS) uses a unique mapping technique to adjust parameters. These are powerful in pre-processing methods that use benchmark data with established denoising to the optimal parameter selected by the objective function and use cell identity isolation and dimension reduction methods. It shows that it maintains a strong cluster and maintains a variance along multiple dimensions of representation. Contrary to earlier heuristic-based approaches, which lead to over-smooth the data distribution, it includes small diffusion and instead practices a fixed-weight KNN-G for denoising. The output of denoising models of gene expression data varies greatly depending on the selection of parameter values, and some methods need a suitable noise model.

VI. KNOWLEDGE ASSISTED APPROACH

The missing value imputation approach combines the field information. This method uses domain knowledge to improve the imputation's accuracy, which makes this approach powerful. This method turned out to be better than a data-driven approach, especially when dealing with datasets with a small number of high missing rates. This technique involves the union of external knowledge. Various algorithms under this approach utilize information about the biological mechanism in the microarray analysis, learning of the underlying biomolecular process, information concerning spot property in the microarray experiment, and knowledge from various external data sets. Projection Onto Convex Set (POCS) is one example of this approach, a flexible structure that utilizes the natural appearance of simultaneity loss and similarity data between genes and arrays [34]. This approach uses the regression technique of local least squares, which efficiently captures similarity gene-wise. It further uses Principal Component Analysis (PCA) imputation to apprehend array-wise similarity. To charge simultaneity loss, this approach limits the squared power of the expression outline. This can achieve an optimal solution using the POCS approach despite the similarity structure present in the data. The most minor but most substantial constraints give better accuracy than more extensive but less significant restrictions. The working of genes is expected to be shown in a modular fashion, displaying huge responses to purposes.

One of the well-accepted models for gene behavior categorization is Gene ontology, and it describes gene outcomes in terms of similar Biological Processes, Cellular Components, and Molecular Functions [35]. This approach improved the imputation accuracy when the balance of interpreted genes was large regarding high frequencies of missing value. Histone Acetylation Information-aided Imputation (HAIimpute) uses techniques like KNNimpute and LLSimpute to enhance the efficiency of missing content calculation [37]. HAIimpute applies the mean depiction of genes from every cluster to develop the pattern expressions. This approach practices a linear regression model to capture missing values. Model is executed with the gene and pattern appearances to determine the missing value. Unification of linear regression technique and a subsequent approach using both KNNimpute and LLSimpute provides the concluding approximations of missing values. HAIimpute has enhanced the KNNimpute or LLSimpute, which could be observed from the improved association in imputed genes and complete original genes.

TABLE 1. MERITS AND LIMITATIONS OF EXAMPLES OF FOUR APPROACHES

Sr. No.	Techniques	Types	Merits	Limitations
1.	SVDimpute [8]	Global	SVDimpute gives the most reliable outcomes on time-series datasets where low noise levels are present.	It does not give the best result on non-time series datasets.
2.	BPCA [9]	Global	The estimation error in BPCA is small and suggests that the bias initiated by the BPCA is smaller than those of existing methods.	If the gene has local similarity structures in a given dataset, then the BPCA technique may not be accurate.
3.	PMI and GAIN-GTEx [10]	Global	GAIN-GTEx approach performs better than various in-place imputation techniques. In inductive imputation techniques, the PMI algorithm gives excellent performance. These algorithms are highly relevant for different levels of missingness.	GAIN-GTEx penalizes the regeneration error of the examined components, and it is suffering from an underfitting problem.
4.	KNNimpute [11]	Local	It gives a better outcome when the number of samples is small in quantity by applying local similarity.	It does not perform better results on the large dataset.
5.	LLSimpute [12]	Local	LLSimpute is more favorable for an enormous value of k (number of the nearest neighbors).	The performance of LLSimpute evolves abysmal when *k* (number of the nearest neighbors) is close to the number of samples.

6.	MICE-CART [13]	Local	It has the ability to apprehend complicated relationships with the least accommodations by a data imputation methodology.	The pair of CART-based and classic MICE outcomes in multiple interruptions does not wrap the corresponding validity, as they depend on inadequate glitch imitation.
7.	GMCimpute [14]	Local	It is more robust because it is proficient in utilizing additional global association knowledge.	It is suffering from a slower fitting problem.
8.	CMVE [15]	Local	CMVE algorithm gives a better outcome when the missing rate is higher in both time series and non-time series datasets.	CMVE does not naturally calculate the most favorable number of concluding genes *k* from the dataset.
9.	Tailored nearest neighbors [22]	Local	This method performs well even though the sample size is small and gives better accuracy in comparing random forest techniques.	Imputation of weighted nearest neighbors requires tuning parameters window width (λ) and selected distance (m).
10.	SLLSimpute [23]	Local	SLLSimpute gives better results in comparing LLSimpute by using the genes with missing data again in this technique.	This algorithm is more favorable only when the slightest missing rates.
11.	Locally Auto-weighted Least Squares Method (LAW-LSimpute) [24]	Local	It is more robust because it optimizes the convergence and is capable of lowering the estimation error.	When the missing rate is high, then this technique is least preferred.
12.	HPM-MI [26]	Hybrid	HPM-MI gives a better outcome in terms of precision, selectivity, and sensitivity.	In the HPM-MI technique, the problem arises due to multi-class imbalanced classification problems.
13.	GA+SVR [27]	Hybrid	This model needs less computational time, and the SVR clustering method gives a more practical outcome.	This imputation technique suffers from a local minimization problem. And this method fails for some outlier data.
14.	KNN+NN [28]	Hybrid	This technique shows better evaluation accuracy than other NN-GA approaches because of noise mitigation methods, improving the computation’s accuracy. Many noise reduction techniques are only suitable for KNN based approaches.	Some criteria have to be decided beforehand, such as choosing the type of neural network and hyperparameters to train the model to meet the performance standards. And Computation time is very high.
15.	Recursive Mutual Imputation (RMI)[30]	Hybrid	This technique is robust and adequate to impute missing values. And It gives a better result for a large dataset.	In the RMI technique, it isn't easy to select a single suitable method in RMI.
16.	Fuzzy c-means [31]	Hybrid	Fuzzy c-means imputation algorithm gives reasonable data estimation for deviated and noisy data. And it is also beneficial for noise reduction and augmentation and provides better efficiency of clustering under noise.	This technique suffers from calculating the predefined cluster number and weighting factor values with high sensitivity.
17.	MIGEC (Multiple Imputation using Gray system theory and Entropy-based on Clustering) [32]	Hybrid	MIGEC outperforms many approaches such as KNNMI, FCMOCS, and CRI, independent of missing data types.	It does not perform well for high dimensionality.
18.	Fuzzy c-means-	Hybrid	This algorithm serves excellently for the	Not suitable for complex

	Multilayer Perceptron (FCM-MLP) [33]		imputation of multiple missing values in the dataset. This method performs better than K-means and multilayer perceptron imputation techniques.	dimensions.
19.	POCSimpute [34]	Knowledge assisted	POCSimpute produces a framework to choose the favored combination method flexibility. And it gives the most favorable solution despite the global or local correlation structure.	This technique has high computation cost in the comparison of LSimpute and SVD.
20.	Gene Ontology Imputation (GOimpute) [35]	Knowledge assisted	The GOimpute technique significantly enhances the accuracy of imputation. Experimental results have proven that this approach performs better where the balance of interpreted genes is enormous at a higher rate of missing values.	GOimpute methodology does not give better results in the comparison of LLSimpute and KNNimpute methods.
21.	HAIimpute [36]	Knowledge assisted	This imputation methodology is more robust when the missing rate is high.	This technique's disadvantage is that the histone acetylation information for every class is not available.

VII. EFFICIENCY MEASURES

If the missing rate is low, then the performance of all methods is nearly equal. And if the missing rate is greater than 40%, then performance is noticeable. When applying any technique for the missing data imputation, and then divide the dataset into two parts, i.e., training data and testing data. When a data set is divided into a training set and a testing set, the majority of the data is used for training, and just a small piece is used for testing. After that, test a model by generating a predictions model against the test set after it has been processed using the training set. And calculate the efficiency of the model based on some performance-based indexes like Normalized Mean Squared Error (NMSE) if minimum bias is occurred then, i.e., good imputation, Mean Absolute Error (MAE) shows the magnitude of overall error, Mean Absolute Percentage Error (MAPE) represents the percentage of absolute error, Root Mean Squared Error (RMSE) measure average root squared deviation of predicted error and Mean Percentage Error (MPE) is similar to MAPE, but this shows the direction of error. Where E_i is the difference between the actual value and predicted value and P_i is the imputed value, and σ^2 is the variance, and n is the number of the samples of the given data set, and i indicate that particular ith samples. And $\sum$ indicates the summation of the calculated value of the expression.

TABLE II. PERFORMANCE MATRIX

Sr. No.	Evaluation Method	Formula	Description
1.	NMSE	NMSE= $\frac{1}{n\sigma^2}\sqrt{\sum_{i=1}^{n}(E_i)^2}$	NMSE panelized extreme errors and the effects of positive and negative errors are not canceled out.
2.	MAE	MAE= $\frac{1}{n}\sum_{i=1}^{n}\lvert E_i\rvert$	In MAE, the positive and negative errors are not canceled out, but they do not provide any results on the direction of the errors.
3.	MAPE	MAPE= $\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{E_i}{P_i}\right\rvert \times 100$	It does not indicate which way the errors are coming from. However, it is unaffected by measurement scale but is affected by data transformation.
4.	RMSE	RMSE= $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i)^2}$	It includes critical errors that occurred during prediction and is sensitive to data scaling and transformation.
5.	MPE	MPE = $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{E_i}{P_i}\right) \times 100$	In MPE, the countersigned errors affect each other and cancel out, and for a good prediction, the obtained MPE must be small.

VIII. CONCLUSION AND FUTURE WORK

Various new researchers have evaluated efficacy relative to existing imputation algorithms. However, there is still no explicit consent around every data set's unique performance methods because many factors can affect the imputation methodology's efficiency. For example, one factor includes a higher missing percentage. Another factor is different types of data types included in imputation. There is no efficient imputation technique that is performing well for every dataset. This paper represents an exhaustive review of many of these methods. This paper classifies various techniques depending on whether they use knowledge from within the data or field information in imputation or data from external references. This paper utilizes distinct imputation methods such as the Local method, Global method, Hybrid method, and knowledge-assisted methods in the discussion. And justify that the technical structure and the postulates behind them produce a brief description of each algorithm's underlying methodology.

The association between various imputation methods grows more beneficial to researchers. The validity of the imputation outcome is an essential step in estimating the performance of any imputation algorithm. Each algorithm's performance is calculated based on three different evaluation scales, and their names are normal root mean squared error (NRMSE), Biomarker List Concordance Index (BLCI), and Cluster Pair Proportion (CPP) [38]. And then, calculate the average of these performance indexes, referred to as average (index). These performance scales are used for distinct intentions. When the need to evaluate the alteration between the predicted value and actual values, then used NRMSE performance index. And to determine the clustering outcome used CPP. And to estimate the effect of finding differentially expressed genes. Based on the performance scale of NRMSE, different algorithms yield different results. ILLS methodology gives better outcomes on the time-series dataset. LS methods produce a better accuracy on non-time series data. The LS algorithm also performs well on mixed-type datasets.

ILLS and LLS give better accuracy on the time series dataset based on the CPP performance scale. And SLLS gives better results on the time series data based on BLCI and average index performance scale. And most of the local learning algorithms undergo the over-fitting problem. Some methods yield a better result, but whenever the missing rate is high, then it does not give a better outcome. Until now, no algorithm gives better accuracy for all the datasets. This paperwork includes a detailed discussion on various approaches proposed to solve missing data imputation in gene expression datasets. Besides the discussion of multiple methods, merits and demerits of different techniques have also been elaborated. This paper expects to help the readers to discover prevailing advancements in this field and encourages the development of new algorithms.

REFERENCES

[1] W.-C. Liew, N.-F. Law, and H. Yan, Missing value imputation for gene expression data: computational techniques to recover missing data from available information, Briefings in bioinformatics, 12(5) (2011) 498–513.

[2] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, Missing data and multiple imputations in clinical, epidemiological research, Clinical epidemiology, 9 (2017) 157-166.

[3] A. Dubey, and A. Rasool, Time series missing value prediction: Algorithms and applications, International Conference on Information, Communication and Computing Technology, (2020) 21–36.

[4] A. Dubey and A. Rasool, Data mining based handling missing data, International conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), (2019) 483–489.

[5] X.Zhu, J.Wang, B.Sun, C.Ren, T.Yang, and J. Ding, An efficient ensemble method for missing value imputation in microarray gene expression data, BMC bioinformatics, 22(1) (2021) 1-25.

[6] A. Dubey, and A. Rasool, Clustering-based hybrid approach for multivariate missing data imputation, International Journal of Advanced Computer Science and Applications,11(11) (2020) 483-489.

[7] A. Dubey, and A. Rasool, Local Similarity-Based Approach for Multivariate Missing Data Imputation, IJAST, 29(6) (2020) 9208 - 9215.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, Missing value estimation methods for dna microarrays, Bioinformatics, 17(6) (2001) 520–525.

[9] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, A bayesian missing value estimation method for gene expression profile data, Bioinformatics, 19(16) (2003) 2088–2096.

[10] R.Vinas, T. Azevedo , ER. Gamazon, and P. Lio, Deep Learning Enables Fast and Accurate Imputation of Gene Expression, Frontiers in Genetics, 12 (2021) 489-500.

[11] M. Celton, A. Malpertuy, G. Lelandais , and A. G. De Brevern, Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments, BMC genomics, 11(1) (2010) 1–16.

[12] H. Kim, G. H. Golub, and H. Park Missing value estimation for dna microarray gene expression data: local least squares imputation, Bioinformatics, 21(2) (2005) 187–198.

[13] LF. Burgette, and JP Reiter, Multiple Imputation for Missing Data via Sequential Regression Trees, American journal of epidemiology, 172(9) (2010) 1070-1076.

[14] M. Ouyang, W. J. Welsh, and P. Georgopoulos, Gaussian mixture clustering and imputation of microarray data, Bioinformatics, 20(6) (2004) 917–923.

[15] M. S. B. Sehgal, I. Gondal, and L. S. Dooley, Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data, Bioinformatics, 21(10) (2005) 2417–2423.

[16] M. S. B. Sehgal, I. Gondal, L. S. Dooley, and R. Coppel, Ameliorative missing value imputation for robust biological knowledge inference, Journal of Biomedical Informatics, 41(4) (2008) 499–514.

[17] A. Farswan, A. Gupta, R. Gupta, and G. Kaur, Imputation of gene expression data in blood cancer and its significance in inferring biological pathways, Frontiers in oncology, 9 (2020) 1442-1451.

[18] A. Wang, J. Yang, and N. An Regularized sparse modeling for microarray missing value estimation, IEEE Access, 9 (2021) 16899–16913.

[19] CA.Mancuso, JL. Canfield, D.Singla, and A. Krishnan, A flexible, interpretable, and accurate approach for imputing the expression of unmeasured genes, Nucleic Acids Research, 48(21) (2020) 1-12.

[20] Y. Hu, B. Li, W. Zhang, N. Liu, P. Cai, F. Chen, and K. Qu, WEDGE: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition, Briefings in Bioinformatics, 4 (2021) 1-13.

[21] YL. Qiu, H. Zheng, and A. Gevaert, Genomic data imputation with variational auto-encoders, GigaScience, 9(8) (2020) 1-12.

[22] S. Faisal, and G. Tutz, Missing value imputation for gene expression data by tailored nearest neighbors, Statistical applications in genetics and molecular biology, 16(2) (2017) 95–106.

[23] X. Zhang, X. Song, H. Wang, and H. Zhang, Sequential local least squares imputation estimating the missing value of microarray data, Computers in biology and medicine, 38(10) (2008) 1112–1120.

[24] Z. Yu, T. Li, S.-J. Horng, Y. Pan, H. Wang, and Y. Jing, An iterative locally auto-weighted least squares method for microarray was missing value estimation, IEEE transactions on nano bioscience, 16(1) (2016) 21–33.

[25] R. Jörnsten, H.-Y. Wang, W. J. Welsh, and M. Ouyang, Dna microarray data imputation and significance analysis of differential expression, Bioinformatics, 21(22) (2005) 4155–4161.

[26] A. Purwar, and S. K. Singh, Hybrid prediction model with missing value imputation for medical data, Expert Systems with Applications, 42(13) (2015) 5621–5631.

[27] I. B. Aydilek, and A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences, 233 (2013) 25–35.

[28] I. B. Aydilek, and A. Arslan, A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks, International Journal of Innovative Computing, Information and Control, vo 7(8) (2012), 4705–4717.

[29] A .Tjärnberg, O .Mahmood, CA .Jackson, GA. Saldi , K.Cho , LA. Christiaen, and RA. Bonneau ,Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data, PLoS computational biology, 17(1) (2021) 1-22.

[30] H. Li, C. Zhao, F. Shao, G.-Z. Li, and X. Wang, A hybrid imputation approach for microarray missing value estimation, BMC genomics, 16(S9) (2015) 1-11.

[31] J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation, Transportation Research Part C: Emerging Technologies, 51 (2015) 29–40.

[32] J. Tian, B. Yu, D. Yu, and S. Ma, Missing data analyses: a hybrid multiple imputation algorithm using gray system theory and entropy based on clustering, Applied intelligence, 40(2) (2014) 376–388.

[33] S. Azim, and S. Aggarwal, Hybrid model for data imputation: using fuzzy c means and multi-layer perceptron, International Advance Computing Conference (IACC), (2014) 1281–1285.

[34] X. Gan, A. W.-C. Liew, and H. Yan, Microarray missing data imputation based on a set theoretic framework and biological knowledge, Nucleic Acids Research, 34(5) (2006) 1608–1619.

[35] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, Improving missing value estimation in microarray data with gene ontology, Bioinformatics, 22(5) (2006) 566–572.

[36] Q. Xiang, X. Dai, Y. Deng, C. He, J. Wang, J. Feng, and Z. Dai, Missing value imputation for microarray gene expression data using histone acetylation information, BMC bioinformatics, 9(1) (2008) 1–17.

[37] Z. Yu, T. Li, S.-J. Horng, Y. Pan, H. Wang, and Y. Jing, An iterative locally auto-weighted least squares method for microarray missing value estimation, IEEE transactions on nanobioscience, 16(1) (2016) 21–33.

[38] C.-C. Chiu, S.-Y. Chan, C.-C. Wang, and W.-S. Wu, Missing value imputation for microarray data: a comprehensive comparison study and a web tool, BMC systems biology, 7(6) (2013) 1–13.