An effective imputation scheme for handling missing values in the heterogeneous dataset

Sowmya Venkatesh[1], Maragal Venkatamuni Vijaya Kumar[2], Ashoka Davanageri Virupakshappa[3]
[1]Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India
[2]Department of Information Science and Engineering, Dr. Ambedkar Institute of Technology, Bengaluru, Karnataka, India
[3]Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India

Article Info

Article history:

Received May 10, 2023
Revised Jun 26, 2023
Accepted Jul 2, 2023

Keywords:

Clinical data
Heterogenous data
Imputation
Missing values
Natural language processing

ABSTRACT

A high level of data quality has always been a concern for many applications based on machine learning, including clinical decision support systems, weather forecasting, traffic predictions, and many others. A very limited amount of work is devoted to exploiting the missing values for effective imputation and better prediction. This paper introduces a unique approach to predicting and imputing missing data fields in the multivariate dataset such as numerical, categorical, and unstructured. The proposed imputation method is a multi-model scheme based on the joint approach of natural language processing (NLP) encoders, machine learning-driven feature extractors, and a sequential regression imputation technique to predict missing values. The proposed system is robust and scalable without requiring extensive engineering. The validation of the model is done on the benchmarked clinical dataset of heart disease obtained from UCI. The results show that the proposed methods achieve better imputation accuracy and require significantly less time than other missing data imputation methods.

Corresponding Author:

Sowmya Venkatesh
Department of Computer Science and Engineering, Dr. Ambedkar Institute of Technology
Bengaluru, Karnataka, India
Email: vsowmyaresearch@gmail.com

1. INTRODUCTION

Most of the real-life data have missing values which is a significant issue in the field of data-driven applications. This can negatively affect the accuracy of the predictive model and introduce bias into knowledge extraction. As a result, the quality of data is critical for accurate modeling for making good decisions [1]. Many reasons can account for missing data, including manual data entry or errors during maintenance or transmission, and incorrect measurements [2]. Clinical data is a significant example of real data which has an important role in discovering useful insights concerning patient treatment and clinical decisions [3]. However, clinical data are often heterogeneous and prone to missing information. Applying knowledge extraction and machine learning algorithms to such datasets can have a serious impact on the final result, so it is important to treat the dataset with the right approach [4]. The proposed study uses a heart disease dataset that is subjected to multivariate missing values. The proposed work considers this dataset as a case study for developing a support system based on predictive analytics which requires effective imputation and missing data handling mechanisms.

Methods for dealing with missing values can be divided into three groups, including imputation, case elimination, and systems based on prediction or approximation. Case elimination involves keeping only the cases with missing values and completing the tasks using the remaining data samples only. The second strategy involves employing techniques like predictive modeling to learn without having to cope with missing data.

Additionally, before using learning, missing data imputation encourages imputing missing values [5]–[7]. Missing data imputation is a process that substitutes certain logical values for missing values based on observed data. It has always been critical to impute missing values when learning from incomplete data since missing values can result in biases that affect the quality of learned models. Due to the fact that multivariate or heterogeneous datasets typically contain both discrete and continuous components, the majority of existing imputation algorithms cannot be used with these kinds of data sets [8], [9]. It also includes mixed-characteristic features such as numerical and categorical features with mixed-independent variables to describe heterogeneity. There are many methods of imputation but which one is better is not clear in the literature. In other words, choosing the most appropriate method is quite challenging [10], [11]. The proposed work reported in this paper introduces a new method for handling missing data to address the above practical requirement. This paper makes a unique contribution that can be applied to any heterogeneous dataset to impute missing values. The proposed imputation scheme analyzes the existing samples and uses that learning to do the imputation of missing values. The proposed scheme adopts an natural language (NLP) encoder that transforms data into suitable representation using word embedding, one-hot encoding, and normalization. On the other hand, the proposed system consists of a learning-based feature vector that represents the latent feature to map missing values with the sequences of existing data. The final module of the system performs imputation using a sequential regression technique to predict and fill in the missing values. The remaining section of this paper is organized as follows: section 2 presents a brief review of the related work on handling missing value in the dataset; section 3 highlights the challenging issue and the motivation the proposed system design and method implementation , the results are presented in section 4 to justify scope and the effectiveness and finally, section 5 concludes overall research contribution and future direction.

2. LITERATURE SURVEY

Recently, the imputation of missing values has attracted more and more attention from researchers. Several missing data imputation methods have been proposed, and they can show significant variations in terms of complexity and quality of the imputation. This section highlights several efforts using different imputation techniques to deal with missing data. The work carried out by Junger and Leon [12] presented an imputation method for handling missing values in the time-series dataset of air pollutants. The authors have used the expectation maximization algorithm to predict the missing values. The outcome of the study showed that the presented method is quite effective when the proportion of the missing value is less than 10%. In the work of Yuan *et al.* [13], the authors have used a long-short-term memory (LSTM) learning algorithm to capture long-term dependencies to impute the missing value and predict accurate PM2.5 concentration using the air pollutants dataset. An interesting work done by Jadhav *et al.* [14] conducted a comprehensive comparison of seven imputation techniques such as mean, predictive mean, median, k-nearest neighbor (KNN), random sample, linear regression, bayesian and non-bayesian linear method. All these methods are evaluated on numerical datasets only and results show the effectiveness of KNN over the other methods. Mean and median-based imputation is the most basic method used in the existing literature. These methods replace the missing value with the mean or median of non-missing values for the attribute. Ravi and Krishna [15] reported that mean and median-based imputation is not a good solution for the predictive model. In this setting, the relationships among features are more important in the predictive or classification task. Although KNN-based imputation is proven to be an effective method to impute missing values as it first identifies k-nearest neighbors, which are the most similar to the missing record among all records within the dataset by using a distance function. However, according to the authors in the study of Liu *et al.* [16] determining the appropriate k value is a challenging task. Also, it is quite expensive for a large dataset because it is required to search within the entire dataset to find the most similar records.

Various works based on KNN and its variants have been presented to address missing data imputation. Jiang and Yang [17] combined the application of c-means and KNN to impute missing values. Similarly, KNN is combined with the expectation-maximization algorithm in the work of Far *et al.* [18] for imputation. Khan and Hoque [19] have presented a hybrid chain equation technique based on single and multiple imputations. The validation of this technique is done on real-world clinical datasets and three public datasets. The outcome shows 20% higher FMeasure and 11% less error for binary and numeric data imputation, respectively. In the study of Nikfalazar *et al.* [20] decision tree and fuzzy clustering techniques are combined to develop an iterative imputation technique. Multiple datasets from the UCI machine learning repository are used in the experimental process. The research study conducted by Rani *et al.* [21] suggested a hybrid imputation technique for handling missing values in medical datasets. The presented technique is developed based on the combination of MICE, KNN, mean, and mode imputation techniques. The presented technique achieved a reduced RMSE score of 17% on the heart disease and 37% on the breast cancer dataset. Apart from this, there are many other recent works published on data imputation see Table 1 for improving the performance of data mining and machine learning techniques.

Table 1. Recent works on missing data imputation

Researchers	Dataset used	Method
Noor *et al.* [22]	Particulate matter dataset	Linear, quadratic, and cubic interpolation
Pereira *et al.* [23]	Medical dataset	Variational autoencoders
Qin *et al.* [24]	Chronic kidney disease	KNN imputation method
Jena and Dehuri [25]	Diabetes, mammographic, automobiles dermatology	Decision Tree and SVM
Mohan *et al.* [26]	Cleveland heart disease dataset	Litwise deletion
Cenitta *et al.* [27]	Ischemic heart disease	Fuzzy-rough sets
Kumar and Kumar [28]	Breast cancer dataset, and lung cancer datasets	Mean imputation, KNN, and fuzzy KNN
Desiani *et al.* [29]	Heart disease dataset	Deletion, mean, mode, and ANN
Venkatraman *et al.* [30]	DiabHealth dataset	Mean and mode imputation method
Rani *et al.* [31]	Heart disease dataset	KNN, MICE, mean, and mode
Kim *et al.* [32]	Photovoltaic dataset	Linear interpolation, KNN imputation, mode imputation, MICE
Howey *et al.* [33]	Early inflammatory arthritis data set	Bayesian networks and nearest neighbor imputation
Muhaideb and Menai [34]	UCI Medical dataset	Litwise deletion
Hu *et al.* [35]	EHR	Mean and MICE

3. METHOD

The missing data samples in the dataset are like a question with no resolution. It can be found in many industrial and research databases, which can reduce the reliability of data-driven tasks and the learning of predictive models from data. Nowadays, 10-50% of records are often missing in a database, making it extremely challenging to analyze and extract valuable insights using data analysis and computational intelligence techniques, which can only be reliable with complete data. Missing data leads to bias, affecting the performance of predictions and the quality of learned patterns. One classification method is to build a classifier ignoring observations with missing values. It is practical only when their significance relative to the class label is negligible. Taking into account correct imputation improves classification accuracy even with a 5% missing rate. Analysis of the literature shows that the researchers left significant scope and direction for improvement. The proposed research work focuses on developing an effective imputation technique to fill in missing values in the training and test datasets to enhance the classifier's overall performance when tested. The proposed system considers labels into account during the imputation of missing samples in the training set. This ensures overall improvement in classification performance and achieves a realistic approach to predictive modeling, benefiting many data-driven real-world applications.

3.1. Proposed imputation method

This section presents the system design of the imputation scheme and its implementation procedure to handle missing data problem which is often encountered in classification tasks. The proposed work has considered a multivariate clinical dataset of heart disease as a case study towards benefiting the development of the predictive model for identifying whether a person is prone to heart disease or not. The study introduces an adaptive and scalable imputation scheme that can be introduced to a multi-variate dataset that exhibits heterogeneous characteristics. The proposed scheme blends combine NLP encoder and learning-based imputation technique offering effective missing data handling functions. The proposed scheme is a fully conditional specification that considers a column with a useful feature as its input and returns imputed value as output in the column i.e., to-be-imputed column. Figure 1 illustrates the schematic architecture of the proposed imputation scheme. As shown in the figure, a five-layer system is proposed to handle missing values using efficient methods of imputing for imputing different variants of data samples. The proposed system is composed of a data processing module, NLP encoder, feature representor, imputer, and predictive analytics. The first module is to analyze the characteristics of the dataset, proportion missing context, and split it into training and testing sets for validation of imputed outcomes. Next, a module is about making the input dataset suitable for further processing using an NLP encoder where numerical data are subjected to encoding using the normalization technique, categorical datasets are encoded using a one-hot encoding mechanism and text data are encoded to sequences of strings. The third module is about extracting important features. In this module, different types of techniques are integrated into a function which according to the type of data extract significant features. The study uses a simple neural network for extracting features of a numerical dataset, word embedding is done to represent encoded categorical data into vectors and recurrent neural network (RNN) is used to capture features considering the long term dependency of the encoded string data. The extracted features are concatenated and introduced to the imputation module which uses an application of sequential regression technique. The final module of the system is predictive modeling using different machine learning classifiers. This module is adopted as a self-validation mechanism to justify and visualize the effectiveness of the proposed imputation scheme in the prediction of heart disease.

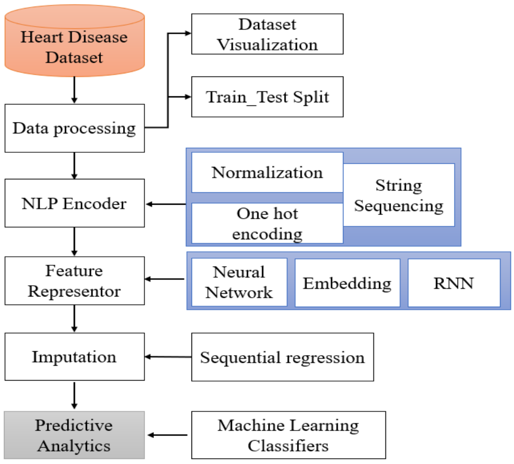

Figure 1. Schematic illustration of the proposed imputation system

3.1.1. Dataset

The dataset used in this research work is a publicly accessible heart disease dataset obtained from the UCI machine learning repository [AR]. The dataset was obtained from 303 patients suspected of having heart disease. It consists of several features but only 14 features or attributes were taken into consideration which are most useful and widely used in the literature for predictive analytics. Table 2 illustrates the complete information on the adopted dataset.

Table 2. Description of the adopted dataset

SI.NO.	Column	Details	Value
1	Age	The age of the patient	Ranging between 29 and 77
2	Sex	Gender of a patient person	Female [0], Male [1]
3	CP	Chest pain	Different values 0, 1, 2, and 3 representing the level of severity of pain
4	RestBP	Blood pressure was measured during the patient was admitted to the healthcare center	Ranging between 94 and 200
5	Chol	Cholesterol level measured during patient admitted	Ranging between 126 and 564
6	FBS	Fasting blood sugar level	Binary value depending on the level if >120mg/dl =1, otherwise = 0.
7	RestECG	ECG	Ranging from 0 to 2
8	HeartBeat	heartbeat count	Ranging between 71 and 202
9	Exang	chest pain caused by reduced blood flow	True [1] and false [0]
10	OldPeak	depression status	Different levels ranging between 0 and 6.2.
11	Slope	The condition of the patient during peak exercise	Upsloping [1], Flat [2], down sloping [3]
12	CA	Fluoroscopy status	Ranging from 0 to 3
13	Thal	Thallium test	Ranging from 0 to 3
14	Target	The response variable (outcome class)	No chance of heat attack 0, higher chance of heart attack [1]

3.1.2. Rationale behind choosing the dataset

The attributes in the dataset describe a range of conditions that can lead to a heart attack. The rationale behind choosing this dataset is that after the covid pandemic, many peoples died from heart disease. Heart disease has become one of the leading causes of morbidity and mortality. The victims are not only the elderly even, young people are becoming prone to heart disease. It is difficult to identify heart disease because of several health conditions like diabetes, blood pressure, and high cholesterol. In this context, predictive analytics using machine learning techniques appears to be one of the hottest topics in clinical data analysis. A large amount of data is generated in the healthcare sector. By using data mining, healthcare data can be transformed into knowledge that can be used to make predictions and clinical decisions. But most of the clinical data are subjected to missing values which impacts the performance of machine learning and its deployment in real-time systems. By applying suitable imputation techniques, the performance and reliability of predictive analytics can be improvised efficiently.

3.1.3. Proposed imputation

The dataset consists of a total 303 number of samples and 14 features. Based on further exploration it has been analyzed that fewer missing values (i.e., approximately 2%) missing values in the data. Before proceeding imputation scheme on the input dataset, the study applies a random masking strategy to simulate missing data with a proportion of 25% see Figure 2.

The next operation involves encoding the dataset where a specific operation is applied to a particular type of the dataset as shown in Figure 3. This process is required to make data understandable to the machine. Also, in the data normalization technique min-max scaling is used and in string sequencing normal encoder and decoder module is used to make text data into sequences. On the other hand, the categorical values are converted into numerical representation using one-hot encoding. Next, the task of feature extraction is done in similar fashion, where different feature learning technique is used for different kind of data types see Figure 4.

	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
182	NaN	NaN	typical	NaN	NaN	0.0	NaN	178.0	NaN	NaN	1.0	2.0	normal	No
216	46.0	NaN	nontypical	105.0	204.0	0.0	0.0	172.0	NaN	0.0	1.0	0.0	normal	No
177	56.0	1.0	asymptomatic	132.0	184.0	0.0	2.0	105.0	1.0	NaN	2.0	1.0	NaN	NaN
290	NaN	NaN	nonanginal	NaN	212.0	0.0	2.0	NaN	0.0	0.8	2.0	0.0	reversable	Yes
50	NaN	0.0	nontypical	105.0	198.0	0.0	0.0	168.0	0.0	0.0	NaN	1.0	normal	No

Figure 2. Sample visualization of the modified input dataset

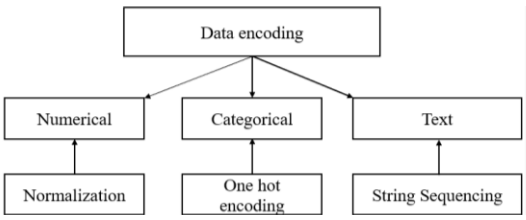

Figure 3. Data encoding with respect to data type

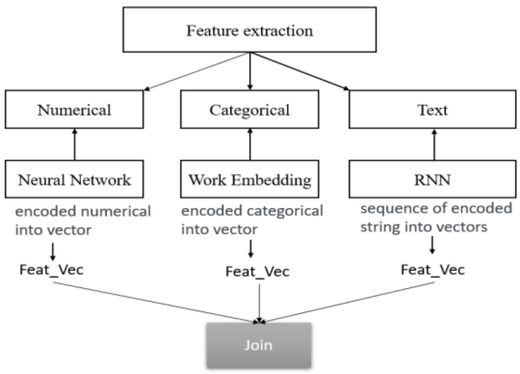

Figure 4. Illustration of the feature extraction process

The neural network implemented here is configured with a single input layer, a single hidden layer, and a fully connected layer. At the hidden layer, 100 neuron units are considered with the Relu activation function. For encoded categorical features are extracted using the word embedding process where the embedding dimension is considered equal to 10. For the string vector, the study implemented two RNN networks with 50 LSTM neuron units. The extracted features based on learning models are then concatenated to form a latent feature vector. The next operation in the proposed system is about applying the imputation technique. Algorithm 1 provides a method for evaluating the missing data imputation.

Algorithm 1. Missing data imputation
```
Input: x column of missing instance dataset (D)
Output: x′ an updated column with imputed missing data
Start
    1.  Init m (number of iterations)
    2.   //First imputation attempt
    3.  For each missing value in x ∈ D do
    4.    Fill in missing values randomly
    5.  end of For
    6.  While a change in predictions do
    7.    For each missing value in x do
```

```
    regression analysis between observed values of missing instances and other
variables
    repeat for n times
    x'← Impute missing values in x
  end of For
  for i: x
  Check the most frequent predicted value in x'
Update D ←x'
  end
end
End
```

The proposed imputation algorithm is a modified version of the MICE technique. It generates m values for a single missing data sequentially imputes all the input features and adds predicted data into an array. Then a missing value is replaced with the most frequent item of the array.

3.1.4. Predictive analytics

The study implements multiple supervised classifiers for identifying whether there is a risk of heart attack or not. In this phase, the modeling is done with both data i.e., original data and imputed data by the proposed system. Also, output from both cases will be compared to see the differences in the classification result. For this both the original and imputed dataset is subjected to train_test split operation with a ratio of 80:20. Before going to predictive analytics, the study performs basic exploratory data analysis from the original dataset to better understand the distribution of output data samples.

From Figure 5, it can be analyzed a greater number of samples are belonging to the disease class compared to the non-disease class. This shows the dataset is a little imbalanced but it will not create any significant biases towards the majority class. From the analysis of Figure 6, it can be analyzed that, most of the patients are in the age between 50 to 60. Also, based on the statistical analysis it has been estimated that the age of the youngest people is 29 and the age of the elderly is 77. From the analysis of Figure 7, it can be seen that male persons are most prone to heart disease.

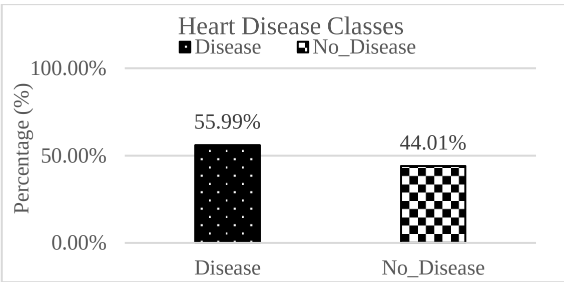

Figure 5. Class label distribution

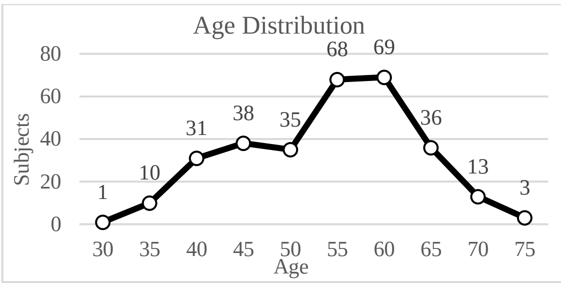

Figure 6. Analysis of age distribution

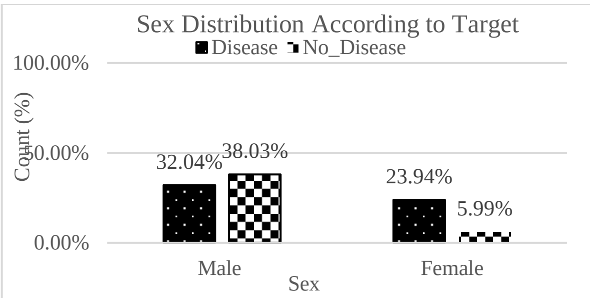

Figure 7. Analysis of gender distribution

4. RESULT ANALYSIS

The design and development of the proposed imputation technique and predictive modeling are carried out using python programming language executed on anaconda distribution. This section presents the outcome and performance analysis to justify the scope of the proposed work.

4.1. Performance analysis of imputation methods

The study considers RMSE as a standard performance indicator for evaluating the performance of the imputation technique on numerical missing data. The outcome from Figure 8 shows the effectiveness of the proposed imputation scheme compared to mean imputation and KNN-based imputation. Similarly, as shown in Figure 9 the proposed scheme introduced on categorical data outperforms other existing techniques namely the random sampler and common imputer technique.

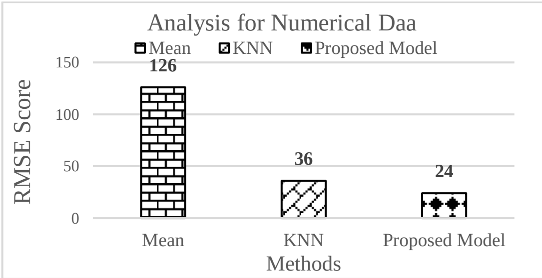

Figure 8. Performance analysis of imputation on numerical feature

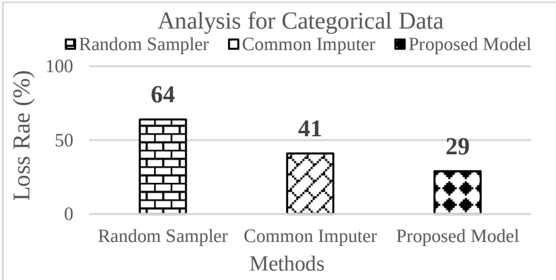

Figure 9. Performance analysis of imputation on categorical feature

4.2. Analysis of imputation for predictive analytics

The study has implemented two supervised classifiers namely SVM and Naïve Bayes (NB) classifier. Both learning models are evaluated on the original dataset and the imputed dataset. The outcomes are analyzed using a confusion matrix and training and testing accuracy. A closer analysis of the above graphs Figures 10-13 reveals that the proposed imputation technique provides better handling of missing samples in the dataset. As it does not exhibit significant differences between the outcome from the original dataset and imputed data. The accuracy has been evaluated using the following equation [36].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

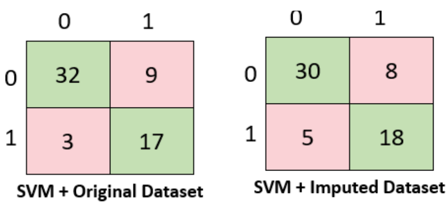

Figure 10. Confusion plot for SVM

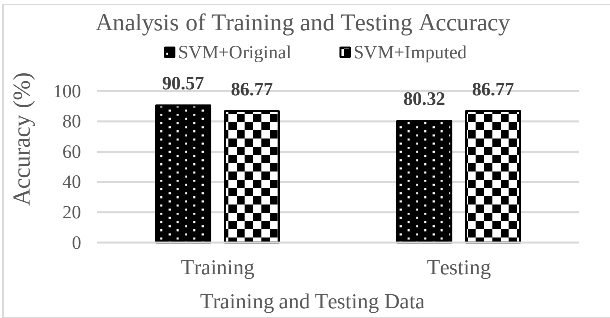

Figure 11. SVM training and testing accuracy

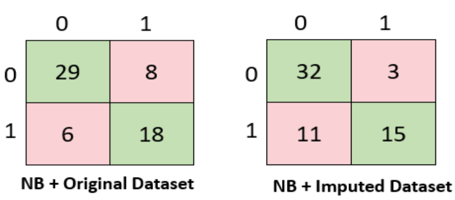

Figure 12. Confusion plot for NB

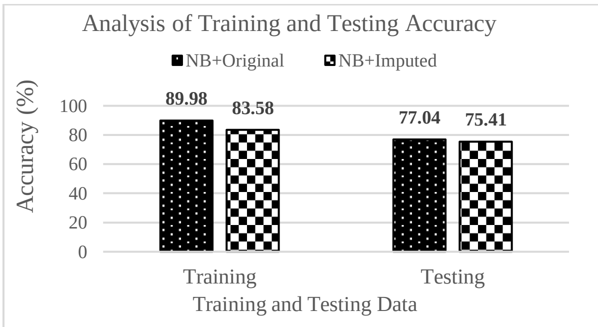

Figure 13. SVM training and testing accuracy

5. CONCLUSION

The proposed study has introduced an efficient mechanism of imputation for handling missing values in multivariate datasets. A systematic implementation procedure is adopted in the literature where first all the data are encoded to numerical representation using specific encoding techniques. Further, the study has implemented a learning-based feature representor. The unique thing about this step is that the feature representor module is based on a neural network and sequence prediction model, which makes it adaptive to fit the dataset. Finally, the imputation is carried out using a customized MICE algorithm. An extensive analysis is carried out to evaluate the performance of the proposed technique. The outcome reveals the effectiveness of the proposed imputation in terms of RMSE and loss rate. In addition, predictive analytics is also carried out to evaluate the proposed scheme. It is to be noted here that predictive modeling is carried out without doing hardcore preprocessing and feature engineering in both cases i.e., predictive modeling on the original dataset and imputed dataset. However, there is a good scope for improvement in improving the performance of the classifier. In future work, the proposed imputation scheme will be evaluated with multiple and complex datasets with more optimization.

REFERENCES

[1] Y. Fu, H. Liao, and L. Lv, "A comparative study of various methods of handling missing data in unsoda," *Agriculture (Switzerland)*, vol. 11, no. 8, p. 727, Jul. 2021, doi: 10.3390/agriculture11080727.

[2] J. Poulos and R. Valle, "Missing data imputation for supervised learning," *Applied Artificial Intelligence*, vol. 32, no. 2, pp. 186–196, Apr. 2018, doi: 10.1080/08839514.2018.1448143.

[3] B. J. Wells, A. S. Nowacki, K. Chagin, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 1, no. 3, p. 7, Dec. 2013, doi: 10.13063/2327-9214.1035.

[4] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013, doi: 10.4097/kjae.2013.64.5.402.

[5] P. Li, E. A. Stuart, and D. B. Allison, "Multiple imputation: a flexible tool for handling missing data," *JAMA - Journal of the American Medical Association*, vol. 314, no. 18, pp. 1966–1967, Nov. 2015, doi: 10.1001/jama.2015.15281.

[6] M. C. M. D. Goeij, M. V. Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker, "Multiple imputation: dealing with missing data," *Nephrology Dialysis Transplantation*, vol. 28, no. 10, pp. 2415–2420, Oct. 2013, doi: 10.1093/ndt/gft221.

[7] B. Suthar, H. Patel, and A. Goswami, "A survey: classification of imputation methods in data mining," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 1, pp. 309–312, 2012, [Online]. Available: www.ijetae.com.

[8] F. Bashir and H. L. Wei, "Handling missing data in multivariate time series using a vector autoregressive model based imputation (VAR-IM) algorithm: Part I: VAR-IM algorithm versus traditional methods," in *24th Mediterranean Conference on Control and Automation, MED 2016*, Jun. 2016, pp. 611–616, doi: 10.1109/MED.2016.7535976.

[9] J. Josse and F. Husson, "missMDA: a package for handling missing values in multivariate data analysis," *Journal of Statistical Software*, vol. 70, no. 1, 2016, doi: 10.18637/jss.v070.i01.

[10] C. Fang and C. Wang, "Time series data imputation: a survey on deep learning approaches," *arXiv*,. 2020, doi: 10.48550/arxiv.2011.11347.

[11] F. Popham, E. Whitley, O. Molaodi, and L. Gray, "Standard multiple imputation of survey data didn't perform better than simple substitution in enhancing an administrative dataset: the example of self-rated health in England," *Emerging Themes in Epidemiology*, vol. 18, no. 1, p. 9, Dec. 2021, doi: 10.1186/s12982-021-00099-z.

[12] W. L. Junger and A. P. D. Leon, "Imputation of missing data in time series for air pollutants," *Atmospheric Environment*, vol. 102, pp. 96–104, Feb. 2015, doi: 10.1016/j.atmosenv.2014.11.049.

[13] H. Yuan, G. Xu, Z. Yao, J. Jia, and Y. Zhang, "Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks," in *UbiComp/ISWC 2018 - Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers*, Oct. 2018, pp. 1293–1300, doi: 10.1145/3267305.3274648.

[14] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, Aug. 2019, doi: 10.1080/08839514.2019.1637138.

[15] V. Ravi and M. Krishna, "A new online data imputation method based on general regression auto associative neural network," *Neurocomputing*, vol. 138, pp. 106–113, Aug. 2014, doi: 10.1016/j.neucom.2014.02.037.

[16] Z. G. Liu, Y. Liu, J. Dezert, and Q. Pan, "Classification of incomplete data based on belief functions and k-nearest neighbors," *Knowledge-Based Systems*, vol. 89, pp. 113–125, Nov. 2015, doi: 10.1016/j.knosys.2015.06.022.

[17] C. Jiang and Z. Yang, "CKNNI: An improved KNN-based missing value handling technique," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9227, 2015, pp. 441–452.

[18] R. R. Far, B. Cheng, M. Saif, and M. Ahmadi, "Similarity-learning information-fusion schemes for missing data imputation," *Knowledge-Based Systems*, vol. 187, p. 104805, Jan. 2020, doi: 10.1016/j.knosys.2019.06.013.

[19] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *Journal of Big Data*, vol. 7, no. 1, p. 37, Dec. 2020, doi: 10.1186/s40537-020-00313-w.

[20] S. Nikfalazar, C. H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowledge and Information Systems*, vol. 62, no. 6, pp. 2419-2437, 2020, doi: 10.1007/s10115-019-01427-1.

[21] P. Rani, R. Kumar, and A. Jain, "HIOC: a hybrid imputation method to predict missing values in medical datasets," *International Journal of Intelligent Computing and Cybernetics*, vol. 14, no. 4, pp. 598–616, Oct. 2021, doi: 10.1108/IJICC-03-2021-0042.

[22] M. N. Noor, A. S. Yahaya, N. A. Ramli, and A. M. M. Al Bakri, "Filling missing data using interpolation methods: study on the effect of fitting distribution," *Key Engineering Materials*, vol. 594–595, pp. 889–895, Dec. 2014, doi: 10.4028/www.scientific.net/KEM.594-595.889.

[23] R. C. Pereira, P. H. Abreu, and P. P. Rodrigues, "Partial multiple imputation with variational autoencoders: tackling not at randomness in healthcare data," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4218–4227, Aug. 2022, doi: 10.1109/JBHI.2022.3172656.

[24] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020, doi: 10.1109/ACCESS.2019.2963053.
[25] M. Jena and S. Dehuri, "An integrated novel framework for coping missing values imputation and classification," *IEEE Access*, vol. 10, pp. 69373–69387, 2022, doi: 10.1109/ACCESS.2022.3187412.
[26] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[27] D. Cenitta, R. V. Arjunan, and K. V. Prema, "Ischemic heart disease multiple imputation technique using machine learning algorithm," *Engineered Science*, vol. 19, pp. 262–272, 2022, doi: 10.30919/es8d681.
[28] R. N. Kumar and M. A. Kumar, "Enhanced fuzzy K-NN approach for handling missing values in medical data mining," *Indian Journal of Science and Technology*, vol. 9, no. S1, Dec. 2016, doi: 10.17485/ijst/2016/v9is1/94094.
[29] A. Desiani, N. R. Dewi, A. N. Fauza, N. Rachmatullah, M. Arhami, and M. Nawawi, "Handling missing data using combination of deletion technique, mean, mode and artificial neural network imputation for heart disease dataset," *Science and Technology Indonesia*, vol. 6, no. 4, pp. 303–312, Oct. 2021, doi: 10.26554/sti.2021.6.4.303-312.
[30] S. Venkatraman, A. Yatsko, A. Stranieri, and H. F. Jelinek, "Missing data imputation for individualised CVD diagnostic and treatment," in *Computing in Cardiology*, 2016, vol. 43, pp. 349–352, doi: 10.22489/cinc.2016.100-179.
[31] P. Rani, R. Kumar, and A. Jain, "Multistage model for accurate prediction of missing values using imputation methods in heart disease dataset," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 59, 2021, pp. 637–653.
[32] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Applied Sciences (Switzerland)*, vol. 9, no. 1, 2019, doi: 10.3390/app9010204.
[33] R. Howey, A. D. Clark, N. Naamane, L. N. Reynard, A. G. Pratt, and H. J. Cordell, "A bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships," *PLoS Genetics*, vol. 17, no. 9, p. e1009811, Sep. 2021, doi: 10.1371/journal.pgen.1009811.
[34] S. Almuhaideb and M. E. B. Menai, "An individualized preprocessing for medical data classification," *Procedia Computer Science*, vol. 82, pp. 35–42, 2016, doi: 10.1016/j.procs.2016.04.006.
[35] Z. Hu, G. B. Melton, E. G. Arsoniadis, Y. Wang, M. R. Kwaan, and G. J. Simon, "Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record," *Journal of Biomedical Informatics*, vol. 68, pp. 112–120, Apr. 2017, doi: 10.1016/j.jbi.2017.03.009.
[36] R. H. Hridoy, A. D. Arni, and M. A. Hassan, "Recognition of mustard plant diseases based on improved deep convolutional neural networks," in *2022 IEEE Region 10 Symposium, TENSYMP 2022*, Jul. 2022, pp. 1–6, doi: 10.1109/TENSYMP54529.2022.9864487.

BIOGRAPHIES OF AUTHORS

Sowmya Venkatesh is an accomplished data scientist with experience in both academia and industry. She currently works for an MNC and is also a research scholar at Dr. Ambedkar Institute of Technology, Bangalore, India. She specializes in data science with a particular focus on NLP, text pre-processing, text generation, pattern recognition, sentimental analysis, and text summarization. Sowmya holds an M.Tech. degree in Software Engineering and a B.E. degree in Computer Science & Engineering from Visweswaraya Technological University, Karnataka, India. She has an indelible reputation for analyzing and transforming natural language data, using features developed by this transformation to build natural language processing applications. She has published in both proceedings and journals, and her research interests include machine learning, deep learning, NLP, computer vision, and text generation and transformation. She can be contacted at email: sowmya512v@gmail.com.

Dr. Maragal Venkatamuni Vijaya Kumar is a reformer educationist, renowned researcher and administrator. Presently working as a professor and HOD in ISE the Dr. Ambedkar Institute of Technology, Bangalore, India. He has 20 years of teaching, administration and research experience. He has been recognized as a best administrator with good track of academic results and research. He completed M.Tech. in computer science and engineering from the department of computer science, KREC, Surathkal, presently named as NITK, India. He is awarded Ph.D. in computer science and engineering, from the University of Hull, England, United Kingdom. He can be contacted at email: dr.vijay.research@gmail.com.

Ashoka Davanageri Virupakshappa having more than 27 years of experience and presently working as a professor in the department of information science and engineering, JSS Academy of Technical Education, Bengaluru. He worked as dean (research) JSSATEB, professor and head, department of CSE/ISE in various reputed Engineering colleges of Karnataka, India. His research area includes: knowledge engineering, operating system virtualization, requirement engineering, artificial intelligence, software engineering and architecture. He is serving as editorial board member, review committee member of various national and international journals, technical program committees/session Chair/organizing committee member for various national and international conferences. Dr. DV Ashoka is one of the National Award winners "Rashtriya Ekta Samman-2013". He can be contacted at email: dr.dvashoka@gmail.com.