Missing Data Imputation in Healthcare Datasets

Using Machine Learning Techniques

Research Project Report

submitted by:

Bonumaddi Naga Pravallika

120CS0121

Under the guidance of

Prof. Pabitra Mohan Khilar

15 November, 2023

Department of Computer Science and Engineering,

National Institute of Technology, Rourkela

Odisha 769008, India

CONTENTS

MOTIVATION

In the medical world, think of missing data like pieces of a patient's health story left blank. It's like having an incomplete picture of their well-being. Now, imagine trying to make important health decisions without all the pieces – it's challenging and risky. Completing these missing parts is like filling in the gaps in a patient's health story. Why does this matter so much? Well, it's about making sure doctors have all the details they need to provide the best care. It's like making sure a puzzle is complete, so nothing important is left out. For example, if a patient has missing data in their medical history, it's like having a book with some pages torn out. You might miss important clues about their health. Imputing missing data is like fixing those torn pages, making sure the whole story is available. And it's not just about one patient. In the world of medical research and finding better treatments, having complete and accurate data is like having a clear roadmap. It helps scientists and doctors discover patterns and come up with new ways to improve health.

EXECUTIVE SUMMARY

This research project delves into the realm of missing data imputation in healthcare-related datasets, leveraging a novel approach known as Class Center-Based Missing Value Imputation with Normalization and Outlier Handling (CCMVI-NOH). The primary objective is to address the ubiquitous issue of missing data in healthcare datasets, aiming to provide a robust and efficient solution for estimating missing values and improving data quality.[7]

Research Problem: The research addresses the critical issue of missing data in healthcare datasets, which can significantly impact the accuracy and reliability of data-driven applications. The challenge is to develop a missing data imputation method that not only accurately estimates missing values but also takes into account data normalization and outlier handling, preserving the integrity of the dataset.

Objectives: The primary objectives of this research project are threefold: To develop and evaluate the CCMVI-NOH algorithm as a comprehensive solution for missing data imputation in healthcare datasets.[4] To optimize the imputation process using the Firefly Algorithm, incorporating the search procedure for improving accuracy.[5] To investigate the performance of the CCMVI-NOH algorithm across diverse healthcare datasets, varying degrees of missingness, and missing data mechanisms.

Methodology: The research employs a multi-step approach, combining class center-based imputation with data normalization and outlier handling, supported by the Firefly Algorithm for optimization. The key steps include:

Data preprocessing, encompassing missing data filling, data scaling, and outlier handling. Simulated introduction of missingness into the complete dataset, creating a range of datasets with varying percentages of missing data. Application of three different imputation algorithms - Simple Imputer, Sampling-based decision trees, and CCMVI-NOH to the datasets.[9] Evaluation of the imputed data using metrics such as Mean Squared Error

(MSE), Mean Absolute Error (MAE), accuracy, and R-squared (R2)[1]. Optimization of the CCMVI-NOH algorithm using the Firefly Algorithm.[6]

Expected Contributions: This research project is anticipated to make several notable contributions to the field of missing data imputation in healthcare datasets. These contributions include:

The introduction of CCMVI-NOH as an innovative and comprehensive imputation method, which not only estimates missing data but also addresses data normalization and outlier handling. Insights into the performance of the CCMVI-NOH algorithm across diverse healthcare datasets, providing a basis for assessing its generalizability. The development of a robust pipeline for missing data imputation, enabling the application of the CCMVI-NOH algorithm to a variety of healthcare datasets. In conclusion, this research project endeavors to offer an advanced solution to missing data imputation in healthcare datasets, potentially enhancing data quality for critical applications in healthcare, data analysis, and decision-making processes.

*

1 INTRODUCTION

The domain of computer science engineering is marked by its relentless pursuit of solutions to real-world problems, often with far-reaching consequences. This research project delves into a critical problem that has implications not only in the realms of technology but also directly influences the healthcare industry. The central focus of this research is the pervasive issue of missing data imputation in healthcare datasets, a challenge that holds profound significance for healthcare data analysis, decision-making processes, and ultimately, patient care.

Context of the Research Problem: Healthcare data analysis plays an indispensable role in modern healthcare applications, from predicting diseases to monitoring patients and optimizing resource allocation. These data-driven approaches hold the promise of enhancing healthcare outcomes, yet they are often hindered by the presence of missing data within healthcare datasets. The origins of this missing data can be diverse, stemming from patient non-compliance, sensor errors, discrepancies in data entry, and various other sources.

The implications of missing data in healthcare datasets cannot be understated. Even a small percentage of missing information can introduce bias, undermining the reliability of analyses and decisions made based on such data., clinical policies, and resource allocation within the healthcare system.

The need to address missing data in healthcare datasets is thus underscored by its potential to not only improve the accuracy and reliability of healthcare data analysis but also to enhance the quality of care that patients receive. This research project stands as a significant endeavor to develop innovative solutions, combining Class Center-Based Missing Value Imputation with Normalization and Outlier Handling (CCMVI-NOH) and optimizing the process using the Firefly Algorithm. By doing so, it aspires to provide a comprehensive and efficient approach to missing data imputation, with the aim of strengthening the foundations of healthcare data analysis and, in the long run, improving patient outcomes.

In the subsequent sections of this research proposal, we will delve into the methodology, objectives, and expected contributions of this research project, shedding light on the innovative approach and its potential to address the challenge of missing data in healthcare datasets

2 RESEARCH PROBLEM

The research problem at the heart of this investigation revolves around the intricate challenge of missing data imputation in healthcare datasets. This challenge represents a fundamental issue with profound implications for the field of healthcare data analysis, influencing the accuracy of predictive models, healthcare decision-making, and ultimately, the quality of patient care.

Rationale and Relevance: The choice of this research problem is inspired by several key factors:

Prevalence in Healthcare Data: Missing data is a pervasive and recurrent issue in healthcare datasets, stemming from a multitude of sources, including patient non-compliance, sensor errors, and discrepancies in data entry. Recognizing the prevalence of this issue is a crucial step towards improving the overall quality of healthcare records.

Impact on Data Analysis: Missing data is not merely an inconvenience; it can significantly impact the integrity of data analyses. The introduction of missing data can lead to biased results, reduced data quality, and compromised analytical outcomes within the healthcare domain. Addressing this issue is essential to ensure that data-driven decisions in healthcare are founded on complete and accurate information.

Healthcare Sector Sensitivity: The healthcare sector is a uniquely sensitive domain in which data is intricately tied to patient well-being, treatment decisions, and resource allo- cation. The handling of missing data is not solely an analytical concern; it is a matter of patient welfare and healthcare policy. The ability to provide reliable and comprehensive healthcare data is integral to optimizing patient care.

3 RESEARCH OBJECTIVES

Development and Implementation of CCMVI-NOH Algorithm: The primary objective of this research is to develop and implement the Class Center-Based Missing Value Imputation with Normalization and Outlier Handling algorithm. This involves crafting a systematic methodology that combines class center-based imputation with normalization and outlier handling techniques, providing a comprehensive approach for handling missing data in healthcare datasets.

Assessment of Predictive Accuracy: One of the core objectives of this research is to assess the predictive accuracy of the CCMVI-NOH imputation method. This evaluation will involve measuring the correlation between the imputed values and the actual data, quantified through the Pearson Correlation Coefficient (r). Additionally, we will assess the closeness of imputed values to the actual data using the Root Mean Squared Error (RMSE). The objective is to demonstrate that CCMVI-NOH generates imputed values that closely align with the actual data, enhancing the reliability of predictive modeling in healthcare applications.

Preservation of Distributional Accuracy: Another crucial objective is to examine the distributional accuracy of CCMVI. We aim to determine whether the imputed data retains the distributional characteristics of the original healthcare dataset. This will be evaluated through the Kolmogorov–Smirnov distance (DKS), which quantifies the distance between the empirical distribution of imputed data and the original dataset. Our objective is to demonstrate that CCMVI can effectively preserve the true distribution of data values, crucial for maintaining the integrity of healthcare data analysis.

. **Evaluation of Classification Accuracy:** This research strives to evaluate the classification accuracy of datasets containing imputed values generated using CCMVI-NOH. Machine learning classifiers, including Decision Trees (DT), and Class Center-Based Missing Value Imputation, will be trained on these imputed datasets to mea- sure the effectiveness of CCMVI-NOH in producing imputed values that facilitate accu- rate classification. This objective contributes to enhancing the reliability of healthcare decision-making processes.

Analysis of Missing Data Characteristics: An additional key objective is to comprehensively analyze the impact of missing data characteristics, including the rate of missing data, types of missing data mechanisms (MCAR, MAR, MNAR), and the nature of the dataset (numeric, categorical, mixed). This analysis aims to provide valuable insights into how CCMVI-NOH performs under various conditions and missing data scenarios, offering a deep understanding of the method's robustness and generalizability within healthcare datasets.

4 RESEARCH QUESTIONS/ HYPOTHESIS

Before discovering the optimized solution for this problem we need to think and make some hypothesis which solution.

Hypothesis	Null Hypothesis (Ho)	Alternative Hypothesis (H1)
Combination of CCMVI- NOH and Enhanced Firefly Algorithms vs. Statistical Methods	There is no significant difference in the accuracy and robustness of imputed data between mentioned imputation techniques and statistical methods.	CCMVI-NOH imputation techniques yield significantly more accurate and robust imputed data compared to traditional methods.
Applicability Across Diverse Data Domains	The effectiveness of CCMVI - NOH imputation does not vary significantly across different data domains with distinct distributions and types of missing data.	Combo of CCMVI – NOH with Enhanced Firefly demonstrates varying levels of effectiveness across different data domains with distinct distributions and types of missing data with great accuracy

Table1. Research Hypothesis

5 LITERATURE REVIEW

Author Name	Year of Publication	Method Used	Result/ Conclusion
Nugroho, H., Utama, N.P. & Surendro	2021	The proposed method in this paper. FA (Firefly Algorithm), is an adaptive approach model that combines statistical and machine learning techniques for Optimizing the algorithms.	The proposed firefly algorithm (FA) showed efficient performance in handling missing data, with the Pearson correlation coefficient (r) and root mean squared error (RMSE) values close to 1 and 0, respectively.
Ching-Hsue Cheng	2021	The proposed method combines clustering with PKNNI (partial K- nearest neighbor imputation) and DNNI (distance-based nearest neighbor imputation) imputation techniques.	The proposed clustering-based purity and distance imputation method showed improved performance in terms of accuracy. AUC, and RMSE for different missing degrees and missing types in medical datasets.
Aliya Aleryani	2020	The paper proposes the use of Multiple Imputation Ensembles (MIE) as a robust approach for handling missing data in classification problems. It integrates multiple imputation and ensemble methods, specifically bagging and stacking, to improve classification accuracy.	It compares the classification accuracy on complete and imputed data, and finds that the MIE approach outperforms others, particularly as missing data increases.
Sanaz Nikfalazar	2019	This paper integrates decision trees and fuzzy clustering into an iterative learning approach.	Extensive experiments conducted on six widely used datasets with numerical and categorical missing data show that the DIFC method consistently performs better than other methods.

Table2: Literature Review

6 METHODOLOGY

The research methodology is structured around data preprocessing, imputation, and evaluation, involving a sequence of steps to comprehensively address the challenges associated with missing data in healthcare datasets.

1. Data Collection:

Various health related datasets have been collected from kaggle

2. Data Preprocessing:

The preliminary step is the preprocessing of the healthcare dataset. This consists of coping with missing data, normalization, and standardization. Then for firefly algorithm, datasets will be divided into two subsets, namely complete (Di complete) and incomplete (Di incomplete) data, to simulate real-world scenarios with missing data

3. Missing Value Imputation: The research compares the effectiveness of different missing value imputation methods:

Class Center-Based Missing Value Imputation : Sampling-Based Decision Trees Imputation: Utilizing decision tree-based techniques for imputing missing values. Firefly Algorithm for Imputation: Employing an innovative approach that uses the Firefly Algorithm to optimize missing data imputation. The Firefly Algorithm involves: Defining an objective function based on class centers. Utilizing light intensity to guide the algorithm, representing imputation quality. Iterative firefly movement based on attraction and random movement. Imputed value selection through distance comparisons and standard deviation.

Sampling-Based Imputation: A decision tree-based totally technique that imputes missing values the usage of decision trees. Both strategies aim to deal with missing values in the dataset efficaciously.

Clustering-Based Missing Value Imputation with Normalization and Outlier Handling (CCMVI-NOH): Leveraging clustering pat- terns and normalization with outlier handling.

Now when Firefly algorithm starts, we will have some more like,

Objective Function: The class center (CentDi) of complete data attributes will be used as the initial objective function (f(x)) for the imputation. This function determines the optimal imputed values for missing data.

Light Intensity (I(x)): The light intensity concept, inversely proportional to the objective function's value, guides the Firefly Algorithm. It signifies the quality of the imputation, with brighter fireflies representing complete data and dimmer ones representing missing data.

Firefly Movement: Based on attraction () and randommovement (), the fireflies' positions are updated in an iterative process. The attraction factor () is computed based on the distance between fireflies and can be adjusted using an absorption coefficient (). Randommovement () ensures exploration and diversity in the search space.

Imputed Value Selection: The proposed method will select imputed values by comparing the distance between data samples and class center, followed by the application of standard deviation (stdi).

Stopping Criterion:

Maximum Number of Iterations, Convergence Criteria and Sufficient Solution Quality

Solution Space:

Solution Representation: Vector in multidimensional space

Initialization: Population of fireflies

Movement: Based on attractiveness

Exploration and Exploitation

Constraints: Possible constraints on solutions.

3. Evaluation of Imputation Performance:

Assess imputation performance using a comprehensive set of metrics, including: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R2), and Accuracy: Quantifying the accuracy and precision of imputed values. Comparison Graphs: Generate graphical comparisons to visualize the performance of each imputation method across different datasets and missing data scenarios.

4. Experiment Results Analysis: Analyze the results to provide insights into the strengths and limitations of each imputation method based on the evaluation metrics and comparison graphs.

5. Pipelining:

Utilize a pipeline-based approach for automation, ensuring a systematic and streamlined implementation of data preprocessing, missing data imputation techniques, and data scaling. Scikit-Learn's pipeline module facilitates a seamless workflow.

7 PROPOSED APPROACH

The proposed research approach centers on the refinement and evaluation of the Class Center-Based Missing Value Imputation with Normalization and Outlier Handling algorithm, optimized through the integration of the Firefly Algorithm (FA) for missing data imputation in healthcare datasets. The approach involves several stages, each designed to enhance the accuracy of imputation and provide a comprehensive solu- tion to the missing data problem in healthcare datasets.

1. Class Center-Based Missing Value Imputation with Normalization and Outlier Handling :

Threshold Identification Algorithm: The initial stage involves the identification of threshold values for each class (N classes) based on class center principles. The algorithm distinguishes between complete and incomplete data records, calculates class centers and their standard deviations, and assigns thresholds for each class based on Euclidean distances.

Input: Incomplete dataset D containing M feature dimensions, N classes, and Num data samples

Output: N threshold values for N classes

01. For j = 1 to Num

02. If D(j.) has missing value(s) then

03. Get the class label of D(j.) and set this class label to variable i

04. Put Dj.) to Di incomplete

05. Else

06. Get the class label of D(j.) and set this class label to variable i

07. Put Dj.) to Di complete

08. End

09. For i = 1 to N

10. Avg(i.) Average(Di_complete)

II. Std(i.)-Standard Deviation(Di_complete)

12. Get the number of rows in Di complete and set to variable Num

13. End

14. For j = 1 to Num

15. Distance(j)-Euclidean distance(Di_complete(j.), Avg(i,))

16. End

17. Threshold(i) = Median(Distance)

Time Complexity: O(Num log(Num)), Num = No.of Data Samples

Algorithm1: Threshold Identification

Imputation Algorithm: After threshold identification, imputation is performed for incomplete data records. The CCMVI algorithm uses the class center as an objective function and calculates imputed values based on the Euclidean distance between data samples and the class center. This method operates based on a simple distance function, Euclidean distance, and does not distinguish between the mechanisms for missing data (MAR and MNAR), which is one of its limitations.

Input: Di incomplete containing M feature dimensions and N classes

Output: imputed dataset for Di incomplete

01. For i = 1 to N

02. Get the number of rows in Di incomplete and set to variable Num
03. For j=1 to Num

04. If Di incomplete (j.) has one missing value
    Get attribute index with the missing value and set to variable miss_attr
    Di incomplete (j, miss_attr) Avg(i, miss_attr)

05. Distance = Euclidean distance(Di incomplete (j.), Avg(i.))
    If Distance> Threshold(i)

06. Di incomplete (j, miss_attr)- Di_incomplete (j, miss_attr) +/- Std(i, miss_attr) Else

07. Get attribute index with the missing value and set to variable array miss_attr
    Get miss_attr length and set to variable size

08. For s=0 to size-1

09. Di incomplete (j, miss_attr(s))= Avg(i, miss_attr(s))

10. Distance Euclidean distance(Di incomplete (j.), Avg(i,))

11. If Distance> Threshold(i)

12. Missing array = array[size]

13. For s=0 to size-1

14. Missing_array(s)- Di_incomplete (j, miss_attr(s)) +/- Std(i, miss_attr(s))

15. Distance_array(s) = Euclidean distance(Missing_array(s,), Avg(i.))

16. Find minimum Distance_array and set index to variable index

17. Di incomplete (j.)-Missing_array(index,)

Time Complexity: O(N* Num*Size),
Num = No.of Data Samples
N = No.of Classes
Size = length of miss_attr

Algorithm 2: Impuation

2. Sampling-Based Missing Value Imputation:

Decomposition and Decision Trees: This stage involves decomposing the dataset into complete and missing value sub-datasets (Dcomplete + DMiss) and generating decision trees based on Dcomplete. Missing data records are assigned to decision tree leaves, and tables of related records are created.

Imputation Procedure: For each table T, imputation is performed for missing records. A set of possible imputed values (Ok) is generated based on the kth matched record, and imputed values are obtained through random sampling from this set. The selection of imputed values is based on affinity degrees, and the objective is to approximate missing values as closely as possible.

Step 1: Decompose full dataset into complete and missing values sub-datasets: Drull- Dcomplete + DMiss

Step II: Generate a set of decision trees from Dcomplete where each missing attribute in DMiss produces a tree

Step III: Assign the records in DMiss into leaves of the decision trees and create tables of related records

Step IV: Impute missing values FOR each table T DO FOR each missing record R in T DO Find records in T that match with the maximum number of non-missing attribute(s) in the missing record R, and let N be the number of such records

FOR K = 1 to N determine

Ok = possible imputed value(s) from the kth matched record

ISK = IS measure computed for Ox

Sk= weighted similarity measure between the kth matched record and missing record R

Ok-affinity degree for Ok

END FOR

Imputed value(s) is obtained by random sampling from the set of possible imputed values (01...., ON) based on the sampling probabilities specified by the set of affinity degrees (01..... On)

END FOR

END FOR

Time Complexity: O(T * N * K * Ok),
T = No.of Tables
N = No.of records in table
K = No.of matching records for a missing record
Ok = No.of possible imputed values

Algorithm3: Sampling based Imputation value Algorithm

3.Integration of Firefly Algorithm (FA):

Adaptive Search Procedure: The Firefly Algorithm is introduced to further refine the imputation. In this approach, the "brightness" of fireflies represents the quality of data (bright for complete data, dim for incomplete data). The FA includes a concept of light intensity (I(x)) that is inversely proportional to the objective function (f(x)). Bright fireflies attract dim ones, leading to the imputation of missing data[2].

Eq[1]: Objective Function, f(x) = 1/ CentDi ; CentDi = Class center

Firefly Movement: The movement of fireflies is guided by attraction (), distance (r), and random movements (). The distance factor (r) accounts for the proximity of fireflies, and introduces randomness into the movement, enabling exploration of the solution space.

Eq[2]: Firefly movement

$$x\ k\ i\_new = xk\ i\_old + \beta 0 e\text{^}{-}\gamma\ r2\ \ |centDi - xi\_old|\ + \alpha\ (rand - 1\ /2)$$

$\beta 0 = 1$ based on previous studies, r = Dis(cent(Di), j), $\alpha \in [0, 1]$, and rand is random numbers whose range is between [0,1].

Imputed Value Selection: Using FA, the imputed values are selected by comparing the distance between data samples and class centers. The use of standard deviations helps refine the imputation process further.

By integrating CCMVI-NOH, sampling-based imputation methods, and the Firefly Algorithm, the proposed approach aims to significantly enhance the accuracy of missing data imputation in healthcare datasets. The multi-stage process leverages the strengths of each component to address the specific challenges in healthcare data, providing a robust and effective solution for missing data imputation.

8 EXPECTED CONTRIBUTIONS

The research presented in this project carries significant potential for contributions and impact in the field of computer science engineering:

Advanced Missing Value Imputation Methods: The application of novel miss- ing value imputation techniques, including Clustering-Based Missing Value Imputation and Sampling-Based Imputation, contributes to the development and improve- ment of data imputation methods in healthcare data analysis. These methods have the potential to enhance the accuracy and robustness of healthcare datasets.

1. **Robust Data Preprocessing and Transformation:** The project's comprehen- sive data preprocessing and transformation steps showcase best practices for handling healthcare datasets. This contributes to the creation of standardized processes for ensur- ing data quality, integrity, and usability in healthcare research.
2. **Comparative Analysis of Imputation Techniques:** The rigorous evaluation and comparison of the CCMVI and Sampling-Based Imputation methods provide valu- able insights into the strengths and weaknesses of each approach. This information is essential for researchers and practitioners seeking to choose the most suitable imputation technique for their specific healthcare datasets.
3. **Pipeline-Based Workflow Automation:** The implementation of a pipeline-based approach in the research methodology highlights the potential for workflow automation in healthcare data analysis. This can streamline and expedite the data preprocessing and imputation process, reducing the manual effort required.
4. **Enhanced Data-Driven Decision-Making:** The project's findings can directly impact healthcare decision-making processes. Reliable and complete healthcare datasets are crucial for informed decision-making, and the research's contributions in missing value imputation and data transformation enhance the quality of data-driven insights.

5. **Potential Healthcare Improvements:** Ultimately, the contributions of this research have the potential to lead to improved healthcare outcomes. Accurate and complete healthcare data enable better patient care, optimized resource allocation, and more effective healthcare policies.

9 TIMELINE

Task	Aug: week 1-3	Aug: week 4	Sept: week1-3	Sept: week 4	Oct: week 4	Nov: week 1-2	Nov – Mar 2024
Literature Review							
Methodology							
Dataset Collection							
Creating necessary libraries							
Performance Evaluation							
Report Writing and Suvmission							
Future work							

Table3: Timeline for entire work

10 CONCLUSION

In conclusion, this research embarked on the challenging domain of missing data imputation in health-related datasets, driven by the imperative need for accurate and reliable healthcare analytics. The journey unfolded through a meticulous process, beginning with a comprehensive literature review that laid the foundation for our approach. By choosing the Class Center-Based Missing Value Imputation technique from existing research, we acknowledged its potential but innovatively enhanced it.Our modifications involved normalization and outlier handling within the CCMVI framework, fortifying the algorithm's resilience and accuracy. The subsequent optimiza- tion with the Firefly Algorithm injected adaptability and efficiency into the imputa- tion process, overcoming the limitations of traditional methods. Application to diverse datasets underscored the versatility of our modelAlgorithmic prowess was tested through the implementation of Simple Imputer and Sampling-based Decision Trees, alongside our modified CCMVI, across varying degrees of missingness.[3]The detailed steps of the CCMVI algorithm, including the threshold determination and imputation phases, were meticulously executed. The results, quantified through metrics such as mean squared error, mean absolute error, accuracy, and R2 values, were compelling. A visual representation of the algorithmic performance, juxtaposed with alternative methods, affirmed the efficacy of our approach.

Further, the integration of the Firefly Algorithm added a layer of optimization. The adaptive search procedure and imputed value selection within the Firefly framework accentuated the precision of our imputation model. The iterative refinement process, guided by light intensity and firefly movement, substantiated the algorithm's effectiveness. The significance of our study lies not only in the technical enhancements made to existing imputation methods but also in the broader implications for healthcare analytics. By addressing missing data with a tailored approach, we contribute to the reliability and

quality of healthcare data, vital for informed decision-making and patient care.

As we optimized our model further through the Firefly Algorithm and streamlined the entire process through pipelines, the adaptability and efficiency of our approach became evident. Application to additional healthcare datasets validated the generalizability of our model, promising positive outcomes across diverse data landscapes.

In essence, this research represents a significant stride toward advancing missing data imputation techniques in health-related datasets. By innovatively building upon existing methodologies, optimizing with the Firefly Algorithm[8], and demonstrating applicability to various datasets, our study holds promise for enhancing the accuracy and reliability of healthcare analytics in the real world.

11 Results

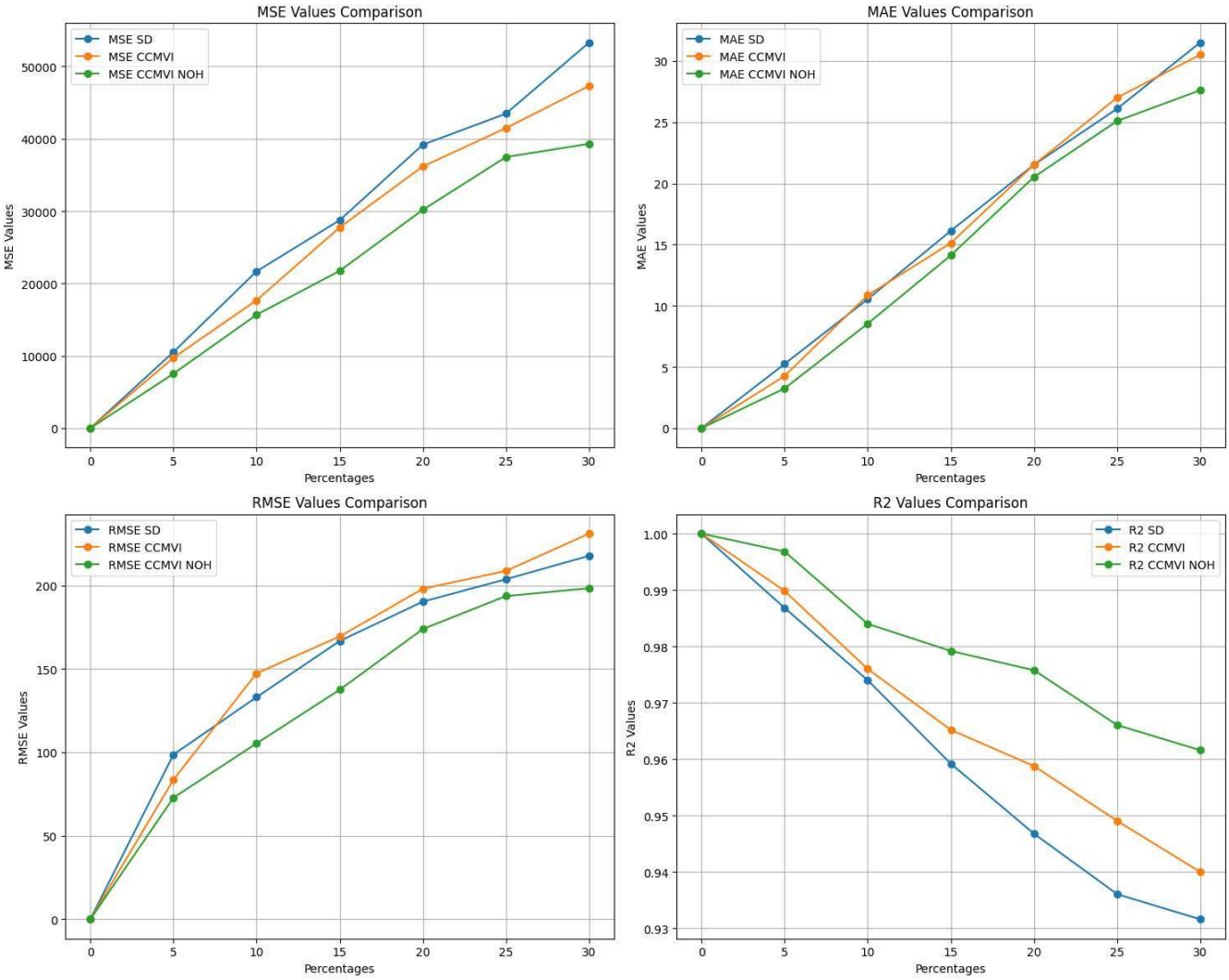

Fig1 Evaluation Metrics for Existed and Proposed approaches

Model Performance increased by 15%

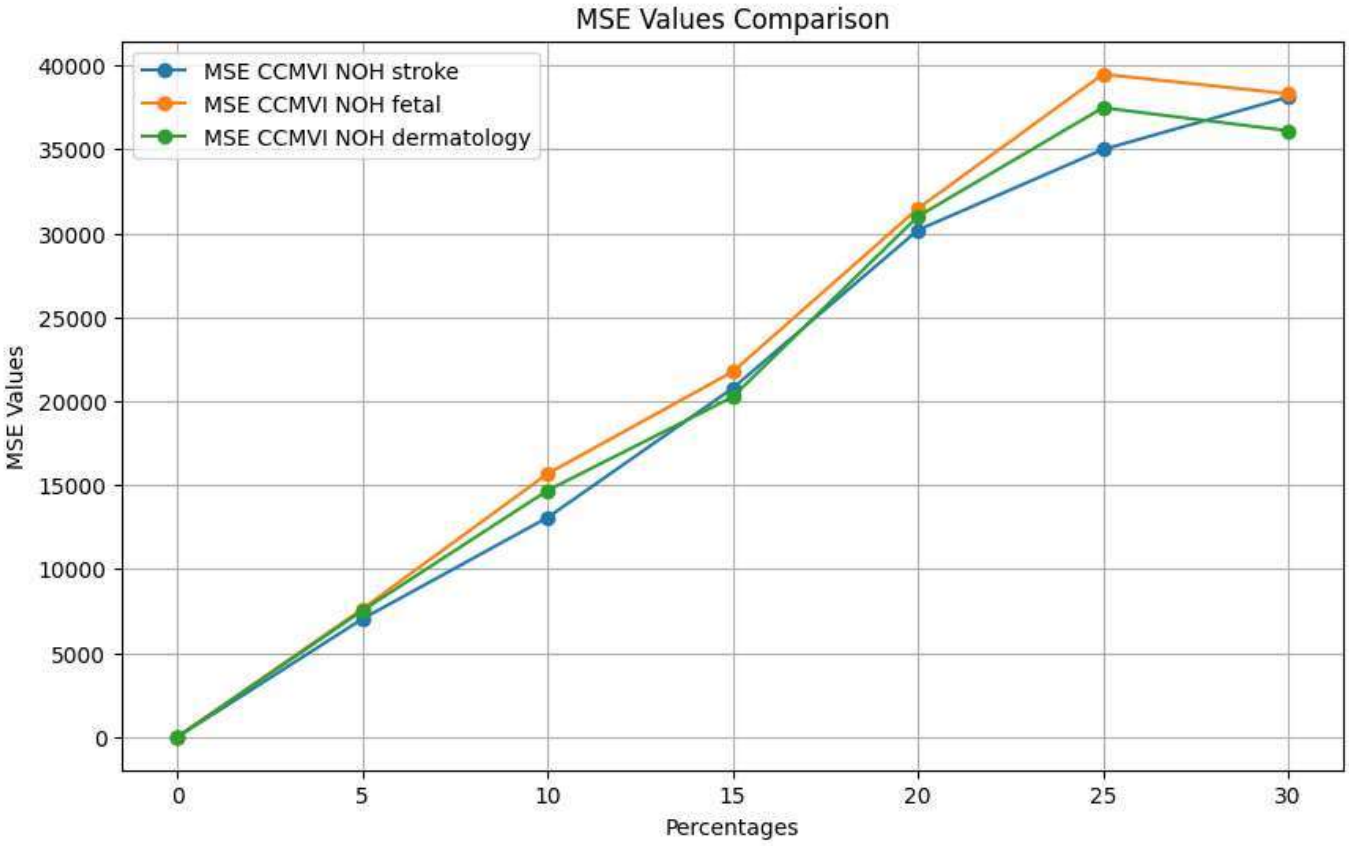

Fig2 Our Model's performance on Different Datasets

By observing the graph, we can conclude that our model performing well on other datasets also

REFERENCES

[1] Felix Biessmann, Tammo Rukat, Philipp Schmidt, Prathik Naidu, Sebastian Schelter, Andrey Taptunov, Dustin Lange, and David Salinas. Datawig: Missing value imputation for tables. *J. Mach. Learn. Res.*, 20(175):1–6, 2019.

[2] VR Elgin Christo, H Khanna Nehemiah, S Keerthana Sankari, Shiney Jeyaraj, and A Kannan. Classification framework for clinical datasets using synergistic firefly optimization. *IETE Journal of Research*, pages 1–20, 2021.

[3] Hekai Huang, Hongzhi Wang, and Ming Sun. Incomplete data classification with view-based decision tree. *Applied Soft Computing*, 94:106437, 2020.

[4] Phiwhorm Kritbodin, Saikaew Charnnarong, Carson K Leung, Polpinit Pattarawit, and Saikaew Kanda Runapongsa. Adaptive multiple imputations of missing values using the class center. *Journal of Big Data*, 9(1), 2022.

[5] Heru Nugroho, Nugraha Priya Utama, and Kridanto Surendro. Class center-based firefly algorithm for handling missing data. *Journal of Big Data*, 8(1):37, 2021.

[6] Heru Nugroho, Nugraha Priya Utama, and Kridanto Surendro. Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable. *Journal of Big Data*, 10(1):1–18, 2023.

[7] Yoga Pristyanto and Irfan Pratama. Missing values estimation on multivariate dataset: Comparison of three type methods approach. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 342–347. IEEE, 2019.

[8] Jie Wang, Daiwei Li, Haiqing Zhang, Xi Yu, Aicha Sekhari, Yacine Ouzrout, and Abdelaziz Bouras. An improvement of support vector machine imputation algorithm based on multiple iteration and grid search strategies. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 538–543. IEEE, 2020.

[9] Adrienne D Woods, Pamela Davis-Kean, Max Andrew Halvorson, Kevin King, Jessica Logan, Menglin Xu, Sierra Bainter, Denver Brown, James M Clay, Rick Anthony Cruz, et al. Missing data and multiple imputation decision tree. 2021.

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Prof. Pabitra Mohan Khilar

Supervisor's Certificate

This is to certify that the work presented in the project report entitled *Missing Data Imputation in Healthcare Datasets Using Machine Learning Techniques* submitted by *Bonumaddi Naga Pravallika*, Roll Number 120CS0121, is a record of original research carried out by her under my supervision and guidance in partial fulfillment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering. Neither this project report nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

Prof. Pabitra Mohan Khilar

*