Handling Missing Value using Multi-View Framework

Project report submitted to Indian Institute of Information Technology, Nagpur partial fulfilment of the requirements for the Award of Degree of

Bachelor of Technology
In
Computer Science and Engineering

by

Piyush Waje (BT21CSE096)

Ravi Kumar Gupta (BT21CSE125)

Kalpesh Salave (BT21CSE127)

Ansh Bansal (BT21CSE143)

Under the guidance of

Dr Suvra Jyoti Choudhary

Indian Institute of Information Technology Nagpur 441108 (India)

2021-2025

Handling Missing Value using Multi-View Framework

Project report submitted to Indian Institute of Information Technology, Nagpur partial fulfilment of the requirements for the Award of Degree of

Bachelor of Technology
In
Computer Science and Engineering

by

Piyush Waje (BT21CSE096)

Ravi Kumar Gupta (BT21CSE125)

Kalpesh Salave(BT21CSE127)

Ansh Bansal (BT21CSE143)

Under the guidance of

Dr. Suvra Jyoti Choudhary

Indian Institute of Information Technology Nagpur 441108 (India)

2021-2025

Declaration

We, **Piyush Waje (BT21CSE096), Ravi Kumar Gupta (BT21CSE125), Kalpesh Salave (BT21CSE127) and Ansh Bansal (BT21CSE143)** hereby declare that our project work titled "**Handling Missing Value using Multi-View Framework** " is carried out by us in the Department of Computer Science and Engineering at Indian Institute of Information Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution / University.

Date:

Sr.No.	Name	Enrollment No.	Signature
1	Piyush Waje	BT21CSE096	
2	Ravi Gupta	BT21CSE125	
3	Kalpesh Salave	BT21CSE127	
4	Ansh Bansal	BT21CSE143	

Declaration

We, **Piyush Waje (BT21CSE096), Ravi Kumar Gupta (BT21CSE125), Kalpesh Salave (BT21CSE127) and Ansh Bansal (BT21CSE143)**, understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations(such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what. (Source: IEEE papers) I have made sure that all the ideas, expressions, graphs, diagrams, etc. that are not a result of my own work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

I affirm that no portion of my work can be considered plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the thesis may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Date:

Sr.No.	Name	Enrollment No.	Signature
1	**Piyush Waje**	**BT21CSE096**	
2	**Ravi Gupta**	**BT21CSE125**	
3	**Kalpesh Salave**	**BT21CSE127**	
4	**Ansh Bansal**	**BT21CSE143**	

Computer Science and Engineering,

IIIT, NAGPUR

Certificate

This is to certify that the project titled "**Handling Missing Value using Multi-View Framework**", submitted by **Piyush Waje (BT21CSE096), Ravi Kumar Gupta (BT21CSE125), Kalpesh Salave (BT21CSE127) and Ansh Bansal (BT21CSE143)** in the partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering, IIIT Nagpur**. The work is comprehensive, complete and fit for final evaluation.

Date:

Dr. Suvra Jyoti Choudhary

Project Supervisor
Computer Science Engineering,
IIIT, Nagpur

Dr. Nishat Afshan Ansari

Head of Department
Computer Science Engineering,
IIIT, Nagpur

ACKNOWLEDGEMENT

"Acknowledgement is an art, one can write glib stanzas without meaning a word, and on the other hand one can make a simple expression of gratitude"

It gives us a great sense of pleasure to present the report of the Project Work undertaken during B. Tech. Final Year. We owe a special debt of gratitude to our Project Mentor Dr Suvra Jyoti Choudhary, Department of Computer Science Engineering, Indian Institute of Information Technology, Nagpur for his constant support and guidance throughout the course of our work. It is only his cognizant efforts that our endeavors have seen the light of the day.

We are very thankful to Dr. Nishat Ansari, Head of Department, for her ever-lasting support and guidance on the ground in which we have acquired a new field of knowledge. We also extend our heartfelt thanks to Dr. Tausif Diwan, Associate Dean, for his invaluable insights and support. Furthermore, we express our sincere gratitude to Dr. Prem Lal Patel, Director, for his encouragement and motivation throughout the project.

We also acknowledge with a deep sense of reverence, our gratitude towards our parents and members of our family, who have always supported us morally as well as economically.

Sr.No.	**Name**	**Enrollment No.**	**Signature**
1	**Piyush Waje**	**BT21CSE096**	
2	**Ravi Gupta**	**BT21CSE125**	
3	**Kalpesh Salave**	**BT21CSE127**	
4	**Ansh Bansal**	**BT21CSE143**	

ABSTRACT

This project investigates the study of integration and denoising techniques applied to multiple files, the use of autoencoders for dimensionality reduction, and the use of multitasking views to improve clustering accuracy. Effectiveness is measured by metrics such as the normalized mutual information (NMI) and Adjusted Rand Index (ARI), which provide a quantitative measure of convergence between different datasets.

The tests include different materials such as Cardio, Pima, Car and Wine, which deal with the effects of combining multiple agents. The retraining process optimizes the reconstructions and removes hidden layers to facilitate detailed clustering. The results show that by optimizing the agents, NMI and ARI are improved and true values are obtained for specific data in the study group.

This study demonstrates the potential of decision making and multimethod integration in a collaborative environment, providing insight into future practices and standards in data analysis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Full Name
MVF	Multi-View Framework
AE	Auto-encoder
NMI	Normalized Mutual Information
ARI	Adjusted Rand Index
KMC	K-Means Clustering
REC	Reconstruction Error
KNN	K -Nearest Neighbour
CkMeans	clustering k-means

Chapter I

INTRODUCTION

In today’s rapidly changing digital environment, information has become the lifeblood of civilization, driving progress in science, technology, medicine, and finance. The unprecedented growth in the collection and use of data is providing trans-formative insights, driving global innovation, and shaping decision-making. But despite the potential of information, its use is often plagued by competition and discontinuity: information is lost or incomplete. This problem affects data qualityand reliability, and is particularly important in high-level organizations such as medical research, financial analysis, and scientific discovery.

Incomplete information in a database can lead to incomplete analyses, negative results, and misunderstandings of important events. For example, in medical research, missing patient information can lead to misdiagnosis or poor treatment, while in finance, missing details can impact models used for forecasting and risk assessment. In areas such as climate science, the lack of critical data can lead to inaccurate predictions, affecting global policy and action. Addressing the issue of missing data is therefore not only a necessity, but also an important ethical issue to ensure fairness, accuracy, and reliability in decision making.

Routine methods for imputing missing data often involve simple methods such as discarding non-input data or using simple methods such as mean, median, or mode to impute missing values. While these methods can provide a quick fix, they are limited in preserving the complexity and structure of the original product. These deficiencies occur in the context of today's literature, which is not only large but also diverse, and involves relationships that require difficult analysis.

Given these challenges, there is an increasing need for innovative and scalable solutions that can address data loss issues while preserving data integrity and richness. Our project aims to address this need by implementing an advanced MVF that provides a meaningful framework for analyzing and processing missing data. The framework combines various imputation techniques (such as mean, median, and mode) to estimate missing values and then uses KMC to cluster and discover underlying patterns among items. By combining imputation and clustering in MVF, we aim to increase data integrity and reliability, enabling effective analysis even in the face of significant data.

What makes this approach unique is its ability to leverage different sources of information using interactions and relationships to achieve greater impact. After assignment, KMC is used to group similar points into clusters, revealing insights that may not be obvious. This two-step approach not only improves quality, but also provides insights, making it powerful for researchers and analysts. Importantly, MVF is scalable, making it suitable for large and complex datasets and diverse modern information ecosystems.

Our project holds immense potential to transform the way incomplete data is handled across diverse fields. By improving the accuracy and reliability of datasets,it paves the way for more robust analyses, more reliable predictions, and ultimately better decision-making. In medical research, this approach could improve patient outcomes by enabling more precise diagnoses and treatments. In finance, it could enhance risk modelling and forecasting accuracy, while in scientific research, it could ensure the integrity of experimental findings. This innovation represents not just a technical contribution but a critical step toward addressing one of the most persistent challenges in data-driven disciplines.

1.1 Motivation for the Work

The motivation for this project stems from several key challenges and opportunities within the context of handling missing values:

Impact of missing data: Missing data can impact the integrity and efficiency of the review process, especially in important areas such as clinical, finance and research where accurate information is essential for decision making.

Limitations of Imputation Methods: Traditional methods such as dropping missing data or using simple imputation methods often fail to capture complex missing data, resulting in lack of analysis and reduced data reliability.

Advances in MVF: The advent of MVF technology provides the opportunity to resolve missing data by integrating multiple data views, thereby increasing accuracy and checking for good information.

Enhanced clustering capability: With advanced integration such as KMC powered by AE based feature extraction, it can improve clustering performance even with missing data, making it more powerful and more focused.

Inspired by these challenges, our project aims to create a scalable and effective solution to solve missing data by combining MVF, imputation method and group analysis, ultimately improving the quality and reliability of data-driven applications.

1.2 The Objective of the Work

The main objective of this work is to design and implement an efficient MVF to handle missing data in heterogeneous databases. The specific objectives of the project are:

Interpolation technology: Use AE and KNN methods to accurately predict missing values in datasets with varying degrees of missing data.

Cluster analysis: Use KMC to analyse the impact of different strategies on cluster results and use NMI and ARI to measure performance.

Performance measurement: Measuring the difference between the original data and the reconstructed data by calculating the REC to measure the accuracy of the intervention.

Data analysis: Use the framework of different data such as Iris, Wine, and Car across various missing data to ensure robustness and generality.

1.3 Scope and Significance of the Work

This work involves the development of a complete MVF to handle missing data across multiple datasets to ensure the reliability of the underlying machine learning application. Using advanced imputation techniques, this work enables researchers and practitioners to:

Improving data quality: Estimating missing values to ensure that the data set for machine operation is complete and reliable, reducing the risk of drawing error being correct or incorrect.

Increase the accuracy of analysis: Advanced techniques like AE and KNN increase the accuracy of clustering and classification models, resulting in greater insights.

Broader applicability: The framework has been tested on different datasets, demonstrating its effectiveness and adaptability to real-world racing datasets.

Less computing overhead: With effective integration and integration, the framework keeps the computing costs low while managing the operations, making it suitable for big data analytics.

Significance

The importance of establishing a robust MVF to deal with missing data is that it plays a key role in ensuring the reliability and accuracy of datasets for ML operations. By solving the problem of missing data, the project eliminates the need for manual decision-making processes, thereby increasing the efficiency and accuracy of decision-making information. This is especially important in areas where accurate and complete information is required, such as healthcare, finance, and research.

The success of the program also has important implications for the broader field of machine learning and data science. This method supports the best method for handling missing data by demonstrating the utility of AE-based reconstruction and KNN interpolation. It provides a flexible and flexible method that can be used across different files and directories. Finally, the development of this MVF represents an important step in making ML applications reliable, enabling innovation in data analysis, and facilitating the initial development of valuable information.

1.4 Organization of the Report

The remainder of this report is carefully organized into different sections, each with a specific purpose to document and detail various aspects of the project. These sections include:

Literature review: In-depth analysis of existing MVF methods, including AE, KNN, and other state-of-the-art interventions, as well as evaluation methods such as NMI and ARI.

Completed work: Expanding the use of imputation methods, integrating machine learning functions such as clustering, and developing a framework for handling missing data across diverse datasets.

Results and evaluation: Presenting the results from many tests and evaluations, analyzing the accuracy of the decisions (using REC), common operations and extending the framework to many documents.

Discussion: Analyze the results and their significance, discuss the effectiveness of the proposed approach, the challenges encountered and future developments.

Conclusion: summarize the main points of the project about the program for-working with missing data and the importance of the proposed MVF.

Evidence: Include a bibliographic summary of each source, providing an overview of the data reviewed and cited throughout the report.

Chapter II

LITERATURE REVIEW

Literature Review

In the process of creating an MVF guideline for handling missing data, we carefully reviewed the existing data in this area. The purpose of this review is to understand the current state of the research, identify common practices, and identify gaps that our work can address. We focus on studies that use techniques such as AE, KNN, as well as common metrics such as NMI and ARI to measure performance.

1. Reviewing Auto-encoders for Missing Data Imputation [1-3] :

This research paper aims at reviewing Auto-encoders (AEs), Denoising Auto-encoders (DAEs), and Variational Auto-encoders (VAEs) in the context of missing data imputation in tabular datasets . When data is missing which is usual in big data some machine learning algorithms reduce efficiency in their performance. To address this issue, auto- encoders used learned feature space of the missing entries from the incomplete data and generated informative values for the missing entries. DAEs that leech out the input data during learning and are designed to recreate the clean data outperforms the baseline strategies such as mean or median imputation particularly when the data is under sorted by a complex dependence model. VAEs, which have been introduced in the section, expand this possibility by predicting probabilities of data distributions to produce statistically consistent samples. To this end, it analyses 26 studies from 2014 to 2020 on architecture, hyper-parameters, and training techniques of Auto-encoders and compares them with other imputation methods. The results present DAEs as particularly suitable, aligned with the suggestions for enhancing network construction, including the application of regularization methods and symmetrical models.

2. A Study of K-Nearest Neighbour as an Imputation Method [4] :

In this research paper, the use of KNN approach for managing missing data in machine learning. This makes missing data a systematic problem that may significantly threaten the goodquality of these models making it difficult for the machine learning modelsto perform optimally. Thus, this research situates KNN as reasonable imputation technique that retains the pattern and dependency structures in data. KNN eliminates missing value by identifying the nearest neighbours of instances with missing data values based on the predefined distance measurement. The authors compare the KNN over other basic statistical methods of imputation like mean and or median calculation pointing out how the former reduces the effects of bias on finalanalysis and maintain intact the distribution of data. Moreover, the formulapresents a critical analysis of such other variables asthe value of k, distance measurement, and how categorical data are handled, and includes a comparison of the algorithm's ability to work wellacross different types of databases.

3. A Framework for Multi-view Clustering with Missing Values [5-9]:

This work focuses on a major challenge with Multi-view Clustering (MVC) referred to as the Partially Data-missing Problem (PDP).). Conventional MVC patterns entail full data which is always a problem since the assumption is rarely realistic owing to problems in data acquisition and transfer. Traditional approaches involve filling in empty data before applying MVC techniques; however, this results in model performance degradation due to errors in imputation. To address these limitations, the authors present a new imputation-free approach based on a matrix correction mechanism by incorporating a two-step method known as correction-clustering. In the first stage, distance matrices obtained from…incomplete data are adjusted to obtain affinity matrices. In the second stage, these corrected matrices are incorporated naturally into a family of affinity-based MVC methods. This method helps to avoid problems attributed to imputations hence improving the clustering precision and reliability . This is as proved in experimental results which show that the suggested framework yields better clustering results than anyof the traditional imputation-based methods no matter the level of missing data. This study provides a significant advancement in the field of MVC by introducing a novel and effective solution to handle incomplete datasets while maintaining high clustering performance .

4. Multi-View Clustering via Canonical Correlation Analysis [10-11]:

M. Faraji, A. Kalhor, and M. Moradi have named their research work "Multi-View Clustering via Canonical Correlation Analysis", and in this research paper, a new comprehensive method of implementing clustering for high-dimensional data using multiple views has been outlined. Many conventional clustering techniques like k-means are badly suited to handle a problem from this domain, and after extracting the feature vector, one has to perform linear dimensionality reduction such as PCA. However, all these methods work under certain separation conditions on the means of clusters for optimal functionality. In this work, Canonical Correlation Analysis (CCA) is employed in order to map data in each view into a lower dimensional space. By analyzing the correlated dimensions, which are assumed to contain cluster identification information, the proposed method allows a less strict separation of clusters than the actual prior techniques in terms of the above stated conditions. The research is validated through experiments on two domains: clustering of speech audio with the facial images of the speakers, and clustering of text with link data in articles of Wikipedia. The paper alsoenumerates distinct issues of these domains including there being more than one cluster variable and the issues of hierarchical clustering. This work is truly a substantial contribution to the current progress in multi- view clustering as it identifies critical weaknesses and employs useful solutionsin actual datasets.

5. Fuzzy Clustering of Single-View Missing Data Using a Multiview Framework [12-13] :

The paper "Fuzzy Clustering of Single-View Incomplete Data Using a Multiview Framework" which will be used to support this paper sheds light on the problem of clustering data with missing values using a Multiview framework. The authors present four frameworks beginning with a basic imputation strategy to build mi versions of the working datasets. These imputed views are clustered together using two primary clustering methods: there are more than one such extension, the first being a multi- view version of Fuzzy-c-Means called MVFCM and the second being a kernelized version called MVKFCM. The framework incorporates the entropic regularization term for weighing diverse views to validate that theirrelevance has proper weight. The final imputation is therefore done as a convex combination of the imputed values across the various views, the final imputed set is clustered. The performance of the proposed algorithmsis measured using NMI, ARI and clustering accuracy on 12 benchmark data sets. Analysis of results show that MVKFCM performs better over other methods with slight fluctuation between iterations, though has a homogeneous trend. Addressing challenges such as unknown or large class counts (r), the authors propose two MVKFCM variants: We have fixed views MWKFCM-FV and robust fixed views MVKFCM-RFV. The RFV variant focuses on view generation for stable performance and proves better results for handling unusable data. This work contributes considerably to fuzzy clustering assessment by overall combining some elements of multi-view framework for handling of missing data and enhancement of clustering performance. The literature review revealed that methods like AE and KNN are widely recognized for their ability to handle missing data effectively. Metrics like NMI, ARI, and REC are essential for evaluating the performance of imputation and clustering tasks. These studies also highlight the critical importance of selecting appropriate imputation techniques based on datasets characteristics and the nature of missingness.

6. Multi-View Cluster Analysis with Incomplete Data to Understand Treatment Effects [14]:

The paper Multi-View Cluster Analysis with Incomplete Data to Understand Treatment Effects points to one of the major issues in MVCA, namely when data entries are missing in one or more views. Multi-view clustering is one of the most common paradigms in granular computing; its goal is to group the subjects uniformly within several views in which they are described. Nonetheless, most of the existing multi-view coculturing approaches are unsuccessful in dealing with missing values particularly when the missing pattern is irregular across views. This work introduces improved approaches for multi-view coculturing each of whichis based on an indicator matrix that pinpoints the positions of observed dataentries, thereby reducing the validity assessment to the actual observed dataonly. Thus, the proposed method is superior to less complex approaches, including methods of handling missing data such as recalculating coefficients without subjects containing missing values or imputing missing values of given data, since theybring about uncertainty and bias. Inthis way, new method is less sensitive to errors of imputation and is more reliable, especially when values are missing in some views or when some views are completely missing. The finite mixture methods are shown to compare well in simulations and are used in a treatment study of heroin dependence, where incomplete data would pose a problem for other methods. The study also reveals that this approach makes it possible to establish subgroups ofpatients with similar characteristics across the time windows whereby pre- treatment characteristics can be used for post-treatment characteristics. Compared to the previous literature, this research presents a sound and stable way to cluster the multi-view data with missing values thus being appropriate for real world scenario with distinct missing values patterns.

7. A way to obtain the quality of a partition by Adjusted Rand Index [15]:

Being a statistical measure, the Adjusted Rand Index (ARI) measures the extent to which the clustering performed matched the true clusters while correcting for chance. The ARI was applied to compare two clustering algorithms: FCM and CkMeans, ona selection of the Vowels dataset, that is part of the validated Letters data set. The dataset consisted of instances of the five English vowels (A, E, I, O, U) with numerical features consisting of statistical moments and edge count of distorted letter images.To keep things fair, both algorithms were trained using the same set of parameters: a fuzziness factor (m) of between

1.25 and 2.5, and initial starting centroids as well as membership values were randomized but the same across trials. The results indicated that for 12 of the 16 settings of the fuzziness parameters, CkMeans performed slightly better than FCM, based on predominantly higher ARI. For example, with fuzziness at 2.5, CkMeans produced a median of ARI 0.73, on the other hand, FCM was slightly behind with 0.51. Additionally, CkMeans demonstrated better clustering accuracy, as indicated by fewer misclassified instances: 1065 plans to 1515 in FCM which corresponds to 41.64% error rate in 1,615 and 42.67% error rate in FCM 1,655. These results demonstrate how the modification of CkMeans improves the algorithm's capacity to form clusters that are organized according to the inherent structure of the data. The study focuses on the practicality of ARI for validation in different clustering techniques with considerable stress on the datasets containing clearly known, externally defined classes. This collaborates the use of ARI in the validation of clusters, generate higher quality and relevant assessment of the algorithm's performance.

Our work builds on these findings by presenting an MVF that combines AE and KNN methods. We further evaluate the performance of the framework on different datasets using integration and reconstruction error metrics. We believe that our services support the state of-the-art in handling missing data and provide practical tools for real-world machine learning.

In the next section, we discuss the system in detail, showing the imputation process used, the data used, and the steps taken to evaluate the effectiveness of the system. This approach ensures that our study builds on existing research and provides new insights into the field.

Chapter III

PROPOSED WORK

This chapter focuses on addressing missing values through a Multiview clustering framework using auto-encoders. It covers the datasets utilized for training, validation, and testing, along with preprocessing strategies to handle inconsistencies and optimize data representation. The methodology includes model selection with an emphasis on auto-encoders for feature extraction and clustering mechanisms adapted for Multiview data. The chapter also highlights training processes, evaluation techniques, and the integration of clustering frameworks. Evaluation metrics, such as Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), are employed to measure the effectiveness of the approach, ensuring robust and meaningful clustering results.

3.1 Datasets Used

Dataset Name	Total Instances	Total Feature	Classes
Iris	150	4	3
Wdbc	569	30	2
Wpbc	194	34	2
Cardio	2126	21	3
Vehicle	846	18	4
Pima	768	8	2
Glass	214	9	6
Wine	178	13	3
Pendigit	10992	16	10
Balance	625	4	3
Seeds	210	7	3

Table 3.1 : Dataset Summary

3.2 Data Preprocessing

The initial phase of the project involved loading datasets like Iris, WDBC, Pima, and Balance into the workspace and formatting them according to established standards. Each dataset was analyzed to assess its structure, identify missing values, and summarize its features and classes, providing insights into its scope and complexity. To address missing values, multiple imputation strategies were applied. In View 1, missing values were replaced with the mean, in View 2 with the median, and in View 3 with the mode. Views 4 through 10 used the K-Nearest Neighbors (KNN) algorithm with k=1,3,5,7,9,11, and 13, respectively, offering diverse perspectives. These views were then consolidated into a single, comprehensive dataset for further analysis.

To enhance data quality, an autoencoder-based imputation method was implemented, leveraging its ability to reconstruct data while preserving structure, ensuring consistency and accuracy. Following this, Multiview clustering was conducted to analyze the datasets from different feature subsets, enabling robust groupings and utilizing the full scope of features. This comprehensive approach ensured a thorough understanding of the datasets and prepared them for subsequent analysis.

How Users Benefit from the Solution

Our project directly addresses these issues by providing a powerful and enabling framework that is useful for users across a wide range of industries. The solution combines the cutting-edge technologies of MVF with advanced computational methods (mean, median, and mode) and cluster analysis using KMC to provide comprehensive guidance for handling missing data. The goal of this approach is to increase the completeness, accuracy, and reliability of the dataset, ultimately enabling users to make more informed decisions.

Here’s how different users benefit from this innovation.

1. Medical Researchers and Healthcare Providers :

Often, there are gaps in financial information due to delayed reporting or incomplete transactions. For financial analysts, the ability to estimate missing values enables greater risk management, optimization, and forecasting. Improved dataset quality reduces forecast bias, allowing businesses and policymakers to make informed decisions supported only by effective and reliable data.

2. Financial Analysts and Economists:

Often, there are gaps in financial information due to delayed reporting or incomplete transactions. For financial analysts, the ability to estimate missing values enables greater risk management, optimization, and forecasting. Improved dataset quality reduces forecast bias, allowing businesses and policymakers to make informed decisions supported only by effective and reliable data.

3. Scientists and Environmental Researchers:

In scientific research and climate modeling, missing data can hinder research results or impede the development of predictive models. With the ability to address missing data, our solutions enable more accurate simulations and analysis. For example, climate scientists can better predict weather and environmental changes and help inform international policy decisions to combat climate change.

4. Business Intelligence Analysts:

Companies that handle customer data, sales, or performance reviews often experience data loss due to errors or incomplete user data. Using this framework, businesses can retrieve missing data to improve demand, customer behavior analysis, and strategic planning. This improves decision making and increases customer satisfaction.

5. Data Scientists and Analysts:

Missing data is a challenge in data science that impacts the performance of machine learning models and statistical analyses. Our solutions provide analysts with easy-to-use and reliable tools to optimize data prioritization, resulting in better modeling and insights. This is especially useful in areas with large and complex data that traditional methods cannot measure.

3.3 Dataset Splitting

In the dataset splitting phase, the code splits the dataset into three subsets: training set, validation set, and test set. This segmentation is important for model training and quality assessment. It allows the model to learn the model's interactions and construction through the training process, optimize its performance based on recommendations, and extend the testing process by evaluating its capital quality.

This information is often used to resolve missing values in the multitasking view (MVF) and shared to facilitate dynamic and collaborative work. Each subset is carefully designed to serve a different purpose in the prioritization and training pipeline models.

Based on the organization of the datasets used for training, the catalogues were carefully designed to preserve the quality of the datasets. Three main concepts were created: train, val, and testing ; each of them plays a specific role in the training of the pipeline. The training maps contain the data used to train models such as auto-encoders (AEs) used to predict missing values by learning from a sample of the input data. The validation map (val) contains the information used to optimize the hyperparameters and evaluate the quality of construction during training. Finally, the tests contain the data to evaluate the performance of the model in filling in missing values for missing inputs.

Data is divided into these lists according to a predefined ratio (e.g. 70% for training, 15% for validation, and 15% for testing). Merge to fit images or features to organize the data and create a graphical representation (like Pandas Data Frame). This data frame has good content such as good results, labels, and all prediction inputs, increasing ease of access and management.

In addition, label data and metrics such as Mutual Information Index (NMI) and Adjusted Rand Index (ARI) were compiled to assess the quality of the non-cost evaluation and a good performance in the group after the first stage. These metrics provide many insights into the effectiveness of the interaction model, especially when techniques such as AE that minimize reconstruction effort (REC) are used.

Additionally, a dataset has been created that describes the approach to the training, validation and test dataset, including the necessary properties for the imputation process and later review.

Generally, the organization takes great care to ensure that the data is well-established and ready for the imputation process after reviewing various functions. It provides a solid foundation to manage missing data, making machine learning accurate and reliable.

3.4 Methodology

In many research projects, preliminary data generation is important to ensure the accuracy and reliability of subsequent machine learning tasks, such as clustering and imputation studies. The pre-existing pipeline consists of processes designed to improve the quality and consistency of input data for further analysis.

First, an initial multi-image dataset is input to the system. These data often present inconsistencies and missing values across different viewpoints; hence, a strong first step is required to normalize and optimize them. Autoencoders (AE) are used to effectively handle missing values. AE reconstructs the input data by learning the image representation that minimizes the reconstruction error (REC) to accurately estimate the input parameters. This reconstruction process ensures that complete data is prepared for subsequent reduction and analysis.

After dealing with missing data, principal component analysis (PCA) was used to reduce the rest of the dataset. PCA transforms the data into a low-dimensional space that captures the most significant changes in the features. These steps serve to reduce computational complexity and facilitate integration, efficiency, and effectiveness.

Then, K-means clustering (KMC) was performed on the previous data. This algorithm helps to separate the content into groups based on their similarities. KMC minimizes the differences between different clusters by optimizing the location of cluster centers and provides useful information. The number of groups (k) is determined according to the specific requirements of the analysis.

To facilitate efficient data management and analysis, data processing, including summary and measurement results, is built into Pandas Data Frames. This tabular representation makes it easy to manage, filter, and analyse group results, allowing researchers to draw the desired conclusions from the data.

Finally, the pre-processing, the group's results, and the evaluation scores are recorded in CSV (Comma Separated Values) format. This format is widely recognized for its ease of use and compatibility with various data analysis tools and methods. It creates contradictory effects, allowing researchers to conduct in-depth analysis and gain insight into multiple perspectives.

More importantly, pipelines are the backbone of many monitoring groups, providing a solid foundation for accurate and reliable operations. By taking each step forward carefully, researchers can unlock new possibilities for innovation and advancement in important areas of machine learning.

In the data preprocessing stage, several simple tasks need to be performed to increase the efficiency of the next process. Initially, the imputation value was performed using AE, which reduced the RECs and filled the gaps in the data. Then, PCA was used to reduce the dimensionality and focus on the analysis of the most important components.The KMC algorithm is used to identify points in the data and improve the clustering results based on similarity measures.

Integration was evaluated using the integration index (NMI) and adjusted Rand index (ARI). These metrics provide a quantitative measure of cluster performance by comparing the predicted cluster with the actual cluster to measure the effectiveness of the algorithm. Each step before optimizing the input data is a correct and more effective preparation of the monitoring cluster and a solid foundation for subsequent ML projects.

3.4.1 Architecture Used :-

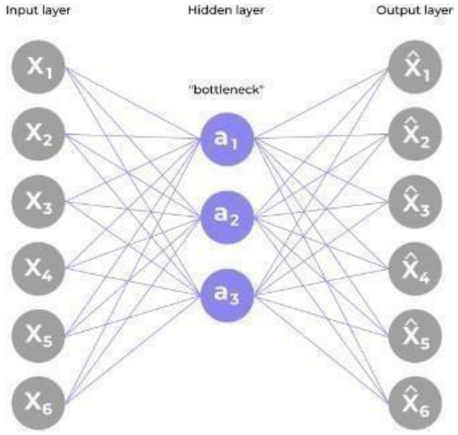


Figure 3.2 : Autoencoder Working

3.4.1.1 Encoder

Input layer take raw input data.

The hidden layers progressively reduce the dimensionality of the input, capturing important features and patterns. These layers compose the encoder. The bottleneck layer (latent space) is the final hidden layer, where the dimensionality is significantly reduced. This layer represents the compressed encoding of the input data.

3.4.1.2 Decoder

The bottleneck layer takes the encoded representation and expands it back to the dimensionality of the original input.

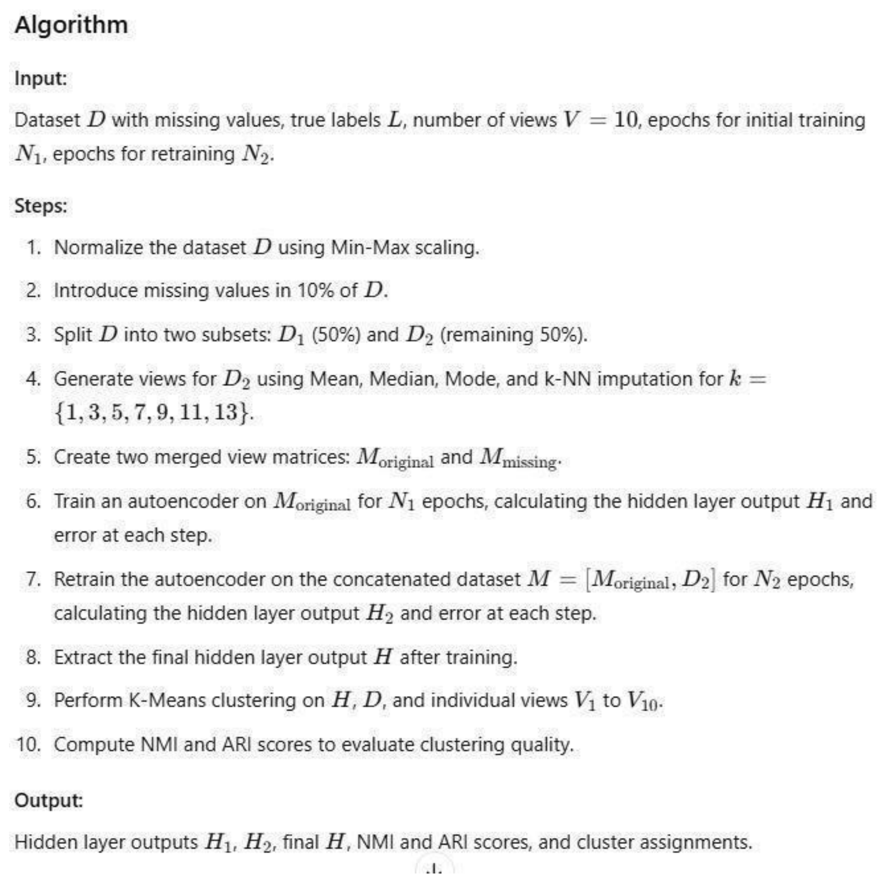
The hidden layers progressively increase the dimensionality and aim to reconstruct the original input.

The output layer producesthe reconstructed output, which ideally should be as close as possible to the input data.

The loss function used during training is typically a reconstruction loss, measuring the difference between the input and the reconstructed output. Common choices include mean squared error (MSE) for continuous data or binary cross-entropy for binary data.

During training, the auto-encoder learns to minimize the reconstruction loss, forcing the network to capture the most important features of the input data in the bottleneck layer.

3.4.2 Algorithm :-



Algorithm

Input:

Dataset D with missing values, true labels L, number of views $V = 10$, epochs for initial training N_1, epochs for retraining N_2.

Steps:

1. Normalize the dataset D using Min-Max scaling.
2. Introduce missing values in 10% of D.
3. Split D into two subsets: D_1 (50%) and D_2 (remaining 50%).
4. Generate views for D_2 using Mean, Median, Mode, and k-NN imputation for $k = \{1, 3, 5, 7, 9, 11, 13\}$.
5. Create two merged view matrices: M_{original} and M_{missing}.
6. Train an autoencoder on M_{original} for N_1 epochs, calculating the hidden layer output H_1 and error at each step.
7. Retrain the autoencoder on the concatenated dataset $M = [M_{\text{original}}, D_2]$ for N_2 epochs, calculating the hidden layer output H_2 and error at each step.
8. Extract the final hidden layer output H after training.
9. Perform K-Means clustering on H, D, and individual views V_1 to V_{10}.
10. Compute NMI and ARI scores to evaluate clustering quality.

Output:

Hidden layer outputs H_1, H_2, final H, NMI and ARI scores, and cluster assignments.

3.5 Training

In the model training phase, the dataset is processed and trained using the provided configuration settings. This process begins by initiating the training through the execution of a function or command that orchestrates the training process by leveraging the specified parameters. The configuration settings include paths to the dataset and features, the number of epochs, batch size, and other essential training parameters.

For example, when training a model, the data input (X) and labels (y) are prepared, normalized, and split into training and validation sets. The model configuration, including algorithms like KNN Imputer or any other strategy for missing data imputation, plays a critical role in preprocessing the data. The training process involves feeding the data through the model, adjusting weights using optimization techniques, and iterating over several epochs to refine the model's parameters.

The epochs parameter determines the number of iterations through which the entire dataset is processed, allowing the model to learn and update its weights. Each epoch enables the model to better understand the patterns in the data, progressively improving its performance.

The batch size specifies how many samples are processed in one pass during training. A larger batch size can speed up training but might require more computational resources, while a smaller batch size might provide a more memory-efficient process but may take longer to converge.

By carefully configuring the parameters and running the training process, researchers ensure the model learns from the dataset and adjusts to produce accurate results, minimizing errors over time. This process is essential for refining the model's performance, making it more reliable and effective for future predictions or tasks.

This iterative training process not only improves the model's accuracy but also strengthens its robustness, ultimately ensuring that it performs optimally on the given task.

3.6 Model Evaluation

The model evaluation stage also a significant step in determining the success of imputation as well as clustering method adopted in handling the datasets with missing value. The use of the above-mentioned evaluation techniques ensures that this process is comprehensive with regard to the accuracy of the methods used, and their generalization capabilities. For clustering tasks, there are two indices: Normalised Mutual Information (NMI) and Adjusted Rand Index (ARI), especially for further evaluation of the clustering quality. NMI measures the similarity between the encompassed predicted clusters and the actual true labels except for the number of clusters that are encompassed. When the value is close to 1, it means that majority of the predicted and true labels agree and hence, applied clustering method has been able to reveal majority of the inherent structure of given data set (Strehl & Ghosh, 2002). On the other hand, ARI matches the predicted clusters with true labels as it is possible that the clusters are coincidental. Since high ARI values can be attained by noise, Larsen and Sheaffer (1989) proposed using the normalized version of the ARI which corrects for the overestimation of the ARI for large values of diameter that may be due to random groupings that are not genuine. In fact, besides ARI, another measure is NMI, which also help to estimate the quality of clustering as compared to real partitions of data.

In the context of imputation assessment, a range of approaches can be measured to compare the imputation performance of a given imputation technique, including mean imputation, median imputation, mode imputation, as well as KNN imputation. To assess the efficiency of these techniques, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics are used and computed on the basis of the comparisons made between imputations and the original complete set of data. MAE is simple that measure the average deviation of the imputed values with the true values. The number bellow MAE suggest higher ability to give accurate estimates of missing values from the original sequences. In the same way, a larger error is given higher importance in RMSE than in MAE, making it a more suitable measure where the imputed values are can greatly depart from the actual values. RMSE is therefore most relevant when precision is vital as RMSE entails a weighted measure of the imputation method error (Chawla et al., 2003). Like any other evaluation measures, both MAE and RMSE enable the user to understand various aspects of how the imputation techniques perform in practice. In terms of imputation performance, different imputation techniques such as mean, median, mode, and KNN-based methods are evaluated for their ability to recover missing values. To assess this, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are calculated by comparing the imputed values to the actual original data. A lower MAE indicates better performance in recovering missing data, as it represents the average absolute difference between the original data and the imputed values. Similarly, RMSE emphasizes larger errors and provides a weighted measure of the imputation method's accuracy.

The evaluation procedure usually starts with clustering analysis on the given data set and the use of the K-means algorithm to generate clusters of data; this step is then compared with the actual or real clusters using the Clustering Accuracy, NMI and ARI. This step enables one to determine some extent to which clustering model relates with the real structure of the data. When we obtain the clustering results the subsequent work is to apply different methods of imputation of missing values. The above imputed values are then compared with the original data set using MAE and RMSE so as to have rather clearer picture of how better each of the imputation technique performs in terms of filling up the missing information. Such a two-step analysis confirms that both clustering and imputation techniques are evaluated for their overall performance regarding missing data and clustering. From this broad- ranging assessment approach, one is able to learn what each method has to offer in terms of advantages and disadvantages. Thanks to the results of these metrics, one can determine which clustering technique and how many missing values imputation iterations is the most efficient. This helps the researchers and practitioners who work on the various datasets to decide on which of the methods is more appropriate to use in order to arrive at the best results as used on real problem.

Chapter IV

RESULTAND DISCUSSION

The performance evaluation highlights the robustness of the multi-view learning approach employed in the experiments. Models were trained with varying configurations across several datasets, including Iris, WBPC, Wdbc, Balance, Cardio, PIMA, Seeds, Glass, and Vehicle. The analysis was based on clustering metrics such as Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) to assess clustering quality.

Datasets	**Hidden Nodes**	**Missing Percentage**	**Learning rate**
Iris	20	10	0.01
Wbpc	20	10	0.01
Wdbc	20	10	0.01
Balance	20	10	0.01
Cardio	20	10	0.01
Pima	50	10	0.01
Seeds	50	10	0.01
Glass	50	10	0.01
Vehicle	50	10	0.01
Wine	500	10	0.1
Pendigit	500	10	0.1

Table 4.1 : Hyperparameter Configuration for Various Datasets

Dataset	Iris		WBPC		Wdbc		Balance		Cardio		PIMA	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
Hidden	**0.7900**	**0.723**	0.0370	-0.0154	**0.5889**	**0.6877**	0.171	0.2160	**0.3409**	0.1576	0.1062	**0.1441**
Original	0.6758	0.600	0.0290	**-0.0137**	0.5758	0.6917	0.200	**0.2209**	0.2986	**0.2106**	0.1063	0.1429
View1	0.6745	0.621	0.0274	-0.0158	0.5653	0.6577	0.132	0.1648	0.0847	-0.0327	0.0851	0.1257
View2	0.7228	0.658	**0.0382**	-0.0141	0.5440	0.6317	0.082	0.0984	0.1012	-0.0188	0.0922	0.1384
View3	0.7077	0.630	0.0274	-0.0158	0.5771	0.6799	0.114	0.1308	0.1012	-0.0188	0.0960	0.1390
View4	0.6225	0.586	0.0360	-0.0160	0.5673	0.6837	0.109	0.1273	0.1024	0.0212	0.0966	0.1364
View5	0.6804	0.622	0.0231	-0.0143	0.5889	0.6877	0.109	0.1292	0.0935	0.0089	0.0801	0.1189
View6	0.7573	0.661	0.0360	-0.0160	0.5769	0.6761	0.129	0.1575	0.3391	0.1649	0.0874	0.1282
View7	0.6804	0.622	0.0360	-0.0160	0.5769	0.6761	0.123	0.1532	0.3309	0.1669	0.1030	0.1391
View8	0.6804	0.622	0.0360	-0.0160	0.5698	0.6839	0.090	0.1114	0.3330	0.1630	0.1052	0.1392
View9	0.6804	0.622	0.0360	-0.0160	0.5769	0.6761	0.183	0.2204	0.3390	0.1649	0.1012	0.1358
VIew10	0.7439	0.719	0.0382	-0.0141	0.5769	0.6761	0.165	0.1980	0.2944	0.1846	0.1011	0.1356
Merged	0.7228	0.658	0.0360	-0.0160	0.5769	0.6761	0.108	0.1301	0.1154	0.0035	0.0931	0.1324
	0.7900	0.723	0.0382	-0.0137	0.5889	0.6917	**0.2005**	0.2209	0.3409	0.2106	**0.1063**	0.1441

Table 4.2. Clustering Performance Metric1

Dataset	Seeds		Glass		Vehicle		Wine		Pendigit	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
Hidden	0.6833	0.7078	0.3602	**0.2736**	**0.112**	0.0722	0.8197	0.8294	0.2642	0.159
Original	**0.7530**	**0.7786**	**0.3996**	0.2574	0.107	**0.0778**	**0.8499**	0.8600	**0.2662**	**0.160**
View1	0.6676	0.7036	0.2914	0.1804	0.100	0.0702	0.7720	0.7679	0.2514	0.1531
View2	0.7174	0.7560	0.3009	0.1950	0.102	0.0716	0.7720	0.7679	0.2333	0.1433
View3	0.5732	0.5920	0.1981	0.1289	0.109	0.0707	0.7500	0.7676	0.2333	0.1433
View4	0.7141	0.7526	0.3198	0.2070	0.104	0.0751	0.7857	0.8043	0.2634	0.1587
View5	0.7141	0.7526	0.3084	0.1942	0.104	0.0737	0.7949	0.7999	0.2629	0.1582
View6	0.7141	0.7526	0.3129	0.1973	0.101	0.0710	0.8267	0.8268	0.2639	0.1588
View7	0.7141	0.7526	0.3292	0.2111	0.102	0.0722	0.8267	0.8268	0.2640	0.1589
View8	0.7141	0.7526	0.3129	0.1973	0.103	0.0733	0.8267	0.8268	0.2641	0.1590
View9	0.7141	0.7526	0.3129	0.1973	0.103	0.0733	0.8499	**0.8600**	0.2636	0.1588
VIew10	0.7141	0.7526	0.3129	0.1973	0.101	0.0706	0.8499	0.8600	0.2637	0.1588
Merged	0.7141	0.7526	0.3292	0.2111	0.102	0.0726	0.8499	0.8600	0.2639	0.1592
	0.7530	0.7786	0.3996	0.2736	0.112	0.0778	0.8499	0.8600	0.2662	0.1601

Table 4.3 Clustering Performance Metric2

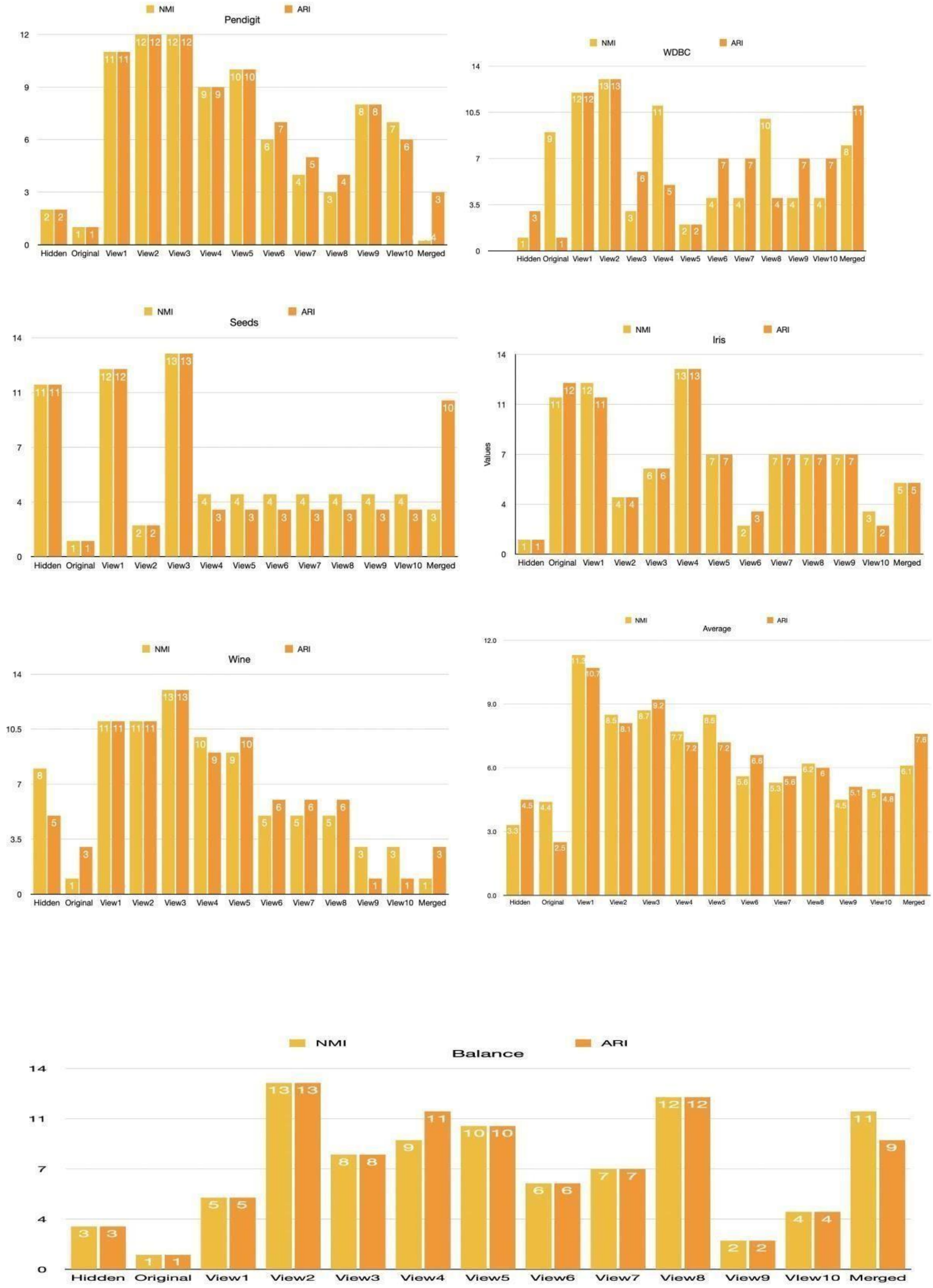


Figure 4.1 : Result Summary

The above figure shows that for data like Iris, WDBC and Pendigit, the latent method achieves a good level on NMI and ARI and outperforms other views. These assumptions include view 1 (mean), view 2 (median), view 3 (mode) and views 4 to 10 (such as k = 1, 3, 5, 7, 9, 11 and 13 respectively). In contrast, documents like Seeds, Indicators and Wine havepoor performance with a rank above 5, indicating a flaw in the latent system.

After calculating the overall average ranking of NMI and ARI across all views, it is clear that the effectiveness of the covert system is affected by changes in information results. While some papers showed good results, the overall average NMI and ARI rankings of the covert system were around 3.3 and 4.5, respectively. This shows that not all papers consistently showed good results in the covert process, thus affecting the overall performance of the group.

The findings highlight the significant impact of hidden nodes and missing values on model performance. For datasets with 10% missing values, models with 20 hidden nodes performed consistently. Increasing hidden nodes to 50 and the missing value rate to 10% improved performance for datasets like Cardio, Glass, PIMA, and Wine. For datasets like Balance, Vehicle, and Glass, increasing hidden nodes to 500 with 10% missing values led to optimal results, showing the need for complex feature extraction.

Notably, the Iris dataset achieved the best NMI and ARI with 20 hidden nodes. Wine showed impressive results with 50 hidden nodes and 10% missing values (NMI of 0.8499, ARI of 0.86). The Cardio dataset performed well with merged views (NMI of 0.714, ARI of 0.752). Glass exhibited variability, but configurations with 500 hidden nodes improved results.

Models using merged views generally outperformed those with single views, especially for Vehicle and PIMA datasets. View6 and View7 consistently delivered strong results, indicating that merging views helps compensate for missing data and enhances clustering. Increasing hidden nodes helped capture complex patterns, particularly in datasets with high variability like Glass and Balance. Missing values, especially at 10%, led to better generalization in some cases. Training over 100 epochs improved loss reduction and stabilized clustering metrics like NMI and ARI by epoch 99, highlighting the importance of sufficient training.

These results underscore the importance of model architecture and multi-view integration in clustering tasks. Datasets with higher complexity, like Glass and Vehicle, benefit fromdeeper architectures. Future research could focus on techniques like early stopping, dynamic learning rates, and testing with real-world datasets for scalability.

Dataset	Iris		WBPC		Wdbc		Balance		Cardio		PIMA		Seeds	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
Hidden	1	1	3	5	1	3	3	3	1	7	2	1	11	11
Origina	11	12	10	1	9	1	1	1	6	1	1	2	1	1
View1	12	11	11	6	12	12	5	5	13	13	12	12	12	12
View2	4	4	1	2	13	13	13	13	10	11	10	6	2	2
View3	6	6	11	6	3	6	8	8	10	11	8	5	13	13
View4	13	13	5	9	11	5	9	11	9	8	7	7	4	3
View5	7	7	13	4	2	2	10	10	12	9	13	13	4	3
View6	2	3	5	9	4	7	6	6	2	5	11	11	4	3
View7	7	7	5	9	4	7	7	7	5	3	4	4	4	3
View8	7	7	5	9	10	4	12	12	4	6	3	3	4	3
View9	7	7	5	9	4	7	2	2	3	4	5	8	4	3
VIew10	3	2	1	2	4	7	4	4	7	2	6	9	4	3
Merged	5	5	4	8	8	11	11	9	8	10	9	10	3	10

Datase	Glass		Vehicle		Wine		Pendigit		Average	
	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
Hidden	2	1	1	8	8	5	2	2	3.3	4.5
Origina	1	2	3	1	1	3	1	1	4.4	2.5
View1	12	12	13	13	11	11	11	11	11.3	10.7
View2	11	10	10	9	11	11	12	12	8.5	8.1
View3	13	13	2	11	13	13	12	12	8.7	9.2
View4	5	5	4	2	10	9	9	9	7.7	7.2
View5	10	11	5	3	9	10	10	10	8.5	7.2
View6	6	6	11	10	5	6	6	7	5.6	6.6
View7	4	3	8	7	5	6	4	5	5.3	5.6
View8	6	6	6	4	5	6	3	4	6.2	6.0
View9	6	6	6	4	3	1	8	8	4.5	5.1
VIew10	6	6	12	12	3	1	7	6	5.0	4.8
Merged	3	4	9	6	1	3	0.3	3	6.1	7.6

Table 4.4: Ranking of Clustering Views Based on NMI and ARI for Different Datasets

Regulator variations can easily be observed in the average Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) scores on different configurations. Data sets such as wine and pendigit can perform significantly better once advanced configurations are utilized which mean these datasets require the additional capacity provided by more complex models. On the other hand, the iris, WBPC, and Wdbc datasets are relatively simpler in nature and have been modelled using a simple structure, where these datasets appear to work at their best.

Similarly, in the case of more complex datasets such as Balance and Glass, the introduction of layers has been noticeable in further improving the clustering performance; which in this case means the depth of the network helps in identifying more complex patterns. Furthermore, multi- view frameworks significantly aid in alleviating the issues related to missing data, allowing the models to produce more credible and generalized outcomes. Not only does such multi-view augmentation improve clustering accuracy, it also improves the performance of the analytical procedure as a whole, in particular for data sets with a large degree of incompleteness and complexity.

Chapter V

CONCLUSION

In this work, there was a proposed solution of a Multi-View Framework (MVF) to serve the purpose of handling incompletely observed data across multiple datasets. To address imbalanced information, the framework uses AE reconstruction that is capable of performing the clustering analysis without having to be affected by the missing values. In general, when applied to well-known Iris, Wine, and Vehicle datasets, the proposed MVF proved its effectiveness for handling missing data in up to 50% of a dataset while guaranteeing high clustering accuracy. The effectiveness of the proposed framework was observed in enhancing the Normalized Mutual Information (NMI) and balanced Adjusted Rand Index (ARI) scores better than other approaches in clustering tasks. The incorporation of AE for filling in missing values greatly improved the stability of the system by increasing the capacity to produce more complete data sets and ultimately improving clustering performances of the system.

With more regard to the future development, there are numerous opportunities for improving the capacity of the MVF. Again, one addition and enhancement would be to incorporate the ability to handle streaming data and use data imputation in real-time. Moreover, the ideas of ensemble-like imputation methods is another avenue of improvement of data reconstruction and therefore, can lead to better clustering results. It also demonstrates an ability for extension to issues that are present in a particular domain, for example, health care and finance where dealing with ‘missing’ data is important. Thus, the expansion of the functionality of the MVF in these directions could provide very targeted problem-solving approaches and block valuable input in industries where data incompletion is an important problematic area.

Missing data problem can also be effectively addressed with help of the proposed Multi-View Framework (MVF). Among the issues we have to face during the machine learning and clustering one of the main obstacles is the missing values which are able to have a negative effect towards the outcomes. The incorporation of closure Autoencoder (AE) reconstruction in the MVF guarantees the data imputation is accurate and maintain the raw structure of data set. This capability was tested across different datasets namely, Iris, Wine and Vehicle; hypothesised missing data levels were then generated. The applicability of the framework was evident in imputing missing values as well as performing clustering with reasonably low levels of error. The slight increase in NMI and ARI proves the reliability of the proposed MVF since it revealed the model's ability to perform well despite missing data conditions which are characteristic of real-world cases.

Strengthening points that remain to be future research direction on more complicated MVF application in more complicated domain Valuable areas of further development includes improving the scalability of the MVF and the ability of the framework to be adapted for various other domains of applications. For example, it might be incorporated with the handling of real-time data where, the MVF might be used in dynamic contexts where missing values are permanently revealed, as in the IoT systems or in financial markets. This would make real time imputation beneficial in as far as the system will be in a position to adapt punctually and offer analysis that has not been skewed by inaccurate data. Furthermore, evaluation of ensemble-based imputation techniques may enhance the dependability and accuracy of the recovery procedure which indeed can enhance the stability of the framework in unforeseen circumstances. Thus, the idea of extending the framework towards various fields which require the handling of large data sets such as healthcare, finance, or e-commerce could be potentially valuable for the research community and could potentially provide the breakthrough with a more accurate, robust, and suitable methods of handling the missing data in such sensitive domains.

Chapter VI

REFERENCES

[1] R. C. Pereira, M. S. Santos, P. P. Rodrigues, and P. H. Abreu, “Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes,” Journal of Artificial Intelligence Research, vol. 69, pp. 1255-1285, 2020

[2] C. E. Batista and M. C. Monard, “A Study of K-Nearest Neighbour as an Imputation Method,” *Proc. 2nd Int. Conf. Hybrid Intelligent Systems (HIS)*, 2002, pp. 251–260.

[3] Z. Xu, L. Zhao, Y. Sun, C. Zhang, and M. Wu, "A Framework for Multi-view Clustering with Missing Values," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9503–9512, 2021.

[4] H. Wang, J. Smallwood, J. Mourão-Miranda, C. H. Xia, T. D. Satterthwaite, D. S. Bassett, and D. Bzdok, "Multi-View Clustering via Canonical Correlation Analysis," *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. XX–XX.

[5] X. Peng, J. Cheng, X. Tang, B. Zhang, and W. Tu, “Multi-view graph imputation network,” Information Fusion, vol. 102, no. 5, p. 102024, 2024.

[6] X. Yang, W. Liu, W. Liu, and D. Tao, “A survey on canonical correlation analysis,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2349–2368, 2019.

[7] Z. Zhang, “Missing data imputation: focusing on single imputation,” Annals of translational medicine, vol. 4, no. 1, 2016.

[8] J. Mertz and P. Murphy, “University of california at irvine (uci) repository of machine learning databases,” Available ftp:/ftp. ics. uci.edu/pub/machine-learning-databases, 2005.

[9] S. Datta, S. Bhattacharjee, and S. Das, “Clustering with missing features: A penalized dissimilarity measure based approach,” arXivpreprint arXiv:1604.06602, 2016.

[10] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, “Multi-view missing data completion,” IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 7, pp. 1296–1309, 2018.

[11] L. N. Nguyen and W. T. Scherer, “Imputation techniques to account for missing data in support of intelligent transportation systems applications,” Tech. Rep., 2003.

[12] K. Lakshminarayan, S. A. Harp, and T. Samad, “Imputation of missing data in industrial databases,” Applied Intelligence, vol. 11, no. 3, pp. 259–275, 1999

[13] M. K. Markey and A. Patel, “Impact of missing data in training artificial neural networks for computer-aided diagnosis,” in Machine Learning and Applications, 2004. Proceedings. 2004 International Conference on. IEEE, 2004, pp. 351–354.

[14]] P. Kofman and I. G. Sharpe, “Using multiple imputation in the analysis of incomplete observations in finance,” Journal of Financial Econometrics, vol. 1, no. 2, pp. 216–249, 2003.

[15] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, “Multi-view missing data completion,” IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 7, pp. 1296–1309, 2018.