Learning from high dimensional healthcare data to improve interpretability and insights

By

Indra Prakash Jha
(PhD15201)

Under the supervision of

Prof. (Associate) Vibhor Kumar

Department of Computational Biology
Indraprastha Institute of Information Technology (IIIT), Delhi
New Delhi - 110020

May, 2023

Learning from high dimensional healthcare data to improve interpretability and insights

By

Indra Prakash Jha
(PhD15201)

A Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy (PhD)

Certificate

This is to certify that the thesis entitled "*Learning from high-dimensional healthcare data to improve interpretability and insights*" being submitted by *Mr. Indra Prakash Jha* to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of Doctor of Philosophy, is an original research work, carried out by him under my supervision. In my opinion, the thesis has met the standards, fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

May, 2023

Signed by: Indra Prakash Jha
Location: Rajasthan, IN
Date: 09/10/2024 11:21:08

Associate Professor, Vibhor Kumar,

Indraprastha Institute of Information Technology, Delhi

New Delhi, 110020

ABSTRACT

The escalating volume and intricate nature of healthcare and social care datasets necessitate the implementation of unconventional feature learning strategies to tackle current challenges. The examination of health and biological datasets enables the assessment of existing computational techniques and fosters the creation of novel algorithms and methodologies that can be applied to difficulties in other fields. By employing these concepts, we have not only developed novel algorithms but also performed meticulous analysis to tackle concerns pertaining to healthcare and social care. It should be noted that the methodologies and analysis processes can be adapted to accommodate supplementary datasets featuring diverse data types and formats.

Firstly, the authors present a novel manifold learning algorithm, named "***Topological Preservation and Distance Scaling***" (TPDS), which aims to enhance classification and visualization of high-dimensional datasets. The proposed method addresses the challenge of the "***curse of dimensionality***". The methodology aims to maintain the hierarchical structure of data by preserving both local topology and distances during linear and non-linear dimension reduction. This approach is designed to prevent the collapse of data points in visualization. In the second study, the authors present a novel matrix factorization-based manifold learning algorithm, "***Network Inference in Reduced Dimensions****" (NIRD),* for inferring very large regulatory networks with a very large number of features. The study revealed that the proposed approach exhibited superior performance compared to existing methods, namely GENIE-3 and GrnBoost2, in terms of both the computational time required to infer the network and the accuracy of estimated edges or connections. The objective was to deduce intricate

dependency and regulatory networks that encompass a vast number of dimensions, with the aim of capturing non-linear dependencies among random variables.

Subsequently, two causal discovery analyses were conducted on high-dimensional healthcare datasets to infer "***explainable***" associations and estimate public health concerns, such as the ***prevalence of mental health***. The hypothesis posits that the utilization of ***generative probabilistic graphical models***, specifically the ***Bayesian network*** and ***Markov network***, in tandem with the ***Markov blanket*** concept of feature learning may yield greater interpretability. The initial investigation involved the utilization of survey data collected from a diverse group of American adults, encompassing various age groups, genders, and socioeconomic statuses. In contrast, the subsequent inquiry employed data from the Longitudinal Ageing Study in India (LASI) Wave-1 survey, which focused on elderly individuals residing in India. The methodology employed facilitated the determination of the most relevant attributes (driver factors) that could effectively represent the incidence of mental health disorders in both studies. The features chosen by our approach demonstrate relevance in facilitating ***actionable interventions*** aimed at promoting mental health well-being among adults during pandemic-induced lockdowns, as well as among elderly individuals in India.

Acknowledgements

I would like to express my deepest gratitude to **Prof** (**Associate**) **Vibhor Kumar**, my thesis advisor, for his invaluable guidance, support, and encouragement throughout the entire process of writing this thesis. His constant availability, insightful feedback, and unwavering dedication have been instrumental in the completion of this work.

I would also like to extend my heartfelt thanks to my colleagues and friends, Dr Neetesh Pandey, Dr Smriti Chwala, Omkar Chandra, Shreya Mishra and Madhu Sharma for their support, encouragement, and inspiration throughout the thesis journey. Their understanding, humor, and camaraderie have made this experience truly memorable.

To my colleagues, Raghav Awasthi, Priyadarshini Rai, Dr. Sumit Patiyal, Abhishek Halder, Dr. Anjali Dhal, Dr. Piyush Agarwal, Sarita Poonia, Dr. Krishan Gupta, Chitrita Goswami and Dr. Manan thank you for your enthusiasm, dedication, and hard work. To Dr. Anupam Mondal, Dr. Anjali Lathwal, Dr. Chakit Arora, Monalisa Jena, Prateek Singh, Shrey Gupta, Parul Sharma, Keshav Bhojak, Dr. Richa, Dr. Dilraj, Sphoorti, Diksha, Karaj: I appreciate all your energy, commitment, and effort. Your contributions have been invaluable in shaping and improving the quality of this research.

I am grateful to my parents, Dr Ramanand Jha and Mrs Preeta Jha, my brothers, Mr. Ravi Prakash Jha and Mr. Divya Prakash Jha, and my sister-in-law, Mrs Simple Jha and Mrs Manjari Jha for their unwavering love, support, and guidance.

To my beloved wife, Priya Mishra, and my daughter Aishwarya Jha, thank you for your unconditional love, support, and understanding. Your patience, encouragement, and belief in me have been essential in this journey.

Lastly, I would like to express my deepest gratitude, again, to my thesis supervisor, **Prof** (**Associate**) **Vibhor Kumar**, for his mentorship, guidance, and expertise. Without his support and encouragement, this work would not have been possible.

Thank you all for your unwavering support and encouragement throughout this journey.

RESEARCH PUBLICATIONS

THESIS PUBLICATION

- Learning the mental health impact of COVID in the United States with explainable artificial intelligence:

  Indra Prakash Jha, R Awasthi, A Kumar, V Kumar, and T Sethi

  JMIR Mental Health, 8 (4), e25097

- Local-Topology-Based Scaling for Distance Preserving Dimension Reduction Method to Improve Classification of Biomedical Data-Sets:

  K Khosla[#], Indra Prakash Jha[#], A Kumar, V Kumar

  MDPI Algorithms 13 (8), 192

  [#] : Joint First Authorship

- Stratified assessment for geriatric mental health using probabilistic graphical model: a cross-sectional observational study in India:

  Indra Prakash Jha, S Mishra, and V Kumar
  MedRxiv

OTHER PUBLICATIONS

- Associating pathways with diseases using single-cell expression profiles and making inferences about potential drugs:
  M Sharma, Indra Prakash Jha, S Chawla, N Pandey, O Chandra, S Mishra, V Kumar
  Briefings in Bioinformatics, 23(4) 2022, bbac241.

- Matching queried single-cell open-chromatin profiles to large pools of single-cell transcriptomes and epigenomes for reference supported analysis:
  Shreya Mishra, Neetesh Pandey, Smriti Chawla, Madhu Sharma, Omkar Chandra, Indra Prakash Jha, Debarka SenGupta, Kedar Nath Natarajan, Vibhor Kumar
  Genome Research 2023;33(2):218-231

- Patterns of transcription factor binding and epigenome at promoters allow interpretable predictability of multiple functions of non-coding and coding genes:
  O Chandra, M Sharma, N Pandey, Indra Prakash Jha, S Mishra, SL Kong, V Kumar
  Comput Struct Biotechnol J. 2023; 21: 3590–3603

- Deciphering the phenotypic heterogeneity and drug response in cancer cells using genome-wide activity and interaction of chromatin domains:
  Neetesh Pandey, Madhu Sharma, Arpit Mathur, George Anene Nzelu, Muhammad Hakimullah, Indra Prakash Jha, Omkar Chandra, Shreya Mishra, Ankur Sharma, Roger Foo, Amit Mandoli, Ramanuj DasGupta, Vibhor Kumar
  BioRxiv

Chapter 1

Introduction

1.1 Background

The advent of new technologies has brought about a data revolution in various fields, including bioinformatics, computational genomics, computational biology, social data, and healthcare AI. These technologies have enabled us to collect and analyze data at an unprecedented scale and resolution, leading to new discoveries and insights in these fields. In bioinformatics, the use of next-generation sequencing technologies has made it possible to generate vast amounts of genomic data, allowing researchers to study the genetic basis of disease and develop personalized treatments. Similarly, computational genomics has enabled the analysis of large-scale genomic data to identify patterns and associations between genetic variations and disease. In computational biology, the use of machine learning algorithms has enabled the analysis of complex biological systems, such as protein-protein interactions and gene regulatory networks, to identify key drivers of disease and develop targeted therapies. Social data has also become a valuable resource for understanding human behavior and social dynamics. Social media platforms and other online communities generate vast amounts of data that can be analyzed to understand trends and patterns in human behavior and sentiment.

In healthcare AI, the use of machine learning algorithms has enabled the analysis of large-scale patient data to develop predictive models for disease diagnosis and treatment. This has the potential to revolutionize healthcare by enabling personalized treatment plans based on a patient's individual genetic and medical history. Overall, the data revolution has enabled us to collect and

analyze data at an unprecedented scale and resolution, leading to new insights and discoveries in various fields.

Machine learning, deep learning, and Bayesian statistics are all methods that enable us to make sense of the vast amounts of data being generated by new technologies. These approaches enable us to identify patterns and relationships in the data that would be difficult or impossible to detect using traditional statistical methods. In machine learning, algorithms are trained on large datasets to identify patterns and relationships in the data. This enables us to develop predictive models that can be used to make predictions or classify new data. For example, in healthcare, machine learning algorithms can be used to predict the likelihood of a patient developing a certain disease based on their medical history and genetic information. Bayesian learning is a statistical approach that involves updating our beliefs about the data based on new evidence. This enables us to make more accurate predictions and decisions based on the data. For example, in finance, Bayesian learning can be used to predict the likelihood of a certain stock increasing in value based on past performance. So, these tools are transforming various fields, from healthcare to finance, and have the potential to drive significant advancements in the years to come.

The data generated by new technologies often suffers from a variety of issues that make it difficult to analyze and interpret accurately. One common issue is high dimensionality, which arises when there are many features or variables in the data. This can make it difficult to identify patterns and relationships in the data and can lead to overfitting and poor generalization of models. Another issue is sparseness, which arises when many of the features in the data are zero or missing. This can be a common issue in healthcare data, for example, where

not all patients may have had the same tests or treatments. Sparse data can make it difficult to identify relationships and patterns in the data and can lead to biased or incomplete analysis. Finally, noisy data can also be an issue, where there is a random error or measurement error in the data. This can be a common problem in social data, for instance, where irony or sarcasm may affect sentiment analysis.

These issues require the use of specialized techniques, such as machine learning and data science, to extract meaningful insights from the data. For example, dimensionality reduction techniques such as principal component analysis can be used to reduce the number of features in the data and identify the most important ones. Sparse data can be handled using specialized techniques such as regularization or imputation, while noisy data can be handled using smoothing or filtering techniques.

Overall, all these challenges are equally important to address, but for this thesis, our goal was to address the issues of high dimensionality and related aspects.

1.1.1 Understanding the high dimensionality in datasets

The high-dimensional dataset may be defined as the case where the number of predictors or features (p) is greater than the number of samples or observations (n). Additionally, if p is slightly smaller than n, then all the discussion may be applied [1].

Mathematically,

The underlying dataset is:

$$M = \left\{ \left(\left\{ x_j \right\}_{j=1}^{p} \right)_i , y_i \right\}_{i=1}^{n} \quad \text{---(1) } (supervised)$$

$$M = \left\{ \left(\left\{ x_j \right\}_{j=1}^{p} \right)_i \right\}_{i=1}^{n} \quad \text{---(2) } (unsupervised)$$

where

n − *the number of samples and*

p − *number of features*

$$M \text{ is high dimensional} \Leftrightarrow ((n << p) \vee (n \sim p))$$

1.1.2 Why is dimensionality reduction challenging?

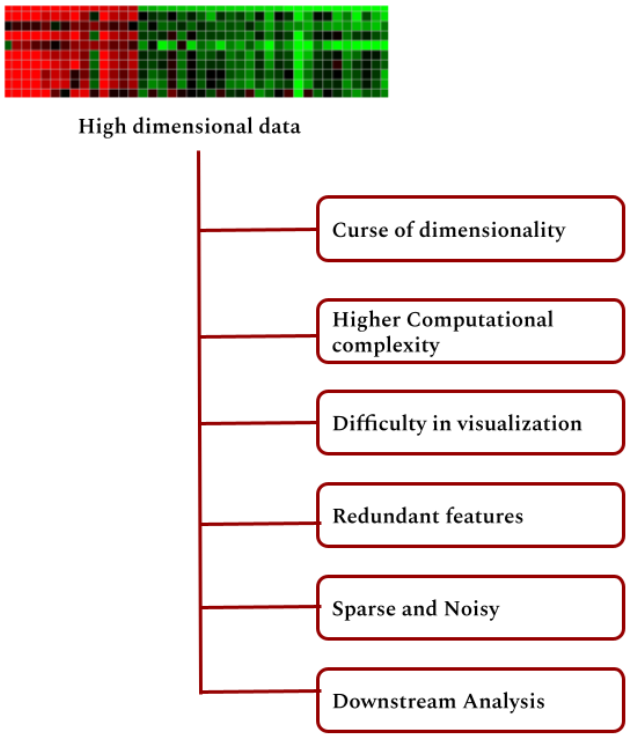

Fig. 1.1: Some challenges with high-dimensional datasets

Mathematically, dimensionality reduction can be challenging for several reasons.

- One of the main challenges is that, as the dimensionality of the data increases, the number of possible combinations of features grows exponentially. This can lead to the curse of dimensionality, where the number of observations required to accurately represent the data increases exponentially with the number of features.
- Another challenge is that, as the dimensionality of the data increases, the number of parameters required to fit a model also increases. This can lead to overfitting, where a model fits the noise in the data rather than the underlying pattern, and can result in poor generalization to new data.
- Furthermore, many common dimensionality reduction techniques, such as PCA, rely on assumptions about the distribution of the data, such as linearity or normality. When these assumptions are violated, the effectiveness of the technique may be reduced or the results may be misleading.

This academic text aims to delve deeper into the challenges encountered in data analysis, particularly in cases where standard methods of analysis are no longer deemed appropriate. To illustrate this point, the use of multiple linear regression will be employed as an example.

The Model for multiple linear regression is

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ . \ . + \beta_p x_p$$

$$f(x) = \boldsymbol{\beta}^T x$$

here $x^{(i)}$ *is* i^{th} *sample and* x_j *is* j^{th} *feature vector*

$$here\ \boldsymbol{\beta}^T = (\beta_0\ \beta_1 \dots\ \beta_p)\ and\ x^T = (1\ x_1\ x_2 \dots\ x_p)$$

The Loss Function for mean squared error (MSE) is

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \boldsymbol{\beta}^T x^{(i)}\right)^2$$

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left([Y]_i - [X\boldsymbol{\beta}]_i\right)^2$$

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}||Y - X\boldsymbol{\beta}||^2$$

The gradient is

$$\nabla\mathcal{L} = -\frac{2}{n}X^T(Y - X\boldsymbol{\beta})$$

By Setting it to zero to derive the analytical solution

$$\nabla\mathcal{L} = 0$$

$$\Rightarrow -\frac{2}{n}X^T\left(y - X\widehat{\boldsymbol{\beta}}\right) = 0$$

$$\Rightarrow X^T y = X^T X\,\widehat{\boldsymbol{\beta}}$$

$$\Rightarrow \widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1}.X^T y$$

It is necessary for the number of features (p) to be smaller than the number of samples (n). The calculation of the estimator ($\boldsymbol{\beta}$), denoted as $\widehat{\boldsymbol{\beta}}$, is contingent upon the condition that the matrix X possesses full rank, indicating that all singular values of the matrix $X^T X$ are non-zero. If matrix X does not possess full rank, then the inverse of the product of X transpose and X, denoted as $\left(X^T X\right)^{-1}$, will not exist [2].

Ultimately, a common trade-off exists between the level of information retention and the extent of dimensionality reduction attained. Stated differently, the act of reducing the dimensionality of the dataset has the potential to lead to the

forfeiture of significant data or the inclusion of noise. Further elaboration on this trade-off will be provided in Chapter 2.

1.1.3 Why are biological and healthcare data uniquely challenging compared to other fields?

Biological and healthcare datasets pose significant challenges due to their high dimensionality. Examples of data types include genome-wide association studies (GWAS) data, human microbiome data, cancer genomics data, single-cell sequencing data, and protein expression data. Throughout the study, we have employed higher-dimensional datasets from the fields of biology and healthcare, which will be elaborated on in the corresponding chapters. The unique challenges of higher-dimensional biological and healthcare data compared to other fields are as follows:

- **High Dimensionality, Sparsity, and Noise**: Healthcare datasets often exhibit high dimensionality, sparsity, and noise, making them difficult to analyze and extract meaningful insights. This complexity stems from the diverse domains encompassed by healthcare data, including bioinformatics, computational genomics, computational biology, social data, and healthcare AI.
- **Rapid Data Accumulation**: The advent of new technologies has led to healthcare data accumulating at an unprecedented rate. Managing and analyzing this vast and rapidly growing volume of data is a significant challenge.
- **Need for Specific Analysis Techniques**: Due to the high dimensionality and complexity of the data, traditional dimensionality reduction techniques, while helpful, can lead to the loss of crucial information. As a

result, feature learning or representation learning techniques are needed to algorithmically discover the necessary representations for feature detection or classification from raw data.

- **Critical Impact on Health Outcomes:** The analysis of healthcare data is directly linked to critical areas such as mental health, disease diagnosis, and treatment efficacy. Therefore, the need for efficient and effective analysis of these large and complex datasets is paramount.
- **Requirement for Reliable and Accurate Results**: Due to the high stakes involved in healthcare, including patient safety and well-being, there is a requirement for highly reliable and accurate results, which makes applying traditional machine learning and AI techniques more challenging in this domain

1.1.4 Patient care and decision modeling by leveraging the higher dimensional datasets:

High dimensionality in healthcare datasets can significantly impact patient care and medical decision-making in several ways:

- **Complexity in Data Interpretation:** High-dimensional data can be challenging to interpret accurately. In healthcare, where data interpretation is critical for diagnosis and treatment decisions, this complexity can lead to uncertainties or delays in decision-making.
- **Risk of Overfitting in Predictive Models:** When developing predictive models using high-dimensional data, there's a risk of overfitting, where the model performs well on training data but poorly on new, unseen data.

This can lead to inaccurate predictions or diagnoses when applied to real patient data.

- **Increased Computational Requirements:** Analyzing high-dimensional data requires significant computational resources and expertise. This can be a barrier in healthcare settings, particularly in resource-limited environments, potentially delaying the integration of data-driven insights into patient care.
- **Challenges in Identifying Relevant or Driver Factors:** With many variables in high-dimensional datasets, it can be challenging to identify which factors are most relevant to a patient's condition or treatment response. This can complicate treatment planning and personalization.
- **Difficulty in Visualizing Data:** High dimensionality hinders the ability to visualize data effectively, which is often crucial for understanding complex patient information and communicating it among healthcare providers.
- **Increased Potential for Spurious Correlations:** In high-dimensional data, the chance of finding false correlations increases simply due to the large number of variable comparisons. This can mislead research and clinical decisions if not properly controlled.
- **Complexity in Integrating Diverse Data Sources:** Patient care often involves integrating data from various sources, like genomic data, clinical records, imaging, and lab tests. High dimensionality can make this integration more complex, potentially obscuring critical insights.
- **Implications for Personalized Medicine:** Personalized medicine relies on analyzing detailed data from individual patients to tailor treatments. High-dimensional data can provide the necessary depth of information,

but the complexity and potential for noise can make it challenging to extract personalized insights.

- **Ethical and Privacy Considerations:** The likelihood of ethical issues or privacy violations increases with the amount of data gathered. High-dimensional data can contain sensitive information that must be carefully managed to protect patient confidentiality.
- **Data Storage and Management Issues:** High-dimensional data requires substantial storage space and sophisticated data management strategies, which can be costly and complex to maintain.

In summary, while high-dimensional data in healthcare holds immense potential for advancing patient care and medical decision-making, it also presents significant challenges. These include complexities in data analysis and interpretation, risks of overfitting, computational demands, and issues related to data integration, privacy, and management. Addressing these challenges is crucial for leveraging the full potential of high-dimensional data in healthcare.
This thesis primarily focuses on complexity in data interpretation, tackling the risk of overfitting in models, the increased computational requirements, the identification of driver factors, data visualization, and the handling of the potential for spurious correlations in studying complex biological systems.

1.2 Problem Statement

The problem statement of this thesis revolves around the challenges and opportunities associated with handling large and complex healthcare datasets,

specifically in the context of machine learning and AI techniques. With the advent of new technologies and the data revolution, healthcare data is accumulating at an unprecedented rate. This data encompasses diverse domains such as bioinformatics, computational genomics, computational biology, social data, and healthcare AI. However, this data often exhibits high dimensionality, sparsity, and noise, making it difficult to analyze and extract meaningful insights.

To overcome these challenges, dimensionality reduction techniques are often used to obtain low-dimensional estimates of the dataset. While these techniques have shown promise in reducing the curse of dimensionality, they can lead to information loss and may not always capture the relevant features needed for effective analysis. Therefore, feature learning or representation learning techniques are required to algorithmically discover the representations needed for feature detection or classification from raw data.

In the context of healthcare, the need for efficient and effective analysis of large and complex datasets is critical for addressing pressing issues such as mental health, disease diagnosis, and treatment efficacy. However, the sheer volume and complexity of healthcare data make it difficult to apply traditional machine learning and AI techniques. Furthermore, the high stakes involved in healthcare require reliable and accurate results to ensure patient safety and well-being.

Thus, the central problem statement of this thesis is to explore and develop novel machine learning and AI techniques that can effectively handle large and complex healthcare datasets while mitigating issues such as high dimensionality, sparsity, and noise. The objective is to develop methods that can extract relevant features from raw data using feature learning techniques and apply dimensionality reduction methods to obtain low-dimensional estimates of

the dataset that capture the relevant features needed for effective analysis. The proposed techniques will be evaluated using real-world healthcare datasets and compared to existing state-of-the-art techniques to demonstrate their effectiveness.

Overall, this thesis aims to contribute to the development of novel machine learning and AI techniques that can handle large and complex healthcare datasets while ensuring accuracy, reliability, and explainability. By addressing the pressing issues in healthcare using advanced data science, this thesis has the potential to significantly improve outcomes and advance the field of healthcare data science.

Hence, features must represent the information in the data in a format that will best fit the needs of the algorithm that is going to be used to solve the problem. Therefore, feature learning is used to overcome this issue. The feature learning or representation learning is a set of techniques that allows a system to algorithmically discover the representations needed for feature detection or classification from raw data. [3]

So, the biggest challenge with any machine learning or AI technique is making sure that existing algorithms can be used with datasets that have more and more dimensions. Handling the "curse of dimensionality," or the increase in the number of dimensions, is a very critical problem. Most attempts to fix it involve a two-step process in which relevant subspaces are first found by making appropriate transformations to the original space, and then standard learning algorithms are used to learn about them. To describe multidimensional data efficiently and in a small space, we need to embed high-dimensional data in lower-dimensional spaces. Our main problem statement in the field of healthcare is to find these low-rank approximations or estimates in such a way

that they mitigate noise, unfold latent relations, and facilitate further processing in such representations.

For underlying dataset given at (1), (2) *is*:

Our objective is to estimate dataset $\widehat{M}$

$$\widehat{M} = \left\{ \left(\left\{ x_j \right\}_{j=1}^{r} \right)_i , \; y_i \right\}_{i=1}^{n} \qquad \textit{(For supervised)}$$

$$\widehat{M} = \left\{ \left(\left\{ x_j \right\}_{j=1}^{r} \right)_i \right\}_{i=1}^{n} \qquad \textit{(For unsupervised)}$$

such that:

$$M \approx \widehat{M}$$

1.3 Research Objectives

To tackle the problem statement, we attempted to achieve the following three objectives:

1. To visualize and classify high-dimensional healthcare datasets
2. To Infer high-dimensional dependency/regulatory networks in lower-dimensional space
3. To discover explainable and actionable public health relationships in high-dimensional healthcare data

1.4 Research Scope

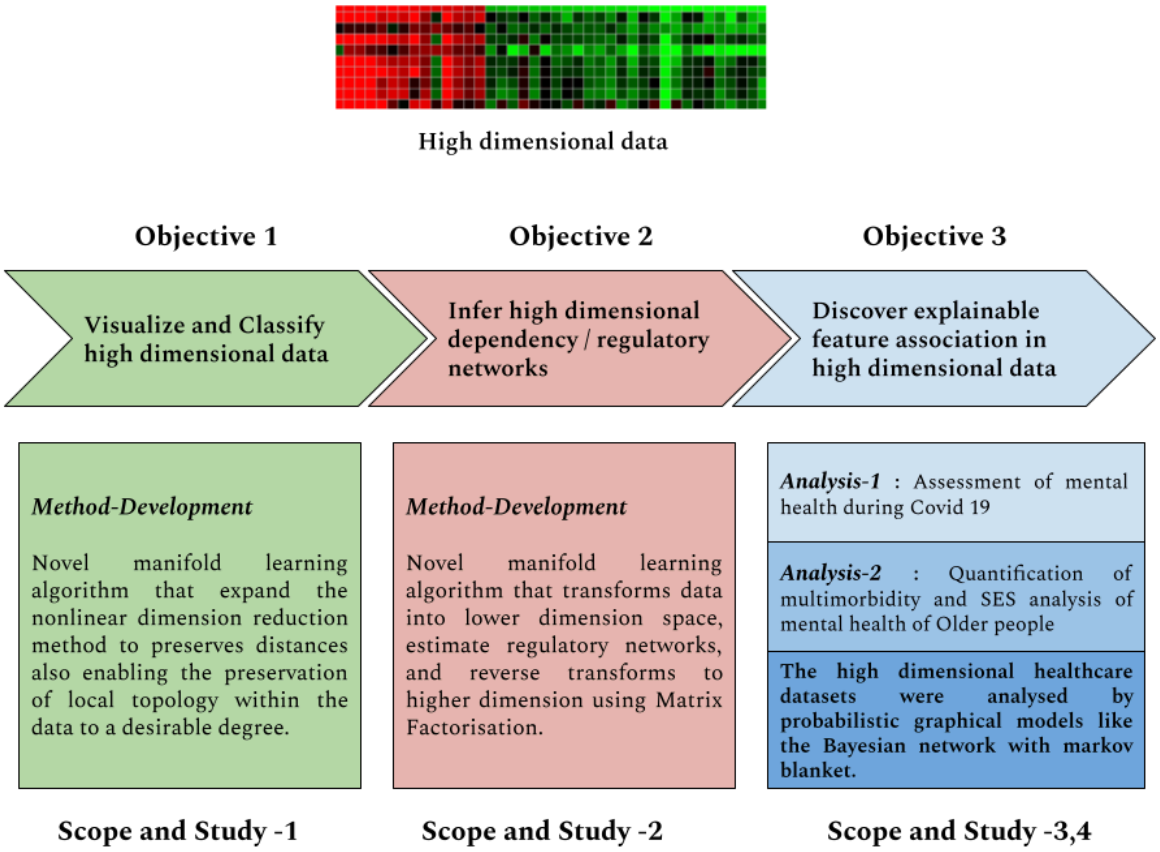

Fig. 1.2: Coherence of research objectives and studies

The scope of each objective is defined as follows:

1.4.1 To visualize and classify high-dimensional healthcare datasets

*STUDY: **Developed a novel manifold learning algorithm to preserve both local topology and distance: Topology Preserving Distance Scaling (TPDS)***

For the classification and visualization of high-dimensional biological and

medical datasets, we require a method that could guarantee the preservation of both local topology and global distances. Current popular existing methods for classification and visualization, such as SNE, tSNE, and UMAP, are not able to preserve global distances. To address the limitations of current distance-preserving nonlinear dimensionality reduction techniques like MDS, a new approach is required that can simultaneously preserve distance information and local topology. This has led to the development of a novel algorithm called Topology Preserving Distance Scaling (TPDS). TPDS builds upon the existing method of non-metric MDS by introducing a mechanism for distance shrinkage, ensuring that local structure within the data is preserved alongside distance information. Unlike traditional non-metric MDS, which emphasizes preserving the rank order of distances and is beneficial in scenarios where relative differences are more critical than exact distances, TPDS aims to improve both the visualization of data, similar to t-SNE, and the efficiency of downstream processes like classification in low-dimensional spaces.

1.4.2 To Infer high-dimensional dependency/regulatory networks in lower dimensional space

*STUDY-1: **Matrix Factorisation based Manifold learning for inferring the Large Gene Regulatory network to capture nonlinear dependencies in sparse single-cell expression profiles: Network Inference in Reduced Dimensions** (NIRD)*

Other than visualization and classification, researchers are also asking questions about the inference of regulatory networks of high-dimensional datasets. In

genomics also, many methods have been proposed for gene regulatory network inference for bulk expression profiles, but their utility for single-cell transcriptome profiles needs further investigation. Recently Pratapa et al. [4] evaluated 12 network inference methods using experimental and simulated single-cell expression profiles and concluded that the performances of all the tested methods were less than ideal.

In this study, we have proposed a novel, simple and powerful algorithm '*Network Inference in Reduced Dimensions' (NIRD)* for inference of gene regulatory networks from scRNA-seq expression profiles. Our method also leads to a conceptual advancement which can have wider application for modeling large numbers of non-linear dependencies. The method is based on the approach of matrix factorisation before using the non-linear regression approach to find the importance of factors using the non-linear regression approach. Further, our method estimates the direct effect of genes using the importance of factors. Worth mentioning that we have incorporated the different matrix factorisation based algorithms which shows the robustness of the modeling.

1.4.3 To discover explainable public health relationships in high dimensional healthcare data

*STUDY-1: **Learning the Mental Health Impact of COVID-19 in the United States with Explainable Artificial Intelligence***

*STUDY-2: **Assessment of mental health using Probabilistic Graphical Modeling among Older People in India***

With the advent of cheap and fast data generation technology, the application in clinical research becomes increasingly popular. The collected datasets often contain tens or thousands of biological features that need to be mined to extract meaningful information. One area of particular interest is discovering underlying causal mechanisms of disease outcomes. Over the past few decades, researchers have proposed several algorithms for causal discovery and extended them to infer such relationships. However, most of such algorithms suffer from the curse of dimensionality and multicollinearity. We hypothesize that applying graphical models such as the Bayesian network with the idea of markov blanket to causal discovery algorithms can solve both the high dimensionality and collinearity problems inherent to most healthcare datasets. Markov blanket refers to a statistical model that encompasses all the nodes directly linked to a specific node, including its parents, children, and the parents of its children, which form the "blanket". This separates the node from the network and provides better insights and accurate predictions or inferences about the node by focusing only on the relevant neighbor of the node.

We performed two studies based on this approach. In the first study, we used survey data of adults in the United States from different age groups, genders, and socioeconomic statuses, whereas, in the second study, we used Longitudinal Ageing Study in India (LASI) Wave-1 survey data of old age Indian people data. Our approach led to finding the most relevant features (drivers) that could model the mental health prevalence with high accuracy in both studies. Most importantly, features selected by our approach also appear to be relevant for providing actionable intervention to rescue mental health wellbeing for adults during the lockdown in pandemics and for Indian old age people.

1.5 Thesis Organization

The thesis is organized into six chapters, each of which addresses a different aspect of manifold learning and its applications in data analysis.

***Chapter** 1* provides an introduction to the thesis and its research objectives.

***Chapter** 2* presents a novel manifold learning algorithm that is designed to facilitate downstream analysis, including visualization and classification.

***Chapter** 3* introduces another novel manifold learning algorithm that is used to infer large regulatory networks and capture nonlinear dependencies.

***Chapter** 4* explores the mental health impact of COVID-19 in the United States through the use of a markov blanket based feature learning analysis.

***Chapter** 5* utilizes a probabilistic graphical model to analyze mental health prevalence among older individuals, with a focus on stratification analysis.

Finally, ***Chapter** 6* concludes the thesis with a summary of the research findings and their potential implications for future work in manifold learning and data analysis.

Chapter 2

TPDS: A manifold learning algorithm to preserve both local topology and distance for improved visualization and classification

*Developed a novel manifold learning algorithm, **TPDS**,*
Topological Preserving Distance Scaling

2.1. Introduction

This chapter is about the investigations that were performed to understand the first objective explained in Chapter 1. As already explained, our aim was to develop a manifold learning algorithm that might be used for better visualization and classification in high-dimensional datasets. Manifold learning is basically a special form of dimensionality reduction. Nonlinear dimensionality reduction is referred to as manifold learning.

The issue of dimensionality reduction in a high-dimensional dataset is of utmost importance across various fields. The challenge of dimension reduction is consistently encountered in the domains of genomics, proteomics, and medical informatics, as it hampers the extraction of valuable insights from datasets with limited dimensions. Reducing the dimensionality of high-dimensional data is a critical challenge across various fields, including genomics, proteomics, and medical informatics. The goal of dimensionality reduction is to distill meaningful insights into a lower-dimensional space to facilitate tasks like visualization, classification, and other downstream analyses. Typically, there are three main approaches to dimensionality reduction, each focusing on preserving one of the following: the distances between data points, the local structure or topology, or the overall information or entropy contained within the data.

“The Principal component analysis (PCA)” [5], which is the most widely used algorithm for the dimensionality reduction, preserves the information or entropy of the dataset; whereas other well known dimensionality reduction

algorithms, “multidimensional scaling (MDS)” [6] and “Sammon mapping” [7] preserve distance. On the other hand, nonlinear dimension-reduction techniques such as t-SNE [8] and “Locally-Linear embedding” (LLE) [9] maintain local topology of the data-points [10]. The different strategies for reducing dimensions each have their own criteria and cost functions, the goal is to optimize both as much as possible.

Considering the Sammon mapping technique, the cost function that was targeted to be minimized was squared difference. This is basically the difference between two distances of the dataset. These scaled distances in higher and reduced dimensions are also normalized by the original distance of higher dimensions. t-SNE is a well-known approach to dimension reduction that aims to minimize the cost function in a manner that is analogous to that of SNE. Here SME is “stochastic neighbor embedding” technique. The associated cost function [11] is derived from KL-divergences of conditional probabilities of the neighborhood. And this conditional probability is calculated by assuming that data points are sampled from multivariate gaussian distribution. Because the mentioned cost function avoids collapse of data-points that are similar to one another and instead places an emphasis on local distances, it frequently results in the information loss regarding global distances. There are non-linear methods for dimension reduction in which the primary focus is to assign or collapse the data-points to a known vector space. Generally this is achieved through the Self-Organizing Map (SOM) algorithm (lattice). SOM is frequently utilized in the process of making a preliminary prediction of manifolds [12].

Methods that are centered on the visualization of the local topology have the potential to, on occasion, result in the loss of essential information [13–15] that is necessary for the subsequent analysis. On the other hand, when it comes to dimension reduction, maintaining the global topology of distances in its original state is an absolutely necessary requirement. This is especially important to keep in mind when performing clustering, phylogenetic analysis, or regression. On the other hand, optimizing the cost function of MDS can frequently result in collapse of data-points in reduced dimensional space [6], which causes the loss of important information. This results from the fact that the data points are being collapsed. In biological data sets, such as single-cell transcriptome and proteome data, where it is essential to preserve cell heterogeneity in order to investigate cell-state gradients, this presents a significant challenge [16,17]. In addition, when conducting research on a group of people, also known as a cohort, it is essential not only to visualize significant differences between diseased and normal cases, but also to make use of the heterogeneity of patients in order to stratify the research. This is because significant differences between diseased and normal cases can only be seen when diseased cases are compared side-by-side with normal cases. We came up with a method that is an extension of MDS and that maintains the distance information as well as the local topology to a degree that is satisfactory as a solution to the issues that have been presented here. Because it can prevent artifacts caused by outlier effects of large distances and avoid the collapse of data-points, also simultaneously maintaining the distances. This may be utilized in clustering and machine learning downstream, this method is useful for biomedical data analysts.

Topology-preserving distance scaling is the name given to our method (TPDS). Before employing the non-metric MDS method, TPDS uses distance shrinkage to reduce the distance between points. As mentioned earlier, t-SNE, which is the manifold learning algorithm, is good in providing better visualization but it is not good in reduced dimension representations of data, which is required to perform machine learning steps such classification in a better manner. The proposed method TPDS is able to perform both tasks as mentioned above. It provides better visualization and also provides better representation in lower dimensions. This section starts off by providing a description of TPDS, and then moves on to provide an explanation of the method's cost function. Next, in the section titled "Results," we will talk about what we learned after applying the four distinct categories of data sets. Following that, we will talk about the benefits and drawbacks of TPDS.

2.2. Methods

As discussed, our objective was to persist not only local topology as well as distance in the data points while performing dimensionality reduction. For distance preservation, we used a different strategy named “distance scaling”.

We use the distance scaling strategy to make sure that the local topology stays preserved even when a distance-based method is used. The MDS algorithm that we use in TPDS makes an effort to maintain distances; however, local topology will also be maintained due to warping, which results in a shrinking of the distances that exist between data points. Yet, in order to accomplish this goal, one must have some prior hypothesis or belief regarding the manifold and local

topology of the dataset. As a result, we begin by assuming a group count based on some priors and doing the first level inference of the manifold using a self-organizing map (SOM). The SOM is for providing the group for data points, which might be inaccurate, but it does give the idea for locality, which is useful for the algorithm to make learning adaptive. Additionally, if there are too many features, data can first be decomposed into singular values using SVD, and then singular vectors can be used with SOM. The distance among data-points was calculated based on two factors for manifold estimation by SOM. First is membership of a particular data-point in the group. Second factor is the distance between means of two groups.

It was important to calculate the likelihood of proximity (LP) for data points. We termed it as m_{ij}. Here i and j are two data points. This depends whether points are in the same group or not. For points associated with the same cluster, this aids in conserving local structure.

- The data-point i and data-point j are in two different cluster: We term M_{ij} as likelihood for proximity of two data-point i and data-point j from two clusters having center "C_i" and "C_j" as

$$M_{i,j} = e^{-dist\,(C_i, C_i)^2 \,/\, \rho} \qquad \ldots(1)$$

here

$i, j \;-\; Datapoints$

C_i, C_j − *centers of different SOM Clusters (calculated by averaging its data − points)*

M_{ij} − *likelihood of proximity between i and j belonging to* C_i *and* C_j

ρ − *spread factor (similar to variance in Gaussian Distribution)*

dist (C_i, C_j) − *distance between two data points i and j*

The *dist* (c_i, c_i) is normalized by dividing the mean of distances between cluster centers.

- The point *i* and point *j* are in same cluster: the value of M_{ij} is estimated by their distance d_{ij}.

$$M_{ij} = e^{-\delta d_{ij}^2 / \rho} \quad \text{M} \quad \text{...(2)}$$

here

i, j − *Datapoints*

d_{ij} − *Distance between data points i and j*

ρ − *spread factor (similar to variance in Gaussian Distribution)*

$\delta \quad - \ multiplying\ factor)$

Then, the scaling factor is as follows:

$$S_{ij} = \frac{1}{1 + m_{ij} - 1/(1 + m_{ij}) + e^5} \qquad ...(3)$$

$Here\ e^5\ is\ a\ pseudo\ -\ count.$

The term e^5 stops this factor to go to infinity.

Now, the distances from one point to other shall be found using:

$$\widehat{d_{ij}} = d_{ij} . S_{ij} \qquad ...(4)$$

Using equation 4, scaled distance can be calculated. Now, it is desirable to transform this $\widehat{d_{ij}}$ matrix into reduced dimension space. To achieve this, non-metric MDS technique was used. Now, let us deep dive into equation 3 again. This equation is giving a spring effect, which means it simulates both attraction and repulsion between data points based on the matrix calculated using equation 4. The relationship between m_{ij} and S_{ij} is quite evident in equation 3. If m_{ij} increases then S_{ij} reduces, which gives an attraction effect. If m_{ij} decreases then S_{ij} increase. It gives the effect of repulsion.

2.2.1. Balancing local and global Topology by modifying the Spread Factor (SF):

Referring to equation 1, spread factor (SF), termed as ρ, and optimality of global and local structure in the data-points may be understood. Low value of SF, causes higher value of LP. Then, this LP causes a small value of local structure. When SF is high, longer distances are taken into account during scaling, which preserves the global structure of distances although it may hide details about the local topology. As a result, optimizing TPDS for SF is a crucial procedure. To proceed this procedure, we chose to balance two costs by using a parameter **λ.** The first cost is global structure strain whereas second is local structure based strain. TPDS selects the value that results in the lowest cost-function (CF). We calculate it as:

$$CF = C + \lambda . E \quad \text{... (5)}$$

Here

λ : *Lagrange multiplier* and

C: *distance based stress used for global topology preservation*

Although the default value for λ in TPDS is 1, other values are conceivable. To find the minimum value of, we conduct grid search by taking the fixed step size, but we do not ascertain the difference between the cost function as shown in Equation-5. In method 1, a grid search is conducted by increasing a factor by the standard deviation of the distance matrix. The sole purpose of the non-metric MDS technique is to identify an appropriate scaled proximity matrix for dimensionality reduction. Consequently, TPDS permits users to customize the level of local topology preservation. MDS utilizes a cost function known as "distance-based stress," which is the sum of high- and low-dimensional distance disparities.

$$C_{ij} = \sum_{i \neq j} [d_{ij} - d'_{ij}]^2 \quad \text{... (6)}$$

here

d_{ij} – *distance between two data points i and j in higher dimension space*

d''_{ij} – *distance between two data points i and j in reduced dimension space*

is distance in reduced dimension. Often referred to as at distance stress, the cost given in equation (5) is what we refer to as MDS-cost in this context.

It is possible to formulate strain from preserving the local topology using various methods utilizing the hard K-Nearest Neighbour or smooth neighborhood-preservation [18]. The cost function connected to "symmetric

SNE'' was employed as a means of testing our model. The notion of joint probability distribution (JPD) is also very relevant and we will refer to it as JPD. As suggested in referred work [8] ”symmetric-SNE” (symSNE) is an adaptation of the SNE method. Then symmetric SNE''s cost function is used, which is termed as E. The E is the Kullback–Leibler (KL) divergence across the JPD, P in space with high dimensions and the JPD Q in space with low dimensions.

$$E = KL(P|Q) = \sum_i \sum_j p_{ij} . log \frac{p_{ij}}{q_{ij}} \quad \text{...(7)}$$

Where:

$p_{ii} = 0,$

$q_{ii} = 0,$

$Symmetric\ SNE \Leftrightarrow p_{ij} = p_{ji} \wedge q_{ij} = q_{ji}$

$$p_{j|i} = \frac{exp(- || x_i - x_j ||^2 / 2\sigma_i^2)}{\sum_{K \neq I} exp(- || x_i - x_k ||^2 / 2\sigma_i^2)} \quad \text{...(8)}$$

and

$$p_{ij} = \left(p_{i|j} + p_{j|i}\right) / 2n \quad \text{...(9)}$$

Where n is the total number of observations. We maintain σ_i at its current perplexity value of 15. The likelihood distance is calculated in the reduced dimensional space:

$$q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{K \neq I} exp(-||y_i - y_k||^2)} \quad \text{...(10)}$$

2.2.2. Parameter Optimizing

Numerous internal parameters of TPDS have been fine-tuned for optimum performance. The number of iterations and the stages used to seek for the spread factor in order to reduce Equation-5 do not need optimization. While distinct five to ten SFs were selected in the grid search, TPDS can accommodate larger steps if necessary. However, the quality of the outcomes produced by a technique depends on two external parameters of crucial importance. The Langrange multiplier (twill be used as LM, now onwards) is the first variable defined within Equation 5. Spread factor can be calculated using value, as we have seen previously. The SOM grid dimension is the second variable of independence. Keep the SOM grid size proportional to the estimated number of classes for enhanced visualization. Nonetheless, this is not a prerequisite for further analytic techniques. Even with a grid size of over 50, TPDS is capable of producing respectable results. Here, grid sizes ranging from 20 to 60 have been evaluated. [19]

2.2.3. Experiment Details

We evaluated the effectiveness of TPDS with two different dimension reduction techniques that had been shown to be effective on various types of data sets. Additionally, the goal of TPDS differs slightly from that of t-SNE in that it aims to enhance only data visualization. We wanted to evaluate the effectiveness of the mentioned technique. We utilized four datasets for this purpose. Two datasets were from physiological characteristics. Other than these datasets, one dataset was of protein expression and one was of single cell gene expression profiles. One datasets used in this study included features taken from Parkinson's disease and normal people's speech [20]. Each replicate for a specific person was treated as a separate data point. "The International Classification of Functioning, Disability and Health for Children and Youth (ICF-CY)" is the basis for the other physiological data set, which is referred to as "Self-Care Activities Dataset based on ICF-CY (SCADI)" [21]. The SCADI data set includes ICF-CY-based self-care attributes for seventy kids with physical and motor disabilities. This protein expression data-set was created from samples having the mouse cortex"s nuclear fraction to measure the expression levels of seventy seven proteins and protein modifications [22]. We also made use of a single-cell expression profile data set that included RNA-seq read counts from cell from different cell-lines [23]. Outliers in the "single-cell expression" data set provided a new perspective for assessing the strength of various techniques for reducing the dimensions. For the purpose of validation we used some datasets. In order to be able to do assessment using these datasets, we used their labels. We used "Density-Based Spatial Clustering of Applications with Noise (DBSCAN)" and

k-means, two different clustering approaches, to assess the usability of dimension reduction techniques [24].

2.2.4 Key parameters

The performance and results of the Topology Preserving Distance Scaling (TPDS) method are influenced by several key parameters, which also differentiate it from other methods. These parameters and their potential impact are as follows:

Spread Factor (ρ):

- The spread factor is critical in the TPDS method. It is adjusted using a grid search approach, typically using 5–10 values, though more can be used if necessary.
- The spread factor directly influences how distances among data points are scaled, impacting the balance between preserving local topology and global distance structures. Its optimal value is crucial for accurately capturing the underlying structure of the data.

Lagrange Multiplier (λ):

- The Lagrange multiplier λ is another essential external parameter in TPDS. It plays a role in the calculation of the spread factor.
- The value of λ influences the cost function used in TPDS, impacting the balance between distance-stress (global structure) and local-topology preserving strain. Adjusting λ can tailor the TPDS method to different datasets and analysis needs.

Grid Size of Self-Organizing Map (SOM):

- The grid size of SOM is used in TPDS to get a preliminary estimation of groups and the initial prediction of the manifold.
- The choice of grid size can impact the visualization quality of the results. A larger grid size (above 50) can still yield satisfactory results, although it

may not be necessary for downstream analysis methods like supervised classification or machine learning.

In comparison to other methods, these parameters in TPDS provide a unique ability to control and balance between preserving local data structures and global distance relationships. This is distinct from methods like PCA, which lacks this balance, or t-SNE, which heavily focuses on local structure preservation but less so on global relationships. The configurability of these parameters in TPDS allows for more tailored and potentially more accurate analyses for complex datasets, especially those common in biomedical research.

2.3. Results

We used to have data sets for TPDS evaluation that required classification as well as visualization in order to draw conclusions that were useful. Naranjo et al. [20] generated the initial data set utilized for this evaluation. It contains features that were taken from speech recordings of healthy people and patients with Parkinson"s disease. There are 44 features in 4 replicated data points for each patient or individual. In the original manuscript, the dataset was categorized into two distinct groups using a supervised classification approach. The first is normal, and the second is Parkinson. We used our unsupervised technique to decrease the dimension.

Figure 1 from the first data set [20] shows the data points for Parkinson"s disease and normal in two different colors. Different other approaches could not provide the separate data-points from separate classes. The TPDS" cost was less than that of t-SNE, and it clearly showed the separation among two cases, namely Parkinson"s and ordinary. In addition, we clustered the result of TPDS

and remaining techniques using k-mean approach(k = 2). With the use of "adjusted Rand index (ARI)" and "normalized mutual information (NMI)" , we determined the purity of the clustering. The ARI and NMI scores for K-mean clustering on TPDS output were nearly 1.7–2 times higher than those for the other methods ("tSNE, non-metric MDS, and Sammon mapping"), indicating that this method had the highest purity. Then by using DBSCAN to cluster data points in lower dimensions, we checked the clustering purity. Even when using DBSCAN for classification, TPDS achieved the highest degree of clustering purity.

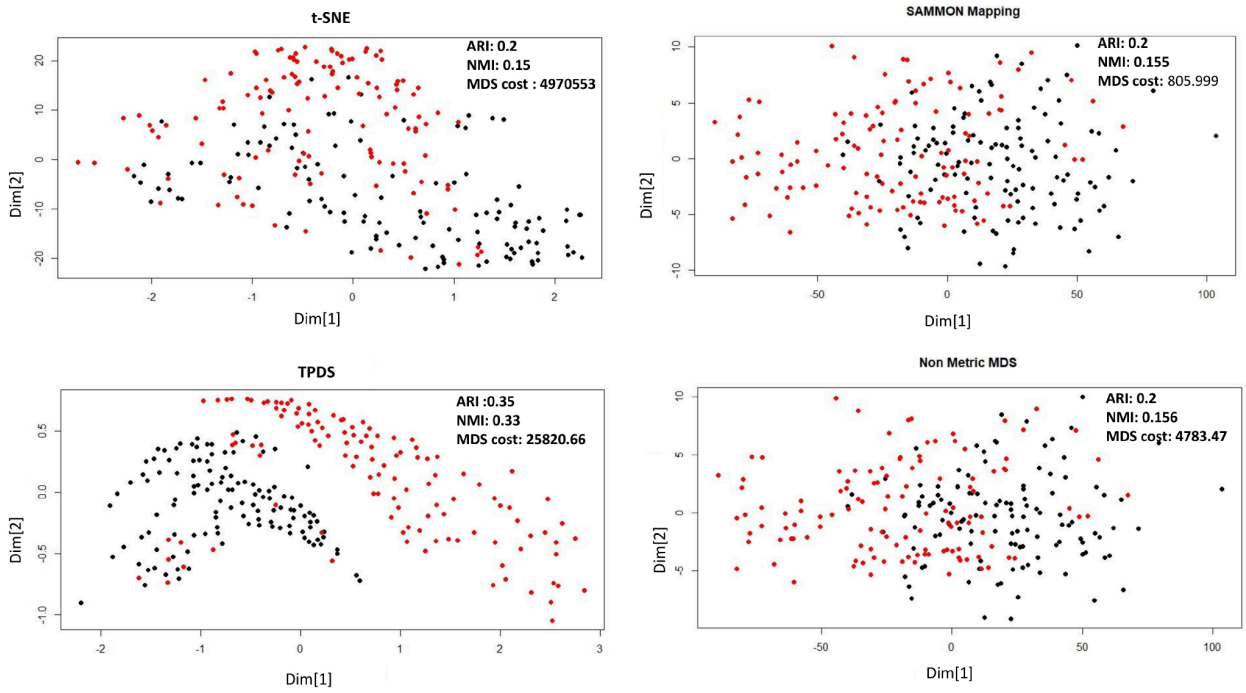

Figure 2.1. Reducing the Dimensionality of Parkinson data set by different algorithms.[25]

Data-Set	TPDS	t-SNE	Non-Metric MDS	Sammong Mapping
Parkinson's	ARI: 0.43	ARI: 0	ARI: 0.0	ARI: 0.0
Naranjo et al.	NMI: 0.33	NMI: 0	NMI: 0.0	NMI: 0.0
Mouse Protein	ARI: 0.133	ARI: 0	ARI: 0.004	ARI: 0.002
Higuuera et al.	NMI: 0.23	NMI: 0	NMI: 0.078	NMI: 0.068
SCADI	ARI: 0.288	ARI: 0	ARI: 0.058	ARI: 0.002
Zarachi et al.	NMI: 0.355	NMI: 0	NMI: 0.060	NMI: 0.068
single cell Expression	ARI: 0.027	ARI: 0.0022	ARI: 0	ARI: 0
Li et al.	NMI: 0.068	NMI: 0.024	NMI: 0	NMI: 0

Table-2.1. Result: clustering purity after applying DBSCAN [25]

Further evaluation of TPDS is done on the next dataset. This dataset is expression data of seventy seven “proteins/protein modification” from mouse"s cortex [22] . Thirty four trisomic mice having down syndrome and thirty eighth control mice were used to create this protein expression data set. The shape of control data was 38 * 15, where as trisomic mice data was 34 * 15. According to genotype, behavior, and treatment, the mice themselves were divided into eight groups. Using genotype mice, control or trisomic groups could be separated [22].

The mice might be categorized as stimulation based or without stimulation to understand the context-shock based on their behavior. Mice could also be categorized as having received drug treatment or not. To start with, we represented samples in reduced dimensions visually by assigning eight distinct groups to various colors. It was challenging to tell which approach worked better just by looking at it. Although the t-SNE method did not overcrowd samples, it dispersed them too widely, which caused group mixing. When using TPDS, some samples appear to be crowded, but for the majority, the neighborhood was based on their local topology based grouping. On the other hand, "Sammon mapping" and "non-metric MDS" enabled the neighborhood of samples as highly

overlapping strata in accordance with different groups. When compared to other tested methods, "K-means clustering" and "purity of classification" calculations produced higher NMI and ARI scores for TPDS. We verified the enhancement in clustering brought about by TPDS using DBSCAN. Therefore, we represented the protein expression data , which actually belongs to high dimensionality space, in reduced three dimensions using TPDS. The TPDS 3D scatter plot was able to clearly distinguish between mice that were trisomic and mice that were not. The TPDS 3D scatter plot was able to clearly distinguish between mice that were trisomic and mice that were not. While TPDS and other methods were unable to provide a comparable level of separability.

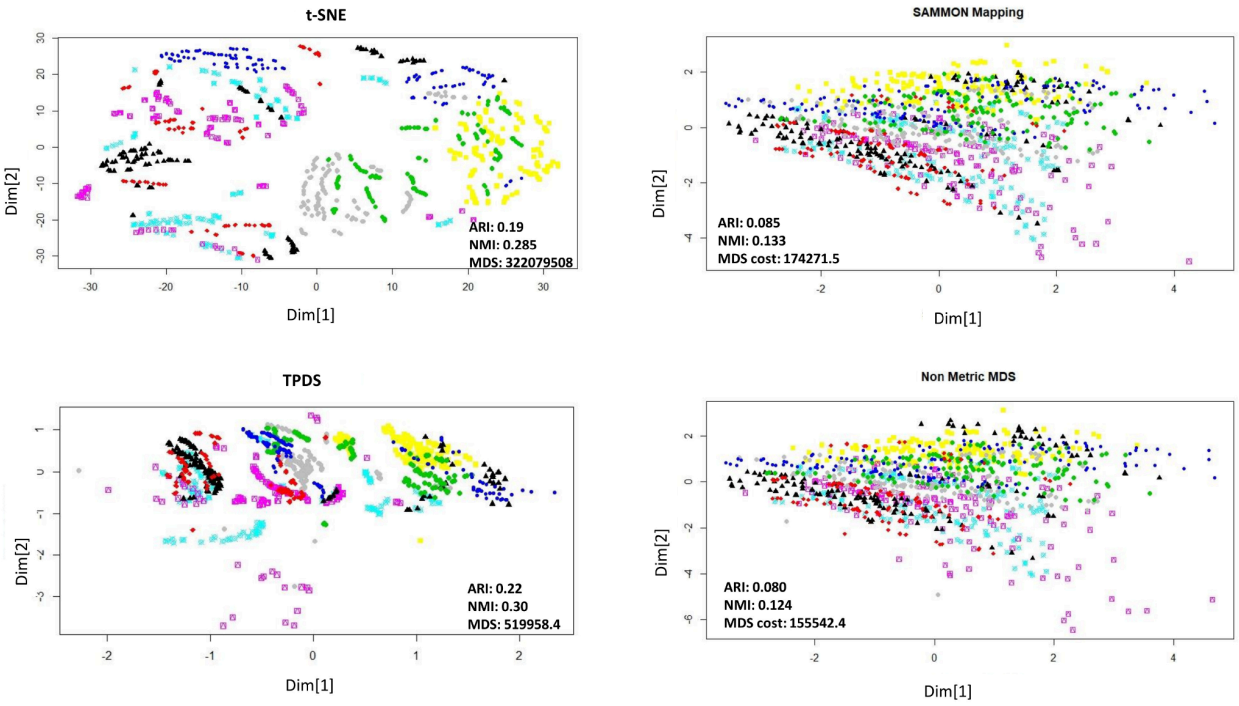

Figure-2.2. Dimensionality reduction applied on mouse protein data.

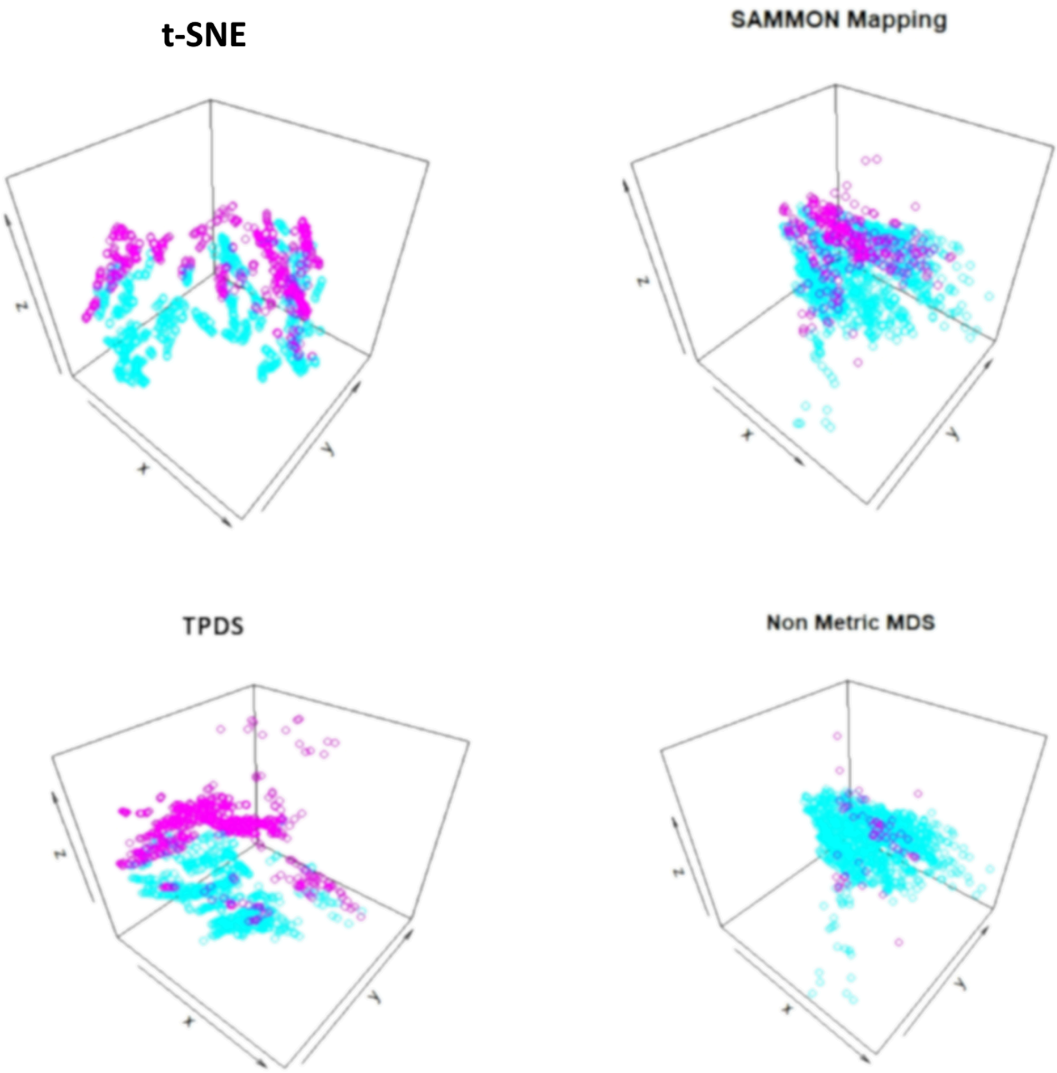

Figure 2.3. Dimensionality Reduction on mouse data with three dimensions.

Then, we moved to one another dataset which is also very relevant and known as SCADI [21]. This dataset may be taken from UCI-ML repository. This dataset is about seventy kids having physical and motor disabilities. There are 206 features for each kid. So, the shape of the dataset is 70 * 206. Then, samples were grouped into 7 classes [21] based on kids' behavior. Using our method along with the other three methods showed that some classes could be clearly separated in

smaller dimensions. An analysis of the separability after dimension reduction of the SCADI data set, on the other hand, revealed that the classification of TPDS output produced significantly higher clustering purity than other tested methods.

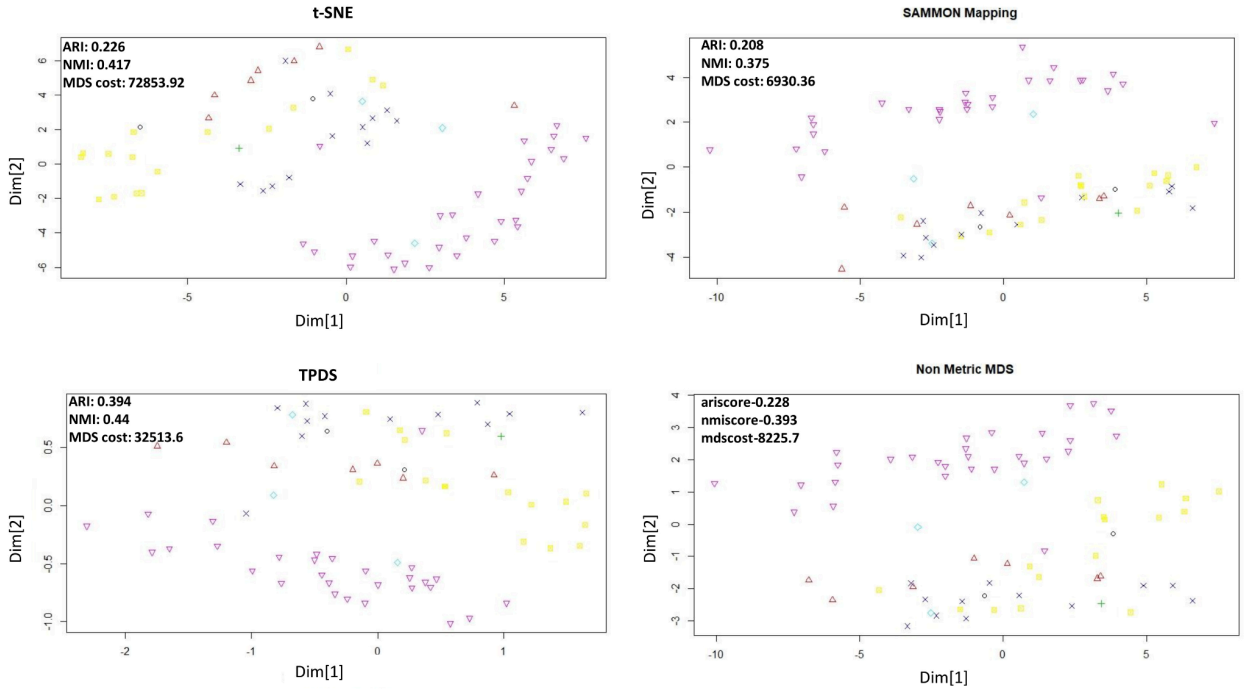

Figure 2.4. Analysis of TPDS by reducing the dimensionality of "SCADI" data set.

Then assessment of "TPDS" was performed by using the "Single Cell Expression" (SCE) data. The SCE data was extracted from a publication[23]. In this dataset, there were 57,242 genes" read counts. These read counts were utilized as features of the data. Then there were 562 cells; which were utilized as samples. "Principal component decomposition" (or "SVD") is frequently utilized for reducing the dimensionality for datasets having high dimensions. As a result, we loaded the top 30 principal components into each method.

The SCE datasets used in this study were having label for every sample and were able to be divided into seven different samples, which was used as class names. Differentiable loci were visible in the "t-SNE" and "TPDS" outputs for cells of various types (classes). "Sammon mapping" and "non-metric MDS" were inferior to TPDS in terms of performance when it came to clustering purity utilizing "k-means" and "DBSCAN". Be aware that the performance of t-SNE varies at various perplexity values. However, TPDS, which makes use of the MDS function, performs about as well as "t-SNE" for SCE data-set. Surprisingly, despite good visualization due to the preservation of local topology among cells, the distance stress cost (MDS-cost) for TPDS for SCE data-set was less compared to other techniques. It raises the possibility that outlier cells could cause convergence problems for Sammon mapping and non-metric MDS methods. It is important to note that despite the fact that TPDS employs the "non-metric MDS" method as the final step, the samples do not collapse because distance scaling maintains "local topology". The "TPDS" model seems to have avoided local-minima by giving low weight to outliers that are far apart.

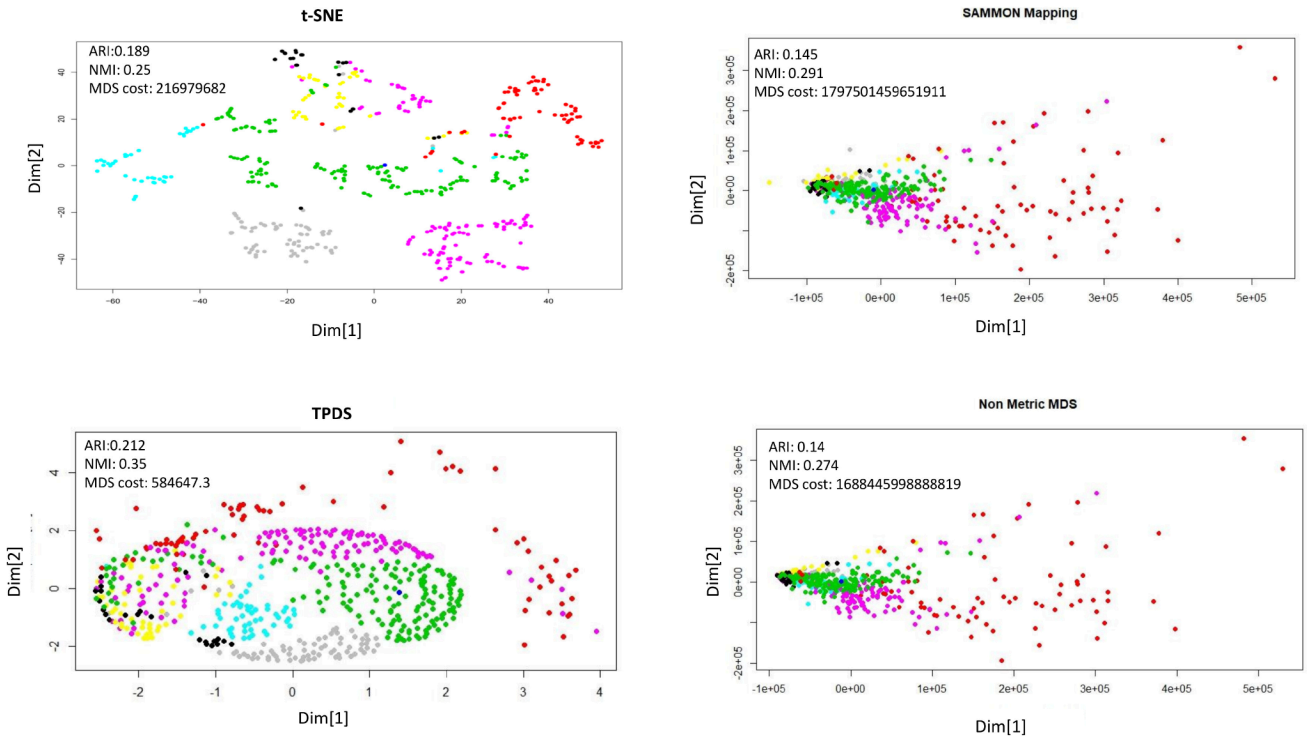

Figure 2.5. TPDS assessment by reducing the dimension for single-cell expression data.

2.4. Discussion

Classification, regression, and anomaly detection are just a few of the processes for large data sets analysis. All these processes could become more efficient and fast by utilizing the reduction of dimensionality. Considering the diversities in the biomedical data set, there are many potential influences on classification. As a result, techniques like t-SNE that are only intended for visualization might not be the best for improving the efficiency of analysis for the majority of datasets. Maintaining distance during dimension reduction might be advantageous for being able to distinguish clearly between different data points. By successfully balancing local-topology and distance preservation, TPDS was

able to produce low "MDS-cost" (distance stress), adequate visualisation, and improved "clustering purity" for tested data sets. This suggests that other analysis procedures may also benefit from achieving this balance.

By keeping some information about local topology, the TPDS method of scaling distances can also help to improve the convergence of "MDS-like" techniques. It can be justified by possibility that outliers caused by greater distances between data points could prevent convergence. Given that the fewer distances between samples within identical group may be numerous, giving them more weight results in a better overall MDS cost reduction. It is also possible to preserve some local topological information during this process for improved visualization. The impact of outlier data points on other "distance-preserving" techniques is very high. It is so great that it leads to data point collapse and longer convergence computation times. As a result, by giving outliers less weight than the standard distance-preserving method, TPDS clearly outperforms it. Along with "separability among data points" of different classes, results from TPDS also have dense "co-localization among data points" of the identical classes, and this is without location collapse. When DBSCAN uses such tight co-localization for density-based clustering, these results lead to noticeably higher classification purity for TPDS results. In order to support such a "density-based clustering approach" as well, TPDS could be used as a different dimension reduction method. In addition to being robust to outliers and possibly producing better classification outcomes, TPDS is discovered to be suitable in terms of time complexity. We are working to further optimize it for speed, though.

Here, we learned a rough estimate of the "manifold" using "SOM". Other methods generally use "KNN" or a modified version of it [18,26,27]. Additionally, we assessed the impact of SOM shape on "TPDS" performance. Every node in the rectangular geometry that TPDS by default employs has 4 neighbors. But SOM"s hexagonal shape was also examined. The outcomes of the SOM"s hexagonal geometry were almost identical to those of the default mode. Data points from the same class were closer to one another in "TPDS" due to the averaging effect when calculating "proximity likelihood", as seen with the protein expression data set. On the other hand, t-SNE was ineffective at determining the "likelihood of proximity for data points" in "protein expression data", most likely because of noise or sample-specific artifacts.

The ability of TPDS to preserve local topology and distances during dimension reduction can facilitate more accurate genomic data analysis, leading to better understanding of genetic basis of diseases and development of personalized treatments. In computational biology, where the analysis of complex biological systems is crucial, the TPDS method can help in better visualization and classification of high-dimensional datasets, such as protein-protein interactions and gene regulatory networks. This can lead to the identification of key drivers of diseases and aid in the development of targeted therapies.

One must make an educated guess regarding the number of SOM clusters (grid size) for visualization with TPDS, which may have the least impact on subsequent analysis steps like supervised classification and regression. Parameters are also necessary for other techniques like t-SNE and Sammon mapping. "Sammon mapping" uses lambda as the initial value of the step size

during optimization while t-SNE depends on the perplexity parameter. TPDS is not an exception to this rule because it depends on one or two parameters. We''ve presented a strategy to address a number of problems with the current distance preservation-based dimension reduction method, ensuring that ''heterogeneity among data points'' is preserved and that classification overall is improved.

Chapter 3

NIRD: Matrix Factorization based manifold learning algorithm, to Infer the large Regulatory Network to capture nonlinear dependency

*Developed a novel matrix factorisation based manifold learning algorithm, **NIRD**, for inferring the large Regulatory network to capture nonlinear dependency*

3.1 Introduction

This chapter is about the investigations that were performed to understand the second objective explained in Chapter 1. Our aim was to infer high-dimensional dependency/regulatory networks in lower dimensional space. We developed a novel matrix factorisation based manifold learning algorithm, NIRD, for inferring the large Regulatory network to capture nonlinear dependency.

Recent advancements with single-cell RNA sequencing (scRNA-seq) technology made possible the analysis of gene expression at the "single-cell level", offering unprecedented possibilities for investigating variation between particular cells, lineage trajectories, fluctuations in differentiation in cells, regulatory networks with high resolution, and the discovery of new cellular states. [28]. One such question is related to inferring the gene regulatory networks (GRN) from sparse scRNA-seq expression profiles [29]. Although many methods have been proposed for gene regulatory network (GRN) inference for bulk expression profiles, they have not proved to be suitable for sparse scRNA-seq profiles.

The gene regulatory networks (GRNs) are also useful for understanding the complex molecular mechanisms that govern cellular functions, for example development, differentiation, and disease. The accurate inference of GRNs from high-dimensional gene expression data has been a major challenge in computational biology and genomics, and various approaches have been developed to address this challenge.

There have been many different computational techniques to infer GRNs from the scRNA-seq data that can be put into a couple of primary groups: techniques

that utilize correlations and techniques that utilize models. The first type of techniques use correlation coefficients or mutual information measures to quantify the strength of associations between pairs of genes, and construct the network based on a threshold or a statistical test. For example, the SCENIC algorithm [30] uses cis-regulatory motif analysis to identify transcription factors (TFs) that regulate gene expression in different cell types, and constructs a GRN based on the co-expression of the target-genes and the TFs. Similarly, the GENIE3 algorithm [31] uses a random forest approach to estimate the importance of each gene for predicting the expression of its targets, and constructs a GRN based on the predicted interactions. GENIE3 also appeared as top performer in DREAM4 network inference challenge. “Gene Network Inference with Ensemble of trees (GENIE3)”, a tree based method, which has been used for reconstruction of gene regulatory networks (GRNs) from high-dimensional microarray gene expression data. Additionally, grnBoost2 [32] is also a well-known algorithm for gene regulatory network (GRN) inference from single-cell gene expression data. It is an extension of the original grnBoost algorithm, which was developed for bulk RNA-seq data. This algorithm utilizes a gradient boosting framework to infer the directed edges of the GRN by iteratively training regression models to predict the expression levels of genes based on the expression levels of their potential regulators.

Model-based methods, on the other hand, use mathematical models to describe the dynamics of gene expression and infer the regulatory interactions among genes. For example, the SCODE algorithm [33] uses a dynamical systems approach to model the time-series scRNA-seq data and infer the regulatory interactions among the genes. The MOGAMIT algorithm uses a generalized

linear model to describe the relationship between the TFs and the target genes, and incorporates prior knowledge from external databases to improve the accuracy of the network inference. Other methods for inference of gene regulatory networks (GRN) from scRNA-seq dataset is Partial Information Decomposition and context (PIDC) method uses multivariate information theory to investigate statistical relationships among triplets of genes and uses partial information decomposition (PID) for exploring gene regulatory relationships [34].

Recently a group of researchers [4] evaluated 12 network inference methods using experimental and simulated single-cell expression profiles and concluded that the performances of all the tested methods were less than ideal. However they also concluded that 3 methods PIDC, GENIE3 and GRNBoost2 [32] were leading and consistent performers in terms of accuracy. This found to have the best performance in terms of both accuracy and computational efficiency among the tested algorithms.

Despite the success of some of the methods, there are still several challenges in inferring GRNs from scRNA-seq data, including the high levels of technical noise, the sparsity and heterogeneity of the gene expression data, and the need for scalable and interpretable methods. In addition, many existing methods only infer the indirect effects of genes on the network, and do not consider the direct effects of genes, which can be important for understanding the underlying mechanisms. Therefore, developing novel and effective approaches for GRN inference from scRNA-seq data is an urgent need.

To address these challenges, we propose a novel approach for inferring GRNs from scRNA-seq data, which is based on matrix factorization and non-linear regression. Our approach captures the essential features of gene expression variation in the reduced-dimensional space obtained by matrix factorization, and estimates the importance of factors and the direct effects of genes using non-linear regression techniques. This allows us to infer the underlying regulatory network at a single-cell resolution, and to validate and interpret the network in the context of the biological system of interest.

In recent years, matrix factorization has been increasingly used in computational biology and genomics for dimensionality reduction, feature extraction, and data integration. Our approach extends the use of matrix factorization to the inference of GRNs, and provides a more generic and flexible framework for modeling both linear and non-linear dependencies.

In summary, our proposed approach represents a novel and promising method for inferring GRNs from scRNA-seq data, and addresses some of the challenges faced by existing methods. We named this algorithm as *Network Inference in Reduced Dimensions* (NIRD). Our approach is simple, yet powerful, and can be applied to other types of high-dimensional data as well. Our method also leads to a conceptual advancement which can have wider application for modeling large numbers of linear as well as non-linear dependencies. Our method is based on the approach of matrix factorisation before using the non-linear regression approach to find the importance of factors using the non-linear regression approach. Further our method estimates the direct effect of genes using the importance of factors. Further for more elaborate understanding, we have utilized different matrix factorisation methods to check generalisability and

robustness of our approach. Here, we have demonstrated our results using mostly SVD (for linear approach) and Kernel PCA (for nonlinear) type of matrix factorisation.

The contributions of this paper are twofold:

- We present a fast procedure for inferring the large network using the NMF, compared with current state-of-art techniques for the network, which take computation time by more than an order of magnitude on the same data.
- We present different NMF variants, in terms of difference in constraints for factorization and also compare both time and accuracy, which leads to a novel model order selection method for NMF.

3.2 Methods

There are a variety of ways for Inferring the regulatory network from bulk data. But our goal is to build a regulatory network from the expression of mRNA in single cells data. Methods for processing single-cell data, however, face a number of obstacles that prevent them from performing optimally or from being feasible in terms of computational time.

3.2.1 Datasets

To evaluate the performance of NIRD, authors utilized three categories of datasets. First category was provided by the “DREAM5 challenge consortium”, which consists of four types of datasets. In Three datasets among those four, were derived from the primary expression profiles of the “bacterium Escherichia

coli", as well as the "single-cell eukaryotes saccharomyces cerevisiae" and "staphylococcus aureus". The fourth dataset was generated through the utilization of "GeneNetWeaver", an "in-silico network" that employs the "chemical Langevin equation" to simulate molecular noise in "transcription and translation". The authentic affirmative connections for each of the four datasets are also accessible.

As a second category of data, we started by using the "scRNA-seq dataset" of "mouse embryonic stem cells (mESCs)" [35]. We used genes-regulation-relationships put together by a different grouping to rate network inference in a fair way. [36]. When compared to "bulk samples", "single-cell expression" corresponds to having more "noise and dropout". Dropouts happen when actual expression is not found because of technology problems.

Then the third data category was young and old pancreatic cells using their scRNA-seq profile [37]. Martin et al. defined three age groups, namely juvenile (1 month - 6 years), young adult (21-22 years) and aged (38-54 years) [37]. The three sets of pancreatic cells were treated individually with NIRD.

3.2.2 Problem Definition

The issue of deducing "regulatory networks" from "gene expression data" is one that we attempt to solve. "Directed graphs" with p nodes, wherein every node denotes a gene and a link from gene i to gene j shows that gene i (directly) influences the expression of gene j, making up the targeted networks. We only look for unsigned edges, so a connection between genes i and j might indicate

that i is a repressor of j or an activator. Recovering the network merely from gene expression data under different situations is the objective of (unsupervised) "gene regulatory network (GRN)" inference. Genetic regulation is dynamic and combinatorial, therefore many types of measurements may be acquired, such as steady-state expression profiles after systematic gene knockdown or knockdown, or time series measurements after random perturbations.

To explain how we arrived at a solution to the network inference issue, we first frame it as a feature transformation problem, then tailor it to the context of tree-based ensemble techniques, and lastly restate it in the opposite form.

Mathematical Representation

In mathematical way, we define the dataset as as a sample of n measurements having p features/genes:

$$V = \left\{ v_1, v_2, v_3, \ldots, v_n \right\}^T$$

$$v_i = \{ v_i^1, v_i^2, v_i^3, \ldots, v_i^p \}$$

where

$n -$ *the number of samples and*

$p -$ *number of features or genes*

Network inference algorithms make their predictions about the underlying regulatory linkages between genes by using this learning sample as their foundation. The majority of approaches for making inferences about networks

rank the regulatory links in a network in order of importance, from most vital to least necessary. The creation of a helpful network prediction is achieved by adding a threshold to this rating. Within the scope of our investigation, not only do we investigate the speed with which data may be transformed into a network, but also the accuracy of inferred dependencies and genes.

3.2.3 Forward Projection: Encoder

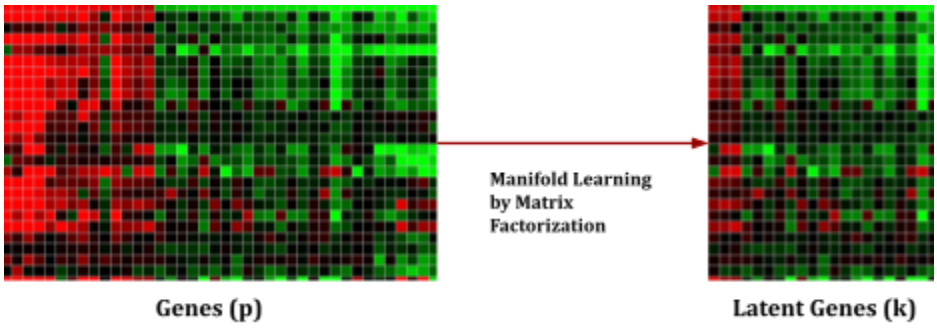

Fig-3.1: Transformation of higher dimensional datasets with n genes into lower dimensional datasets with k latent genes

To achieve the mentioned task, we first transform the dataset from higher dimensional space to reduced dimensional space. This task of transformation can be achieved from multiple ways. We choose a very well known technique called (non negative) Matrix Factorization (MF). There are various ways to perform MF. We used two python packages to achieve this, one is SKLearn and NIMFA. We used multiple matrix factorisation based techniques: Bayesian Decomposition (BD), Bayesian nonnegative matrix factorization (BNMF), Binary Matrix Factorization (BMF), Iterated Conditional Modes nonnegative matrix factorization (ICM), Fisher Nonnegative Matrix Factorization for learning Local

features (LFNMF), Alternating Nonnegative Least Squares Matrix Factorization etc, to this goal.

As discussed in chapter two of this thesis, unsupervised learning algorithms like principal component analysis (PCA) can be viewed as the factorization of a data matrix under various constraints. The nature of these constraints significantly influences the properties of the resulting factors. In the case of PCA, a weak orthogonality constraint is applied, leading to a highly distributed representation that relies on cancellations to capture variability in the data [38,39].

It has previously demonstrated that imposing a nonnegativity constraint during matrix factorization is effective in extracting a "parts-based" representation from the data [40,41]. The learned nonnegative basis vectors are combined in a distributed but sparse manner, enhancing the expressiveness of the reconstructions [42,43]. The different algorithms provide different representation by learning the optimal nonnegative factors from data.

Non-negative matrix factorization (NMF) Given a non-negative matrix V, find non-negative matrix factors W and H such that:

$$V \approx W . H$$

Given a set of multivariate n-dimensional data vectors, the vectors are placed in the columns of an $n \times p$ matrix V where n is the number of examples in the data set. This matrix is then approximately factored into $n \times k$ an matrix W and an $k \times p$ matrix H. Usually k is chosen to be smaller than n *or* p, so that W and H

are smaller than the original matrix V. This results in a compressed version of the original data matrix.

Here each data vector v is approximated by a linear combination of the columns of W, weighted by the components of h. Therefore W can be regarded as containing a basis that is optimized for the linear approximation of the data in V. Since relatively few basis vectors are used to represent many data vectors, good approximation can only be achieved if the basis vectors discover structure that is latent in the data.

Cost Functions

To find an approximate factorization $V \approx W . H$, All the factorization algorithms define their own cost functions that quantify the quality of the approximation. Some of the basic cost functions are defined as follows:

The Euclidean distance between A and B

$$|| A - B||^2 = \Sigma_i^j (A_{ij} - B_{ij})^2$$

This is lower bounded by zero, and clearly vanishes if and only if A = B [44].

Another useful measure is

$$D(A \,||\, B) = \Sigma_i^j (A_{ij} log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij})$$

Like the Euclidean distance this is also lower bounded by zero, and vanishes if and only if A = B . It reduces to the Kullback-Leibler divergence, or relative

entropy, when $\Sigma_i^j A_{ij} = \Sigma_i^j B_{ij} = 1$, so that A and B can be regarded as normalized probability distributions. Proof of convergence has been provided elsewhere [45].

3.2.4 Model : Tree based ensembling

In this step our goal is to decompose the problem of recovering a network of k latent genes/features into k subproblems, each of which identifies a regulation of a latent gene/feature which is formed by linear or nonlinear combination of genes. During this process, the automatic feature score imputation also happens, to some extent. Using reduced expression data, regulatory latent genes/features for a target are identified as the subset of latent features whose expression directly controls or predicts the target feature's expression. This is a feature selection challenge in supervised learning. Our solution is based on tree-based ensemble methods' embedded feature ranking mechanism. To implement these trees, the python implementation uses the scikit-learn library.

3.2.5 Backward Projection (Decoder)

At this point in the process, the inferred network that was derived from the reduced-dimensional expression data is once more converted back into an actual high-dimensional vector space. In order to accomplish this, we computed the importance of features using the preceding step's calculation of the importance of latent features. In addition to this, we computed the feature contribution (LFC).

Feature importance (LFI) for latent features is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature. Normalized feature importance is first calculated on each decision tree and then at random forest level sum of this based on all trees calculated.

Different implementations of matrix factorization provide different ways to calculate the LFI and LFC. There are implementation where one can directly find mixture matrix H which is product of LFI and LFC with shape of $k \times p$ [46].

$$LFI = \{i_1, i_2, i_3, \ldots, i_k\}$$

$$LFC = \{c_1, c_2, c_3, \ldots, c_p\}^T$$

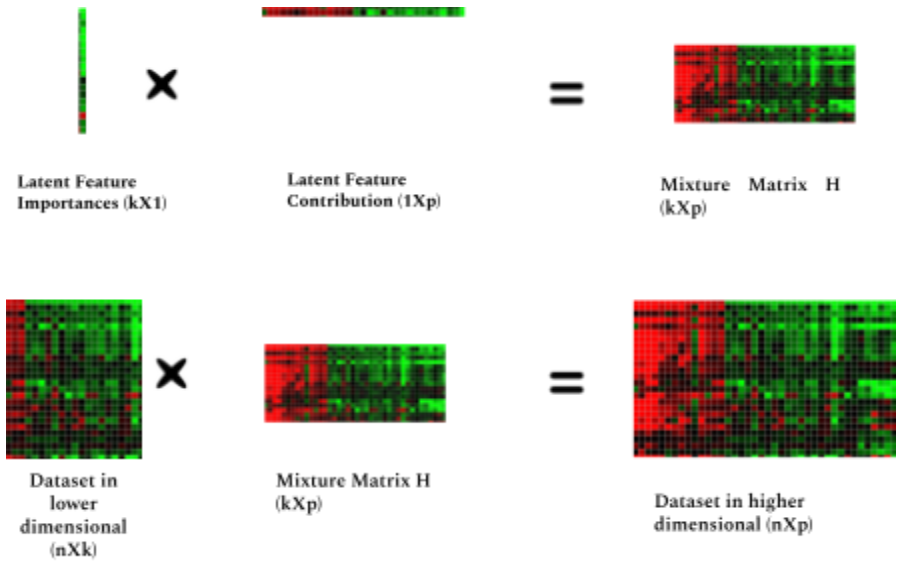

Fig-3.2: Reverse Transformation of lower dimensional datasets with k latent genes into higher (actual) dimensional datasets with p genes

Now, the product of the dataset in reduced dimension, latent feature importance and latent feature contribution are multiplied, the inferred network has been projected back to its actual shape.

3.2.6 Assessment of inferred network accuracy by using overlapping edges and true positives

In network inference, especially in complex fields like genomics or social network analysis, distinguishing between these two concepts is crucial for evaluating the performance and reliability of network inference algorithms or methods. The terms "overlapping edges" and "true positives of known edges" are often used in the context of network analysis, particularly when discussing the accuracy of inferred networks compared to known or established networks.

Overlapping Edges: This term generally refers to the edges (or connections) in a network that are found both in the inferred network and the reference or known network. In other words, they represent the intersections or common connections identified by both the inferred method and the established standard. The "overlapping" aspect highlights that these edges are present in both networks being compared. "Overlapping edges" focus on the comparison between the inferred network and a known network, highlighting the commonalities or agreements between the two. It's a measure of similarity or overlap.

True Positives of Known Edges: True positives refer to the edges in the inferred network that correctly represent actual connections existing in the real-world or known network. In network analysis, a true positive is an edge that is both present in the inferred network and is a valid connection in the real-world

scenario (i.e., the known network). The term emphasizes the accuracy of the inferred network in identifying real connections. "True positives of known edges" focus on the accuracy and validity of the inferred network in identifying real, existing connections. It's a measure of correctness or true discovery in the context of the known network.

3.2.7 Automatic Imputation in NIRD to handle outliers and dealing with noise

The automatic imputation feature in the NIRD method is a significant aspect that helps in managing outliers and noise in the data. This process is essential for ensuring the robustness and accuracy of the inferred gene regulatory networks.

What is Imputation: Imputation in data analysis refers to the process of replacing missing or unreliable data points with substituted values. This is often done to maintain consistency in datasets and ensure that analyses are not skewed by gaps or aberrations in the data.

Handling Outliers: In the context of NIRD, automatic imputation helps in mitigating the impact of outliers. Outliers can significantly distort the results of network inference, leading to incorrect conclusions. By identifying and replacing these aberrant values, NIRD's imputation process helps in normalizing the data, making it more homogenous and representative of the underlying biological reality.

Dealing with Noise: Similarly, noise in gene expression data, which can arise from various sources such as experimental errors or inherent biological variability, can obscure the true relationships between genes. The imputation

process in NIRD helps to 'smooth out' this noise, enhancing the signal-to-noise ratio and making the patterns of gene regulation more discernible.

Matrix Factorization and Imputation:

The matrix factorization step in NIRD is integral to its imputation capability. Matrix factorization breaks down the high-dimensional gene expression data into lower-dimensional representations, capturing the essential features while discarding redundancies and anomalies. This process inherently fills in gaps and smoothes out irregularities in the data, contributing to the automatic imputation.

Implications for Network Inference: This imputation capability of NIRD is crucial for accurately inferring gene regulatory networks. It ensures that the relationships between genes are not misrepresented due to outliers or noise, leading to more reliable and biologically meaningful network models.

Benefits for Downstream Analysis: The cleaned and normalized data resulting from the imputation process can lead to more accurate downstream analyses, such as identifying key regulatory genes or pathways. This is particularly valuable in systems biology and genomics research, where data integrity directly impacts the insights gained from the analysis.

3.2.8 Regularization techniques applied during matrix factorization

Regularization techniques are crucial in matrix factorization, especially when dealing with high-dimensional data, to prevent overfitting and improve the model's ability to generalize to new data. Here are some common regularization techniques used during matrix factorization:

L2 Regularization (Ridge Regularization):

How it Works: Adds a penalty term equal to the square of the magnitude of the coefficients to the loss function. The regularization term is the sum of squares of all feature weights, multiplied by the regularization parameter.

Impact: It shrinks the coefficients but does not reduce them to zero. This discourages large weights in the model, leading to simpler models that are less likely to overfit.

L1 Regularization (Lasso Regularization):

How it Works: Involves adding a penalty equal to the absolute value of the magnitude of the coefficients to the loss function. The regularization term is the sum of the absolute values of all feature weights, multiplied by the regularization parameter.

Impact: It can shrink some coefficients to zero, effectively performing feature selection. This leads to sparser solutions and can be particularly useful when dealing with high-dimensional data where feature reduction is desirable.

Elastic Net Regularization:

How it Works: Combines L1 and L2 regularization by adding both penalties to the loss function. The model is controlled by two parameters: one for L1 and one for L2 regularization.

Impact: It blends the feature selection capability of L1 with the stability and predictive power of L2, making it suitable for models where some coefficients are significant and others should be discarded.

In summary, regularization in matrix factorization plays a pivotal role in controlling model complexity, reducing overfitting, and enhancing the model's ability to generalize from the training data to unseen data. It is a fundamental aspect of creating robust and reliable predictive models, especially in complex data environments.

3.3 Results

3.3.1 Evaluation and insights using DREAM challenge expression dataset

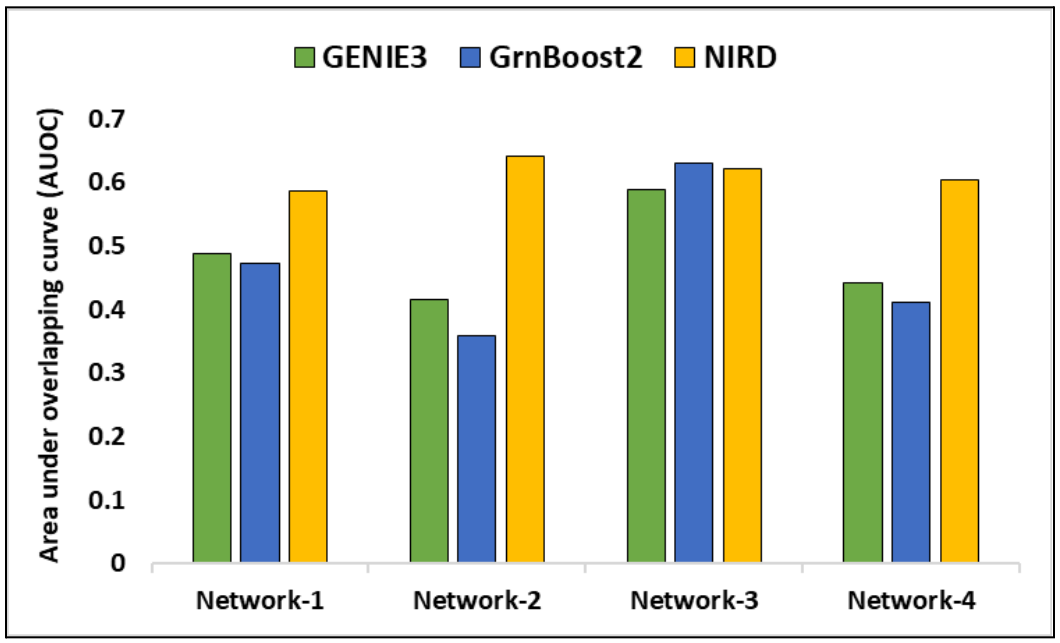

Fig. 3.3. Inferring the "gene regulatory network (GRN)" by NIRD of gene-expression. Performance of network inference using bulk gene-expression data-sets of DREAM5 challenge.

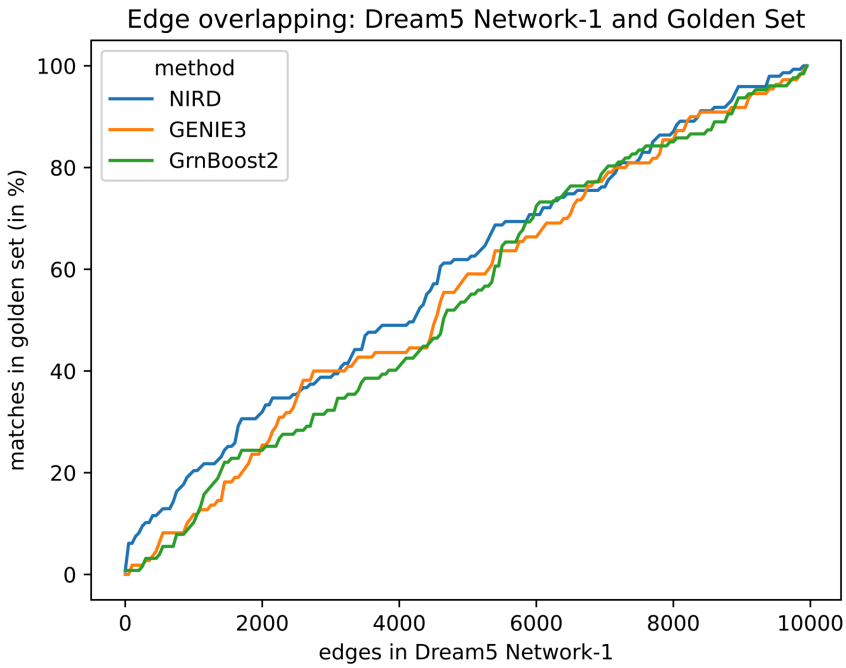

Fig.3.4. Comparison of different methods in terms of consistency in the prediction of the gene-regulatory network. Inferred networks from dream5 network-1 were compared.

In order to compare and evaluate the performance of network inference by the proposed method, which is NIRD, we utilized the concept of overlapping edges compared to the true positives of known edges. The overlapping edges means the number of similar edges, in dream-5 network-1 and the golden set which are true positives of known edges.

We first inferred the network using all three methods, namely, NIRD, GENIE3 and grnBoost2. These three networks encode the node to node regulation (or gene to gene regulation) in the form of edges present in the corresponding network. Then we identified the overlapping edges compared to known true

positives of edges. This may be named as overlapping edges and visualized as area under overlapping curve (AUOC). Hence, higher values of AUOC represent a method's ability to infer an optimal network compared to the true network. It is found that for all Dream-5 networks, our method NIRD is performing better than the state-of-the-art "gene regulatory network (GRN)" inference method GENIE3 and GRNBoost2 in terms of both time taken and edge overlapping with corresponding golden sets.

3.3.2 Evaluation and insights using Single-cell expression profiles of mESC

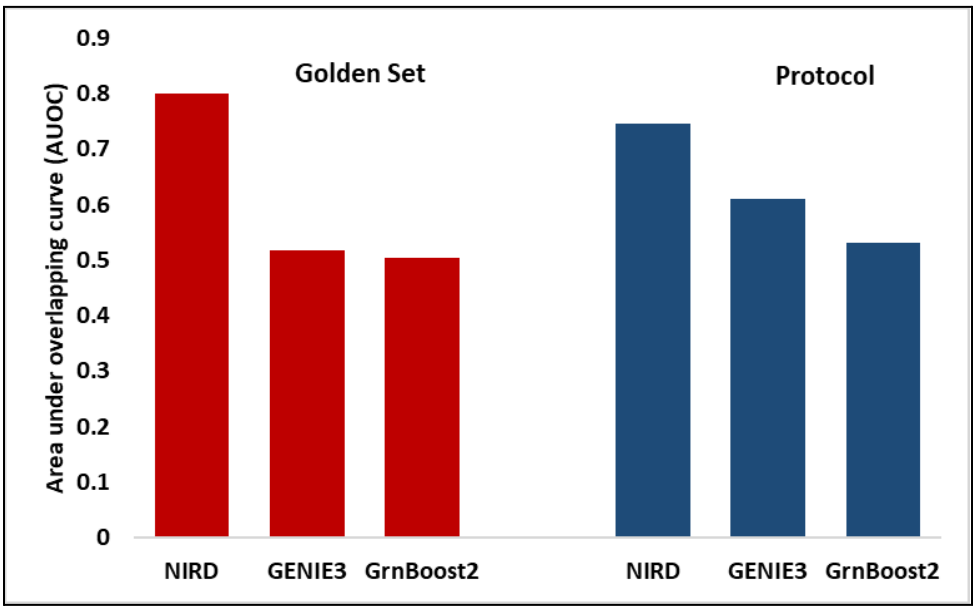

Fig. 3.5. Inferring the "gene regulatory network (GRN)" by NIRD of mESC data.

We validated the effectiveness of our method NIRD using single-cell expression profiles in two ways. In the first way, similar to what was done in previous para of DREAM5 challenge data, the area under overlapping curve (AUOC) compared to known positive interactions - golden set, was calculated using all three

methods. NIRD found to perform better from other methods both by time taken and performance. Second, we verified the accuracy of our technique using an additional strategy. We employed a metric to assess the similarity between two inferred networks for the same cell type, each affected by different technical biases and batch effects. A high degree of overlap suggests that the inferred networks from both datasets more accurately reflect the true gene interaction model. To evaluate this, we used two scRNA-seq datasets of mESC generated by different protocols (SMARTseq and DropSeq). Our findings show that NIRD outperforms other existing gene regulatory network (GRN) inference algorithms in both cases.

3.3.3 Evaluation and insights from single-cell profile for age based pancreatic cells

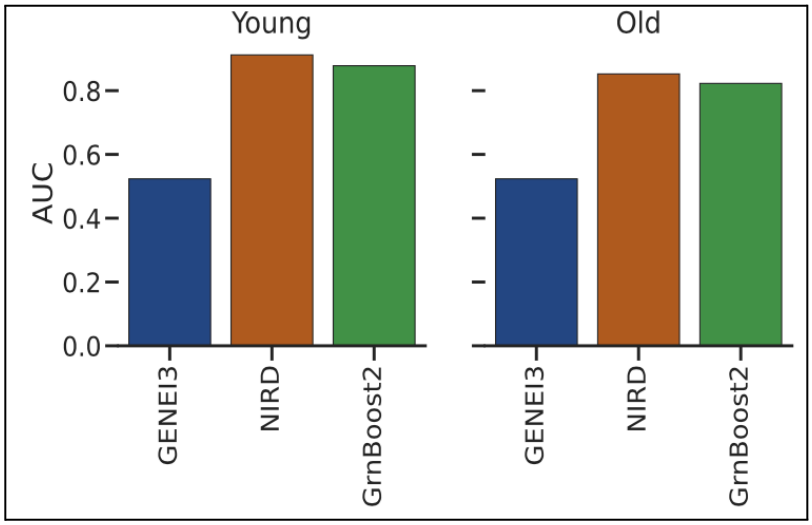

Fig. 3.6. Performance and analysis of noise for single-cell RNA-seq profile of pancreatic cells.

Improvement in overlap among inferred gene-networks from two expression data-set for a cell type also hints that predictions are closer to true

gene-interaction profiles. Such a notion inspired us to compare gene-networks inferred for young and old pancreatic cells using their scRNA-seq profile [37]. Authors utilized two age groups, namely young and old [37]. The two sets of pancreatic cells were treated individually with NIRD. Studying variations in the influence of genes has been done using differential centrality in the co-expression network (refer fig. 3.7). Yet, noise in single-cell expression patterns might result in fictitious centrality discrepancies. As a result, we demonstrated the varied gene expression levels in the network that was inferred from the scRNA-seq profiles of young and old cells. Following matrix factorization, network inference on low dimensional representation does automated imputation, which appears to lessen centrality discrepancies that may be brought on by noise irrationality. Further, we also performed gene ontology enrichment analysis. The enriched pathway terms for top 500 genes based on pageRank were used for this purpose. Refer fig 3.8, 3.9.3.10

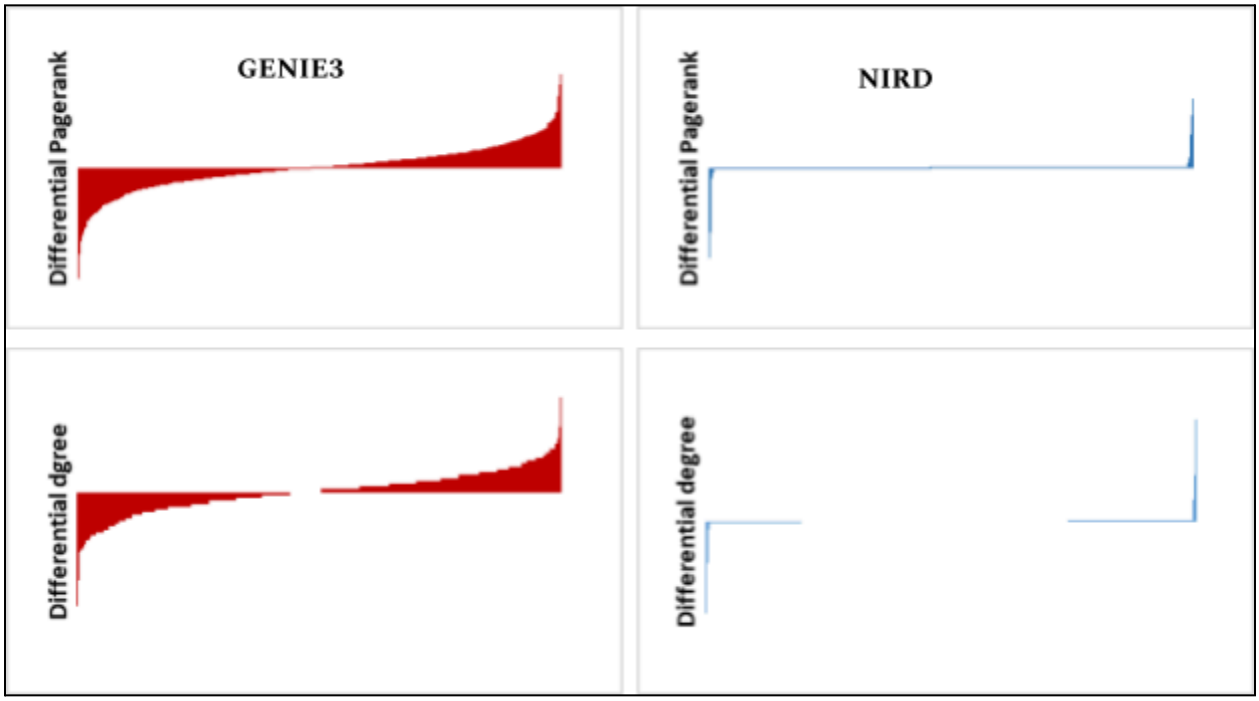

Fig. 3.7. Presence of noise in the estimated differential centrality in the networks inferred using Genie3 and NIRD. Automated imputation, happens Due to matrix factorization, appears to lessen centrality discrepancies that may be brought on by noise irrationality

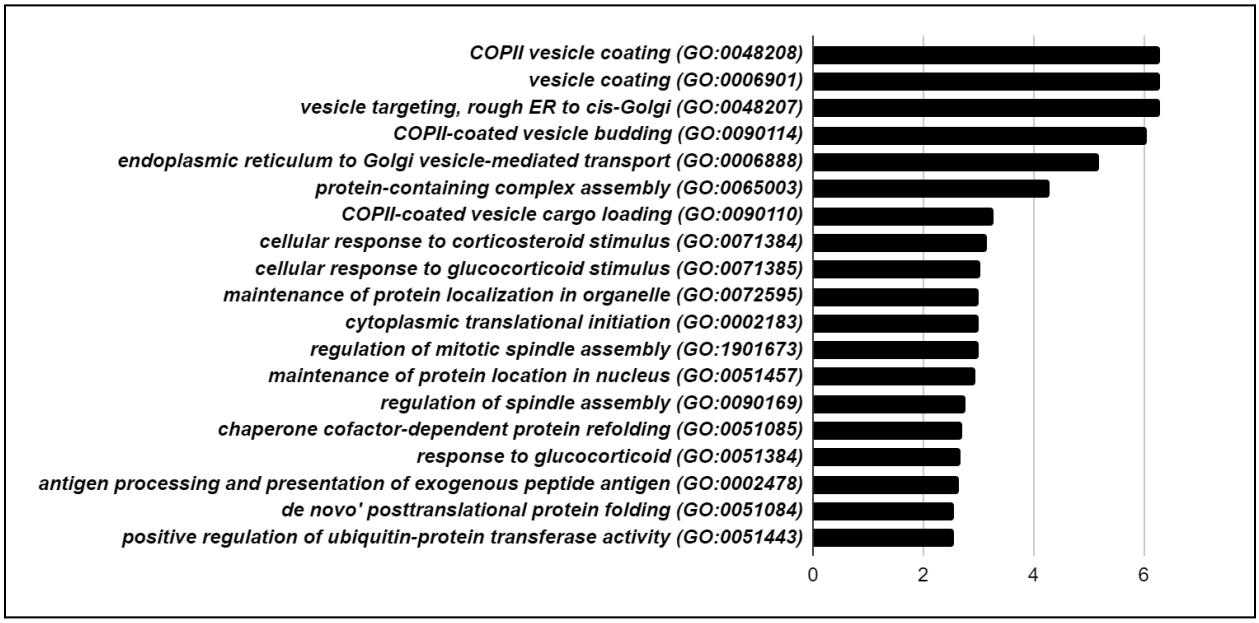

Fig. 3.8. Enriched Pathways in terms of GO biological process

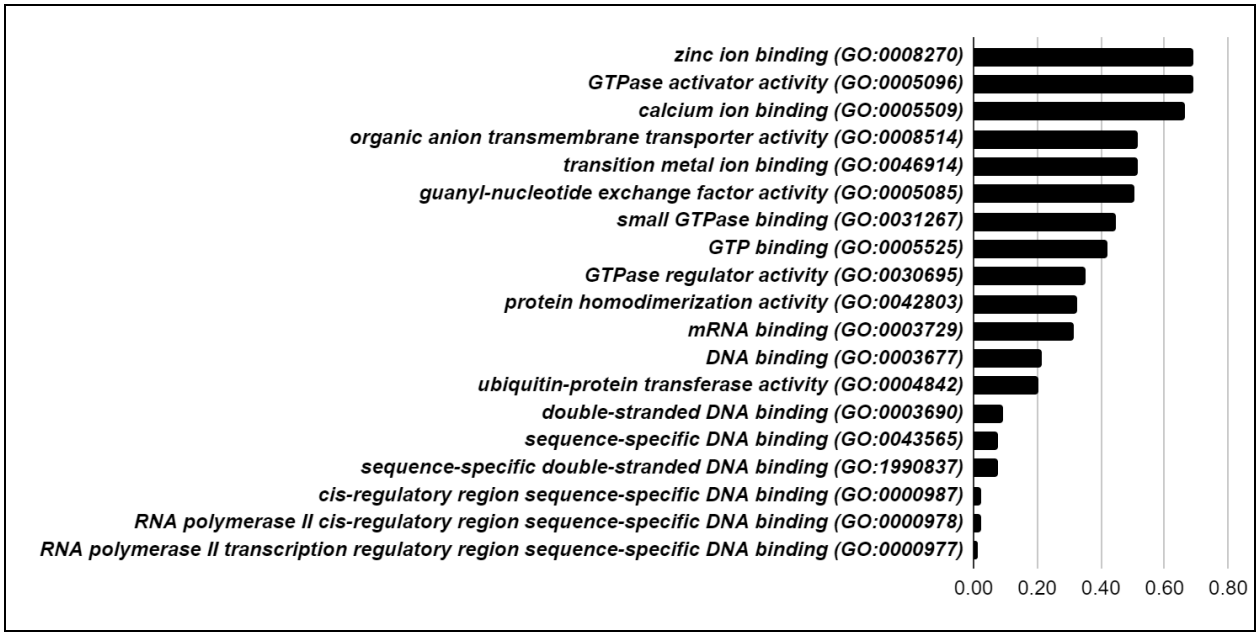

Fig. 3.9. Enriched Pathways in terms of GO Molecular function

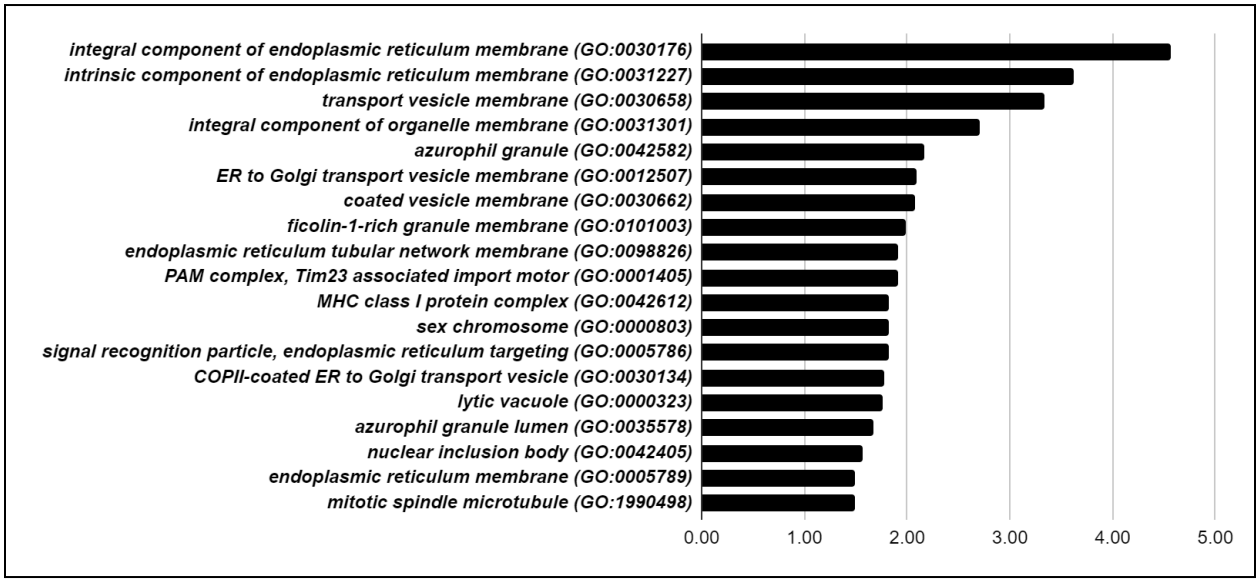

Fig. 3.10. Enriched Pathways in terms of GO cellular component

3.4 Discussion

Our results demonstrate that NIRD can effectively infer regulatory networks from gene expression data with high accuracy. The matrix factorization step helps to reduce the dimensionality of the data, which can improve the performance of the tree-based ensemble method. In our experiments, we compared the performance of NIRD with several existing methods for network inference from gene expression data, including GENIE3, and GrnBoost2. Our results demonstrate that NIRD outperforms these existing methods in terms of both accuracy and computational efficiency.

The proposed approach has important implications for understanding the regulatory mechanisms underlying complex biological processes. Inferring regulatory networks from gene expression data is an important problem in systems biology, as it can provide insight into the interactions between genes and their roles in various biological pathways. The ability to accurately infer regulatory networks from gene expression data can facilitate the development of new therapies and treatments for complex diseases, as well as improve our understanding of fundamental biological processes.

Certainly. Another important aspect to discuss is the significance of the proposed NIRD method in the context of "gene regulatory network (GRN)" inference. The results presented in this paper demonstrate that NIRD outperforms the state-of-the-art methods, GENIE3 and grnBoost2, in terms of both time taken and edge overlapping with corresponding golden sets. This

suggests that the proposed method is a promising approach for inferring "gene regulatory network (GRN)"s from high-dimensional gene expression data. Furthermore, the ability of NIRD to perform well on single-cell expression data, despite the high levels of noise and dropout, is a particularly valuable feature in the current research landscape.

One potential area of application for the NIRD method is in the field of personalized medicine. By inferring "gene regulatory network (GRN)"s from patient-specific expression data, it may be possible to identify specific regulatory pathways that are dysregulated in individual patients, leading to more targeted and effective treatments. Additionally, the ability of NIRD to perform well on single-cell expression data may allow for the identification of cell-specific regulatory networks, which could have important implications for understanding cellular differentiation and disease development.

Another important aspect to consider is the limitations of the NIRD method. While the results presented in this paper are promising, it is important to note that the method is still subject to certain limitations. For example, the NIRD method relies on the assumption that gene expression levels are non-linearly related to regulatory interactions, it relies on the availability of a gold standard set of positive interactions, which may not always be available or may be incomplete.

Another limitation of the NIRD method is that it is computationally intensive, particularly when applied to large-scale datasets. While the use of matrix factorization can help to reduce the dimensionality of the data and improve computational efficiency, it may still be challenging to apply the method to very

large datasets. This may limit the applicability of the method in certain contexts, such as when working with complex tissues or whole organisms.

The proposed approach can be extended to incorporate additional sources of biological data, such as epigenetic data or protein-protein interaction data. Combining multiple sources of biological data can provide a more comprehensive view of the regulatory mechanisms underlying complex biological processes. Second, the proposed approach can be evaluated on different datasets to further demonstrate its effectiveness and robustness. Finally, the proposed approach can be combined with other methods for network inference, such as Bayesian networks or causal inference methods, to improve accuracy and provide additional insights into the underlying regulatory mechanisms.

Despite the advantages of NIRD, there are several limitations to our approach that should be considered. First, matrix factorization is sensitive to outliers and noise in the data, which can impact the accuracy of the inferred network. Second, the choice of hyperparameters in the tree-based ensemble method can impact the performance of the algorithm. Further work is needed to evaluate the impact of hyperparameter tuning on the performance of NIRD. Finally, the proposed approach is limited by the quality and quantity of the gene expression data used as input. Future research should explore ways to overcome these limitations and improve the accuracy and robustness of the approach.

Overall, the results presented in this paper demonstrate that the NIRD method is a promising approach for inferring "gene regulatory network (GRN)"s from high-dimensional gene expression data. The ability of NIRD to perform well on

both bulk and single-cell expression data, despite high levels of noise and dropout, is particularly valuable. However, it is important to consider the limitations of the method, particularly its reliance on certain assumptions and the computational challenges associated with working with large-scale datasets. Future research will be needed to further validate and refine the NIRD method, and to explore its potential applications in various contexts, including personalized medicine and the study of cellular differentiation and disease development.

Chapter 4

Learning the Mental Health Impact in the United States with Explainable Artificial Intelligence

Analysis work utilizing PGM along with Markov Blanket for explainable feature learning

4.1 Introduction

This chapter is about the investigations that were performed to understand the third objective explained in Chapter 1. As already explained, our aim was to discover explainable public health relationships in high-dimensional healthcare data. We attempted to identify the driver factors for mental health.

People's mental well-being was affected by the COVID-19. Over 400 million job and half a million lives have been shattered, and this has triggered widespread fear, anxiety and sorrow. [47] . These effects are seen across the board, and they are sometimes felt more strongly in some groups, such as young individuals, frontline workers, caretakers, and people who have persistent medical illnesses. [48]. The emerging norm has resulted in unprecedented measures, such as statewide lockdowns, attempting to halt the spread, which have only served to further isolate people. Increases in isolation, depression, suicidal ideation, and the use of illicit substances are also expected. The Lancet Psychiatry has lately shed light on the needs of marginalized communities throughout this period, including individuals with serious mental disorders, difficulties with learning, and neurodevelopmental abnormalities, as well as those in jail, the homeless, and refugees. [49].

More early-career psychiatrists [50,51], technologies such telepsychiatry, and a focus on the high vulnerability of the forefront medical professionals itself [52]. have all been proposed as potential solutions to the problem's gravity.

Furthermore, it is expected that initiatives would have different effects on men and women, with women having a greater probability to face challenges related to informal care, economic inequities, and the closing of schools. Aging and

comorbidities status seem to have an immediate impact on susceptibility to mental health problems, as shown by the correlation between COVID-19 morbidity and mortality. Evidence suggests that afflicted communities typically suffer psychic suffering; this will be true of the COVID-19 pandemic's victims [48].

Finally, the impacts of social media are multifaceted [53,54], with some research suggesting an association among usage of social media as well as an increase in the incidence of psychological disorders [55]. However, most of these consequences were only investigated separately, with little effort performed to recreate the cumulative impact of these aspects. This research uses “Bayesian networks (BNs)”, a “explainable artificial intelligence” which represents the “joint multivariate distribution” (JPD) underpinning vast data from surveys collected throughout the United States, to bridge this knowledge gap. We also close the discrepancy in the prediction of psychological episodes like anxiety attacks by using supervised machine learning models.

4.2 Methods

In order to infer factors affecting mental health, we first performed a few statistical analyses and tests to check the sanity of the data-set and make initial inferences. Such statistical analysis and tests involved item reliability analysis and test of independence. Later we used Bayesian network inference and supervised machine learning model as explained below.

4.2.1 Data Set

We gathered information on 17764 individuals [56] using two weeks survey of the adult household population in the US conducted between April 20 and 26 and May 4 to 10, 2020, for 18 regional areas, including 10 states "(California, basic Colorado, Florida, Louisiana, Minnesota, Missouri, Montana, New York, Oregon, Texas) and 8 metropolitan areas (Atlanta, Baltimore, Birmingham, Chicago, Cleveland, Columbus, Phoenix, Pittsburgh)". In this research, we used information from both of the data collecting periods that had been completed as of the date of publication (25-May-2020). The original data's specifications are available elsewhere[57]. In summary, the data set included elements linked to social dynamics, insurance-related policies, economic security, and physical and mental health.

4.2.2 Analysis

The flow diagram for the analysis is shown in Figure 2a. Indicators including mental wellness, remote-work, COVID-19 symptoms, communication, persistent medical conditions, behavioral characteristics, insurance support, and other were included in the survey questions classification.

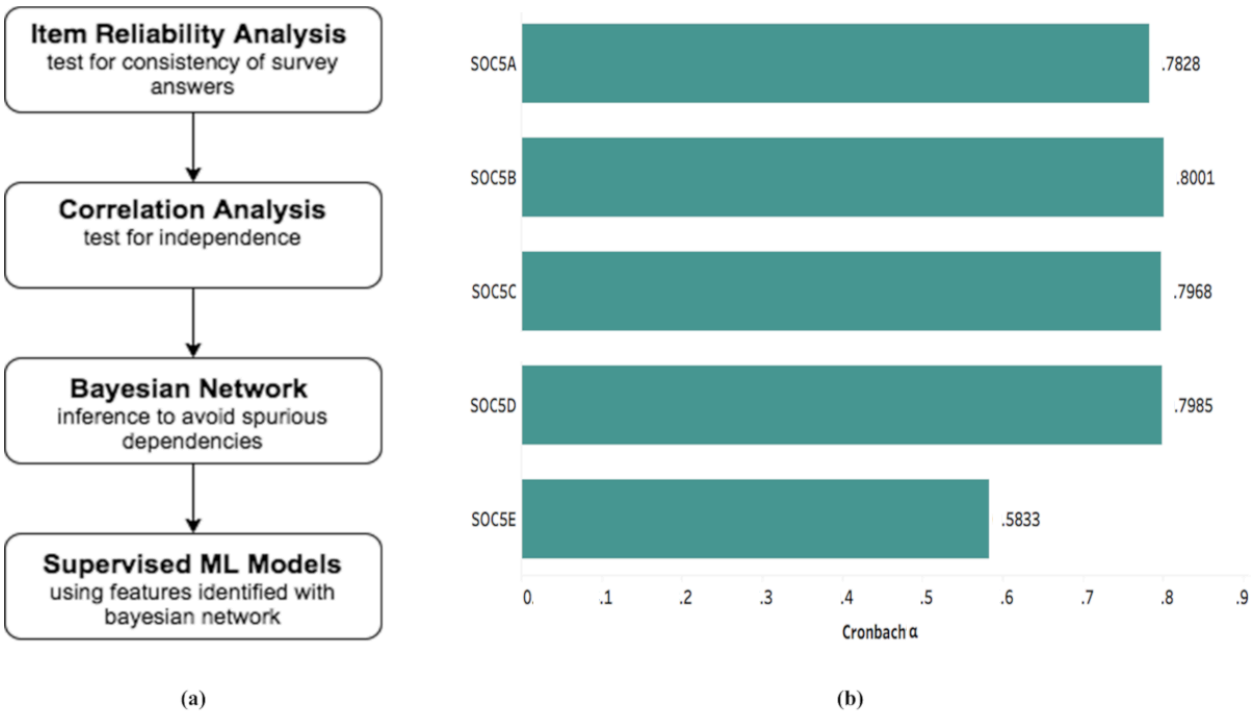

Figure-4.1 a. Methodology Flow. (b) "Item reliability analysis"

4.2.2.1 Item Reliability Analysis (IRA)

We built a mathematical graphical representation for the mental wellbeing factors using the following variables: "Soc5e (sweating, difficulty breathing, pounding heart, etc. in the last 7 days), Soc5b (felt depressed), Soc5c (felt lonely), Soc5d (felt hopeless about the future), and Soc5a (felt nervous, anxious, or on edge)". Consequently, we used an analysis named IRA to gauge the correctness for the response for mental wellbeing queries. The "Psych package in R (R Foundation for Statistical Computing)" was used to calculate "Cronbach alpha", a scale for evaluating the dependability of internal correctness[58].

4.2.2.2 Test of Independence Among the Mental Health Indicator and Other Indicators

Then, a paired “chi-square test of independence” was used to look for links within mental wellness measures and other factors. The 'P value' less than half was used as the threshold as their importance.

4.2.2.3 Data-Driven Bayesian Network Analysis

Since factors related to mental health could have intricate relationships including possible “confounders, mediation, and intercausal dependence”, we included "data-driven BN structure learning" into the association analysis. Through “bootstrapping and ensemble averaging of edge directions”, the learnt BN's structural stability was strengthened. The optimal PGM that estimated the data was selected by utilizing the “hill climbing optimizer” [59] and “Akaike information criterion score” [60]. “Bootstrapped learning and majority voting” were performed on 101 networks. The strength of learned connections was measured using exact inference utilizing the belief propagation approach [61]. All this was carried out by utilizing a “package wiseR” [62].

4.2.2.4 Markov Blanket

The concept of a "Markov blanket" refers to a statistical model in the field of machine learning and Bayesian networks. Essentially, for a given node (or variable) in a Bayesian network, its Markov blanket includes all the nodes directly connected to it, including its parents (nodes directing into it), its children (nodes it directs into), and its children's other parents. This set of nodes forms the "blanket". The key property of the Markov blanket is that it effectively isolates the node from the rest of the network. This means that once you know the states of the nodes in the Markov blanket, they 'shield' the node from the

rest of the network. In other words, the node is conditionally independent of all other nodes in the network, given its Markov blanket. This concept is particularly useful in probabilistic reasoning and learning in Bayesian networks, as it allows for localized computations of probabilities, reducing computational complexity.

In practical terms, understanding the Markov blanket of a node can help in making predictions or inferences about that node based on a limited, manageable set of other relevant nodes, rather than needing to consider the entire network.

4.2.2.5 Mental Health estimation by utilizing Supervised Machine Learning

"The Markov blanket" of mental health factors was then explored to understand the variables that might anticipate response indicators [63]. "The synthetic minority oversampling technique" was utilized to fix "the class imbalance" after dividing the dataset into two parts. First is train data (.8) and second is test data (.2) . Using the "Python Scikit-learn" module, several "supervised machine learning models—random forest (RF), support vector machine (SVM), logistic regression, and naive Bayes"—were learnt for estimating how people would react to mental health factors.

4.3 Results:

4.3.1 Item Reliability Analysis (IRA)

“Attribute soc5a (Felt nervous, anxious, or on edge), attribute soc5b (Felt depressed), attribute soc5c (Felt lonely), attribute soc5d (Felt hopeless about the future)” achieved a “Cronbach's alpha” score of 0.8 [Figure 2(b)], which confirms “internal consistency and suitability” for model.

4.3.2 Variations within mental-health variables with respect to age and gender.

A disparity in "attribute soc5a" was noted based on gender and age, as evidenced by a higher prevalence among females compared to males (as determined by a "two proportion z-test" with a significance level of $P < 0.001$) (Figure 3a). Additionally, individuals in the 18-29 age bracket exhibited a statistically significant difference ($P < 0.001$) (Figure 3b). The age cohort of 18 to 29 years in both sexes exhibited the highest susceptibility to experiencing mental stress for more than five days per each week. This suggests that COVID-19 may have had an overabundance of effect on the mental well-being of young individuals, potentially attributable to a range of factors.

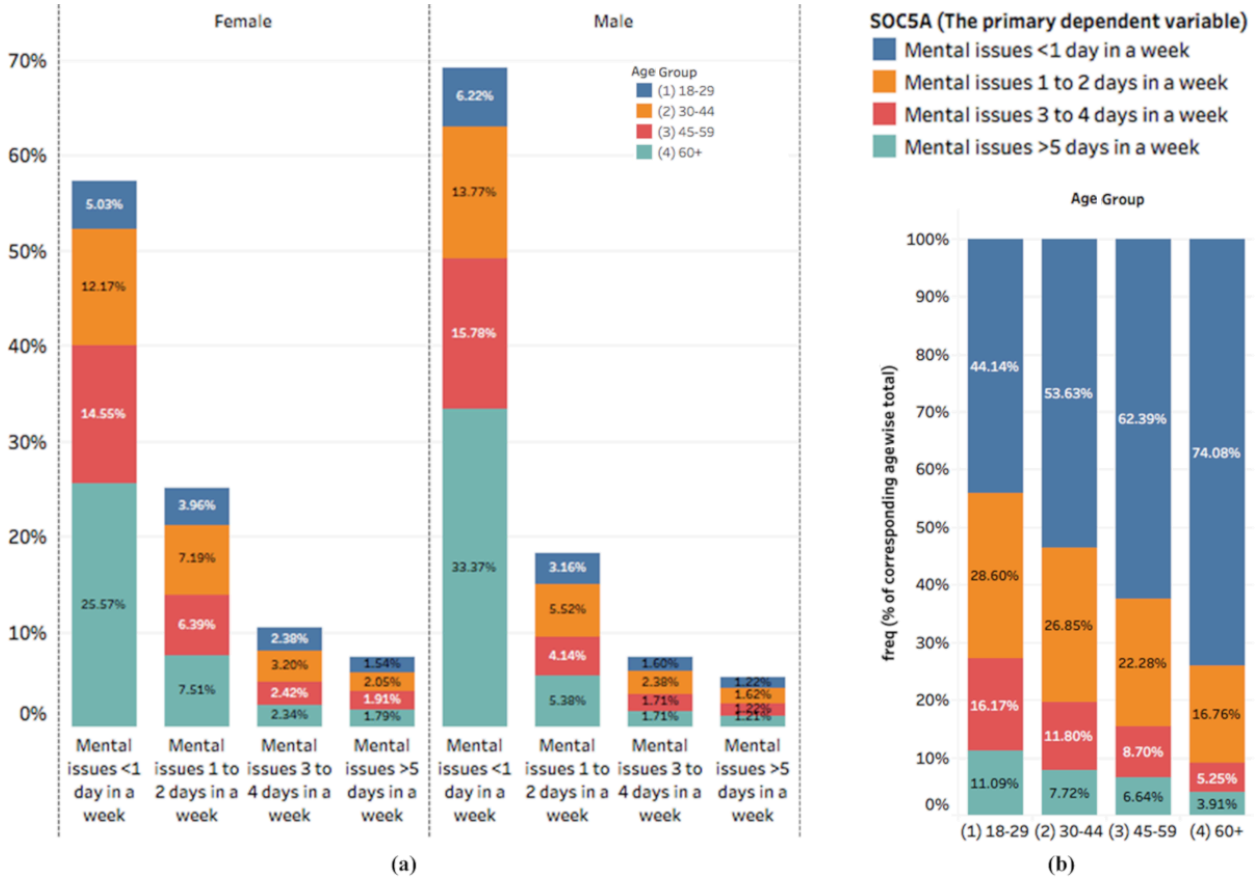

Figure 4.3. a. Gender based and b. Age based analysis

4.3.3 Connections of depression

These mental wellness factors exhibited numerous crucial connections, as indicated by a "Chi-square test" (Sup Fig-1). Nevertheless, this examination fails to consider plausible "confounding" or "explaining away" outcomes.

4.3.4 Data-Driven Bayesian Network Analysis

As a result, a "Bayesian network structure-learning" approach based on data was conducted, yielding noteworthy findings. The majority of self-initiated

structures indicated that the "attribute soc5a", which pertains to experiencing feelings of nervousness, anxiety, or edginess within the past week, was identified as the main factor for additional markers of mental health, as depicted in Figure 4(b) based on the acquired structure. The "attribute soc5a" was selected as the primary "dependent variable" for afterward "modeling analysis", serving as the driving factor within the structure itself.

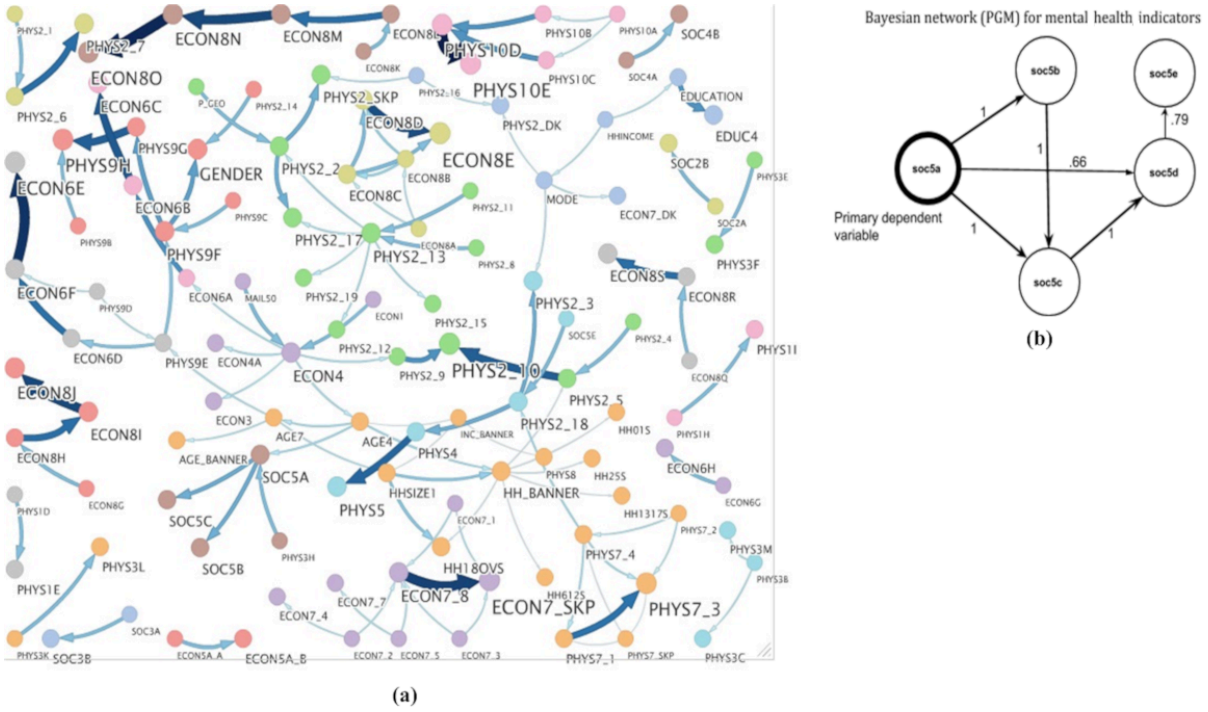

Figure 4.4. (a) **The consensus structure was derived from 101 bootstrapped samples, using hill climbing search combined with the Bayesian information criterion to identify the structures. Only connections with edge strength and direction strength exceeding 90% are displayed. The color of the edges indicates the proportion of networks in which each edge appeared across the 101 bootstrapped samples, reflecting the level of confidence; (b) the attribute soc5a was identified as the parent node for all other mental health variables, which led to its selection as the primary dependent variable. PGM: probabilistic graphical model.**

4.3.5 Effects due to social interaction and job stressor

The findings of our study, which employed "network inference" through the "Exact Inference algorithm", indicate a discernible influence of face-to-face socialising on the mitigation of anxiousness symptoms. The study discovered a statistically significant and positive correlation within the level of anxiousness influence and the frequency of social interaction with nearby residents. Specifically, a monotonous rise of more than 5% with a confidence interval of approximately 1% at both sides was observed for "attribute soc2a", while a monotonous rise of more than 6.5% with a confidence interval of approximately 1.5% at each sides was discovered for "attribute soc2b". The impact observed in the aforementioned study was comparatively less pronounced, approximately 1.5%, and was accompanied by a broad "confidence interval". This was observed in the context of communicating electronically with acquaintances and relatives through various channels such as texting, calling, emailing, or internet-based mediums, as indicated by the "attributes soc3a and soc3b". The aforementioned discovery highlights the significance of social interaction while adhering to recommended precautions, such as wearing masks and practicing “social distancing”, to preserve mental well-being amidst periods of isolation. It was discovered that the existence of children in the household is associated in a decrease in the likelihood of depressive disorders by more than 11%, with a confidence interval of approximately 2% on either side. Moreover, the network analysis conducted Exact Inference and indicated an elevation in the "conditional probability" of anxiousness (specifically, "attribute soc5a") due to work cancellations or postponements (as well as a rise of over 4% and a confidence interval of approximately 1.4%), educational institutions

cancellations or postponements (using a raise of 7% and a confidence interval of approximately 1.5%), working in their homes (using a raise of over 5% and a confidence interval of approximately 1.3), and studying in residence (using a raise of over 7% and a confidence interval of approximately 1.8). It is noteworthy that despite 83% of the participants opting for wearing masks, 77% refraining from visiting eateries, as well and 83% avoiding general population and busy areas, our hypotheses did not detect any significant alteration in anxiousness stages. The aforementioned inferences have been succinctly encapsulated in Fig-5.

Figure 4.5. "Inferences from the Bayesian network were derived by calculating the difference in inferred probabilities after conditioning the independent variables. A positive association indicates a factor that contributes to mental stress, while a negative association suggests a factor

that alleviates mental stress. The red circle represents the mean value, with the green and blue intervals indicating the confidence ranges. COPD stands for chronic obstructive pulmonary disease, SNAP refers to the Supplemental Nutrition Assistance Program, and TANF denotes the Temporary Assistance for Needy Families program."

4.3.6 Consequences of symptom and comorbidity

The present study additionally explored the association among psychological stresses and markers for COVID-19 symptom. According the recommendations of the World Health Organisation, it is advisable to seek medical attention from healthcare professionals if an individual experiences any of the signs related to COVID-19 (including but not limited to "phys1a to phys1q") within the preceding seven-day period. The findings of this research suggest that there was no notable influence of the response on mental wellness (specifically, 'attribute soc5a'). "The conditional probability" of this attribute stayed constant at 62.2% throughout every answer. Despite the established association between health problems ("attributes phys3a to phys3m") and heightened susceptibility to severe COVID-19 outcomes, our analysis revealed a counterintuitive relationship between cancers ("attribute phys3k") and hypertensions ("attribute phys3b") and anxiousness stages. Individuals afflicted with cancers exhibited a "conditional probability" of experiencing a minimum of one day of anxiousness every week that was roughly 8.3% larger, with a confidence interval that was roughly 2%. Hypertension, on the other hand, demonstrated an impact size of greater than 7% with a confidence interval of roughly 1.5%. Furthermore, the medical conditions of "Cystic Fibrosis" (physiological "attribute 3i") and Liver Diseases (physiological "attribute 3j") exhibit broad ranges of confidence and insignificant disparities in their averages. (Fig4.5).

4.3.7 Consequences of financial features

The receipt of financial support via Social Securities was found to enhance the likelihood of experiencing a maximum of continuous 24 hours of anxiousness a week by 10.4% (using a confidence interval of approximately 1.5%), relative to individuals that did not submit an application for or obtain this kind of support. The act of submitting an application for financial support resulted in a 4% enhancement, as illustrated in Figure 4.5. The outcomes were comparable for Supplementary Social Securities, or SSS, which had a confidence interval of approximately 4%, and medical coverage, which had a confidence interval of approximately 2%, with both contributing to a similar effect.

Furthermore, individuals who are classified as elderly (aged 60 years and above) have reported a greater sense of comfort and ease regarding health insurance as compared to their younger counterparts. The COVID-19 pandemic has had a significant impact on the economic well-being of everyone. This could potentially result in psychological strain.

4.3.8 ML Models for people who are prone to anxiousness

The study employed a "supervised machine learning" technique that relied on the "Markov blanket" for "soc5a attribute", which includes age ("age4 attribute"), physical indicators experienced in the past one week ("phys7_4 attribute"), staying within residence ("phys2_18 attribute"), and previous medical diagnosis for any kind of psychological disorder ("phys3h attribute").

Three instances of predictions were evaluated in the present study.

- The study compares individuals who experience mental health problems for a maximum of continuous 24 hours (classified as Label-1) to those who experience mental health problems for over a continuous 24 hours(classified label-0).

- The study compares individuals who experience mental health problems for a maximum of continuous 24 hours (classified as Label-1) to those who experience mental health problems for over a continuous 72 hours(classified label-0).

- The study compares individuals who experience mental health problems for a maximum of continuous 24 hours (classified as Label-1) to those who experience mental health problems for over a continuous 120 hours(classified label-0).

Based upon the conventional modeling evaluation metrics ("accuracy, sensitivity, specificity, AUROC"), which are outlined in the first tableau, "random-forest (RF)" algorithms outperformed "SVM, logistic regression, and naive-Bayes" algorithms. When we proceeded through a scenario with an elevated likelihood for depressive disorder (label three) to a scenario with a small likelihood of depressive disorder (label one) [table-1], one witnessed a decline within the algorithm's predictions ("Accuracy" 0.80 to 0.64; as well as the range of confidences given in the table-1). With each of the four ML methods we employed, this pattern could be seen.

Scenarios	RF	SVM	Naive Bayes	Logistic
Mental issues less than one day in a week (class 1) Vs Mental issues more than five days in a week (class 0)				
Accuracy (+- CI)	.80(+-.016)	.80(+-.016)	.77(+-.017)	.77(+-.017)
Sensitivity (+- CI)	.59(+-.063)	.56(+-.063)	.59(+-.063)	.59(+-.063)
Specificity (+- CI)	.82(+-.016)	.82(+-.016)	.79(+-.017)	.78(+-.017)
AU Roc (+- CI)	.71(+-.026)	.69(+-.026)	.69(+-.025)	.68(+-.025)
Mental issues less than one day in a week (class 1) Vs Mental issues more than three days in a week (class 0)				
Accuracy (+- CI)	.72(+-.018)	.72(+-.018)	.74(+-.017)	.73(+-.018)
Sensitivity (+- CI)	.6(+-.041)	.6(+-.041)	.56(+-.041)	.57(+-.041)
Specificity (+- CI)	.75(+-.018)	.75(+-.018)	.78(+-.017)	.76(+-.018)
AU Roc (+- CI)	.68(+-.022)	.67(+-.022)	.67(+-.022)	.67(+-.022)

Mental issues less than one day in a week (class 1) Vs Mental issues more than one day in a week (class 0)				
Accuracy (+- CI)	.66(+-.019)	.66(+-.019)	.65 (+-.019)	.62 (+-.019)
Sensitivity (+- CI)	.48(+-.027)	.49(+-.027)	.45(+-.026)	.61(+-.026)
Specificity (+- CI)	.77(+-.018)	.76(+-.018)	.77(+-.018)	.64(+-.020)
AU Roc (+- CI)	.62(+-.019)	.62(+-.019)	.61(+-.020)	.62(+-.018)

Table 4.1: conventional modeling evaluation metrics

4.4 Discussion:

The issue of psychological wellness is a significant matter of public health. The prevalence of mental health conditions and suicidal behaviours has exhibited a noteworthy escalation throughout the past decade across all demographic cohorts and both sexes [64,65]. The swift proliferation of coronavirus contagion prompted nations worldwide to implement the closure of communal spaces, educational institutions, dining establishments, and industrial facilities. The current state of affairs has resulted in a shift towards social isolation, reliance on technology for communications, and remote jobs and schooling, which have become commonplace. Furthermore, there have been a significant loss of employment opportunities. Around the world, the worldwide impact of this occurrence has resulted in heightened levels of anxiousness, and depressive disorders. No empirical research has been identified that employs models for the

dual purpose of predicting and elucidating the nuanced impacts of life circumstances on mental well-being. Our study utilized a "probabilistic graphical modeling" technique that is "explainable", incorporating "bootstraps and exact inference methods". This method enabled us to successfully capture a multitude of consequences in a resilient way. The findings of our investigation indicate that individuals with a pre-existing identifying of mental disorders have a heightened risk for developing mental health issues throughout the COVID-19 pandemic. Since a result, it is recommended that national-level regulations be implemented to systematically monitor their mental well-being and provide appropriate treatment. Our findings emphasise the financial basis of an integrated approach to addressing mental health. Our model indicates that the provision of income assistance through "Social Security or Supplemental Social Security" have a measurable impact on reducing anxiety disorders. This represents the initial evidence, to the best of our knowledge, supporting the effectiveness of these initiatives. The impact of the magnitude of these measures can be assessed through modelling studies carried out in different regions of the world, where there are varying levels of support structures during this period.

The results obtained through our study conducted in the US have the potential to inspire additional that cultural and social investigations in other regions that share comparable or distinct societal frameworks. The potential impact of face-to-face communication versus online interaction could differ in regions where communal living and extended families remain prevalent, such as India. The efficacy of digital connection in comparison to interpersonal communication with neighbours has been found to be inferior in the United States. This underscores the fundamental dissimilarities between these two

factors in their impact on mental health, and highlights the need for further organised investigation. It is postulated that the aforementioned variations may be attributed to the processes of evolutions that have influenced human societies to coexist and interact in intimate physical proximity. The manifestation of depressive symptoms has been demonstrated in primates that were previously isolated [19,20]. The correlation between "neuropeptide hormones" and parenthood may offer a partial rationale for our findings that having a number of children is linked to a decrease in anxiety disorders levels. This is in line with previous research [21]. The COVID-19 pandemic has presented a distinctive opportunity to conduct an organic study on the communal psychological reaction of people after an epidemic of public health.

Additional research might be required to examine the life-cycle of such a reaction to be the global community navigates via different stages of the global epidemic until it's ultimate solution. Nonetheless, our research suggests the effects on psychological well-being can be discerned within a short period of time, particularly among adolescents. Additional investigation is required, preferably in a longitudinal context, wherein the same participants can be re-surveyed to comprehend the structure of the communal psychological reaction.

The findings of study underscore the inadequacy of contemporary advances in technology in electronic interactions as a substitute for organic interactions between people. Therefore, it is crucial to develop more effective and compassionate scientific devices that can form a community that discourages seclusion and detachment, while still upholding determinants for social distance and disease prevention to restrict transmission. The customization and

situational adaptation of aforementioned evaluates are going to have significant value, given the fact that individuals having prior psychological ailments might encounter a disproportional impact.

Ultimately, the findings suggest that individuals with the greatest susceptibility to mental health disruptions may be identifiable. The system exhibited highest efficiency in relation to individuals that were deemed especially susceptible, as indicated by experiencing psychological strain for over 5 day(s), in comparison with those whose were considered less susceptible, characterized by experiencing no anxiety or fewer compared to a day of stress per week. The method may aid in the categorization of susceptible demographics, such as the front lines medical professionals, whose are experiencing an unjust degree of strain amidst the current circumstances.

"Transparency and explainability" are crucial elements in both "clinical and public health" models, particularly when dealing with intricate relationships. The intricate relationships between psychological factors as well as probable "confounding variables, mediation, and inter-causal dependency" are anticipated. As a result, we expanded our associative analysis by utilizing "data-driven structure learning of Bayesian networks". This approach was favored over "black-box machine learning "and standard statistical modeling due to multiple reasons. The utilization of "structure learning" enables the identification and representation of "confounding factors" in an open manner, while ML models that operate as black-boxes, such as "random forests and gradient boosted machines", are not optimally designed for "transparent" justification. The complexity of modeling relationships among thousands of different variables using standard statistical techniques presents a significant

challenge for human beings. The process of "structure learning" enables the identification and analysis of various types of effects, such as "mediation, confounding, and inter-causal effects", by breaking down and examining relationships. The issue of erroneous learning is tackled through the implementation of an "ensemble" method, wherein a multitude of "Bayesian networks" (101 in this particular instance) are utilized and the "ensemble's voted structure" is selected. This research showcases the possibility of an "AI" approach in multifaceted psychological wellness circumstances, which has been previously examined for problems related to public health [66,67]

The process of bringing a COVID mental health prediction solution to the market is multifaceted and requires careful planning across various domains. Although providing precise timelines and resource estimates is challenging without specific solution details, key elements of this journey typically include:

Rigorous Clinical Trials and Studies: Essential to validate the solution's efficacy and safety, these trials involve methodical testing and data collection, often spanning several phases and potentially lasting years.

Strategic Partnerships with Healthcare Entities: Collaborations with hospitals, clinics, and mental health experts are crucial. These partnerships not only aid in the development and refinement of the solution but also play a vital role in its subsequent adoption and integration into healthcare systems.

Securing Adequate Financial Investment: The commercialization process requires substantial funding. This encompasses development costs, clinical trials, marketing, and operational expenses. Securing such funding might involve reaching out to investors, applying for grants, or other financial avenues.

Building a Skilled Team: The success of the solution hinges on the expertise and dedication of the team involved. This includes professionals in healthcare, technology, regulatory compliance, and business strategy.

This roadmap outlines the primary pillars necessary for the successful commercialization of a mental health prediction solution in the context of COVID-19. It underscores the need for a balanced approach that combines scientific rigor, strategic collaboration, and robust resource management.

The present investigation exhibits certain constraints. It is widely recognised that setting up "causal inference" in "cross-sectional data" is a challenging task, and the potential for "confounding" must be taken into consideration. The rationale behind our selection of the "structure learning approach" was precisely due to its ability to openly identify and elucidate certain "confounding factors". The present research employed an array of 101 "Bayesian Networks (BNs)" to tackle the issue of erroneous learnings. The "ensemble approach" was employed, wherein a model was chosen based on the majority decide of the "bootstrapped structures". This method was deemed robust, given the adequately significant amount of "bootstrapped structures" used. Our methodology is optimally designed as a "probabilistic reasoning" framework that elucidates the factors that determine psychological wellness. It also has the capability to generate estimates, which is a valuable result in the context of the psychological mortality caused by the COVID-19 pandemic. The reasons behind the potential decrease in anxiety levels among individuals with already present cancers or hypertensions remain unclear. This phenomenon could potentially be attributed to a decrease in stress levels associated with the work environment or an

increase in familial interactions within the home setting. Nevertheless, the present dataset is inadequate for providing a more detailed level of explication. Additionally, the length of time and perseverance of these impacts could not be commented upon. Presently, our findings are constrained to a singular geographic location, specifically the United States. The survey's substantial number of respondents and inclusion of multiple ethnic groups render the model representative of various ethnicities and influences throughout the United States, thus increasing the likelihood of its validity within the country. Ultimately, our research represents a significant advancement in the utilisation of explainable artificial intelligence for the purpose of forecasting mental health outcomes on a population-wide scale through the analysis of survey data, thereby increasing its potential for widespread implementation. The datasets obtained through surveys are known to contain a significant amount of noise. Our methodology has successfully achieved an equilibrium between the discovery of information and predictability, resulting in an accuracy of 80% rate. This has established a baseline for a new and unique scenario. The algorithms have the potential to serve as a screening tool for detecting individuals who require assistance, and the incorporation of supplementary measurements and characteristics may enhance the precision of prognostications. Consequently, the development of predictive models aimed at screening and evaluating the mental health consequences of COVID-19 represents a critical measure in the proactive treatment and avoidance of mental health disorders, as communities persist in their efforts to combat the global epidemic.

Chapter 5

Integrative Probabilistic Modeling for Quantitatively Analyzing Mental Health among Older People in India

Analysis work utilizing Bayesian networks together with stratification analysis for SES and multimorbidity analysis

5.1 Introduction

A key component for healthy ageing is psychological well-being. The elderly are more prone to psychological problems. Understanding the frequency and contributing factors of elderly mental health difficulties is essential given the rapid ageing of the global population. Reports indicate that in this demographic, mental health issues can significantly affect a person's daily routine lifestyle, level of functioning, and overall state of health [68,69]. With projected growth from 703 million individuals who are 65 or older in 2019 to 1.5 billion in 2050, the worldwide population of elderly people is anticipated to grow quickly in the next decades [70]. With an expected 340 million individuals aged 65 or over by 2050, India is predicted to hold the second-highest global population of elderly people[71] [72]. It is projected that as the aged population grows, older people's burden of mental health issues will grow as well, creating a significant public health challenge in India. As older people population is more prone to chronic illnesses, cognitive disabilities, as well as mental health issues, this rapid ageing has consequences for global health and healthcare systems [73,74].

According to research, older people around the world frequently struggle with mental health difficulties. According to research [75], older persons were more likely to experience depression in nations with middle and low incomes (7.2%) than in high-income ones (11.5%). These results emphasise the necessity for research into elderly mental health difficulties, particularly in countries with low or middle incomes where the number of elderly people has been growing rapidly. Numerous physical as well as mental challenges that older people come

across must receive more attention from society. With the exception of headache disorders, over 20% of elderly persons who are 60 years of age or older suffer from mental and neurological conditions [76,77]. According to the latest data, mental, neurological, and substance abuse issues formed 10.5% of the worldwide illness challenge, as assessed by DALYs, while mental, neurological, and substance abuse disorders accounted for 6.6% of all "DALYs (disability-adjusted life years)" among individuals ages 60 and above [78] [79–81]. Between 1990 and 2013, the YLDs for mental health and substance abuse issues expanded by 45% [81]. The two mental and neurological conditions that affect elderly people the most are dementia as well as depression. Approximately five percent and seven percent of the elderly worldwide experience depression and dementia, respectively. Huge social and economic challenges make these issues worse in developing nations like India [82,83]. Physicians and senior citizens may not constantly recognise mental health issues, and the stigma attached to them discourages people from seeking assistance.

In addition, decreases in socioeconomic status associated with retirement and bereavement seem to be more prevalent in older persons [84]. Any of these stressors have the potential to cause elderly people to feel mentally alone, lonely, or upset, demanding long-term care. Numerous investigations have verified the relationship between both mental and physical well-being. [85,86]. Previous research has also demonstrated that socioeconomic status (SES) and mental health are significantly related. For instance, elders who live in nations with higher income disparities have poorer psychological well-being than elders who live in nations with lower income disparities. SES as well as mental health are related, as evidenced by the harmful effects of displacement on adults who

reside in metropolitan regions. In Japan, the frequency of poor mental health among the elderly decreased as socioeconomic level (SES) rose, supporting the idea that SES and mental health are positively correlated. The socioeconomic status (SES) of older individuals has an impact on their psychological resources, especially their perception of autonomy, self-efficacy, and positive self-beliefs [87]. For instance, self-efficacy [88] may encourage elderly people to lead mentally good lives since they believe they are capable of handling the demands and challenges of their immediate surroundings. In a similar vein, older people who are enthusiastic about the future frequently have healthier lives. Additionally, it is thought that older individuals' sense of independence is good for their health since it mitigates the negative effects of stress variables. There hasn't been much research on how much a neighborhood's socioeconomic standing affects individuals' sense of self-efficacy, which subsequently in turn affects their ability to maintain good health. The great majority of research efforts in this field have only looked at how self-efficacy [89] affects the relationship between SES factors and health at the individual's level. Older adults with mental diseases can be shielded from the negative effects of prejudice by societal structures including social protection, interpersonal trust, and solidarity. For instance, if individuals had access to instrumental (e.g., travel assistance), informational (e.g., health information), or affective (e.g., comfort in the face of a health problem) social support, they might be able to obtain resources that they otherwise could not. These examples of how social support keeps people safe also demonstrate the importance of observing one's surroundings. Social environments, such as neighborhood social cohesion, are particularly essential sources of social support.

Having undergone an abundance of literature exploring the link between SES and mental health, models that account for the impact of racial and ethnic diversity have been astonishingly overlooked. In this research, we assessed information from a large-scale survey of Indian seniors on a range of demographic and sociocultural factors. To begin, we used bayesian network inferences to put a number on the correlation among SES and psychological well-being. We also used the identical strategy to the evaluation of mental health comorbidity with other chronic illnesses. Bayesian networks are a form of graphical model, and to the best of our knowledge, this is the first time they have been used to examine the mental health of Indians in their later years. Our findings may help shape targeted treatments and policies that encourage healthy ageing and improve mental health outcomes in this group by identifying the characteristics associated with these outcomes. Our results may be generalizable to other poor and middle-income countries where similar problems are prevalent as a result of the global rise in the senior population.

Additional research included an "age-based stratification analysis". The senior population, who are more likely to suffer from co-occurring mental problems, is a prime example of why "age-based stratification analysis" is so important to society. Policymakers, healthcare professionals, and social service organizations may better allocate resources and develop treatments to suit the special requirements of this group if they have a better knowledge of the patterns and incidence of mental diseases among older persons.

The psychological well-being of seniors is a complex topic, and "age-based stratification analysis" may assist pinpoint potential risk factors and protective

factors. This knowledge may be used to guide the design of early treatments and preventative measures aimed at improving the mental health of the elderly. In addition, health and social care policies that take into account the specific requirements of the ageing population may be made more effective and fair if we have a better grasp of the effects of mental health multimorbidity.

In conclusion, “age-based stratification analysis” is essential to society because it elucidates the distinct mental health needs of the elderly, thereby guiding the development of specific strategies to improve their health and quality of life.

The remainder of the paper will be structured as follows: first, we will provide an overview of the LASI dataset and the methods we employed for this analysis. Then, we present our findings and discuss their significance. We conclude with a discussion of the limitations of our work and recommendations for future research.

5.2 Method

We used data of a large survey conducted in India for people aged above 45 years and performed analysis for finding associations between disorders and diseases in elder population. The details are mentioned below.

5.2.1 Dataset

This information was collected through the "Longitudinal Ageing Study in India (LASI) Wave-1" study. [90] [91], consisting of 31,464 persons aged 60 and more, and 6,749 individuals aged 75 and up in India, within a total sample size of

72,250 people aged 45 and up. The information comes from 35 different states and union territories (other than Sikkim). From April 2017 through December 2018, the aforementioned poll was carried out nationally. Adults aged 45 and above with a partner (without regard to the age of the partner) are the target population for the LASI, a "multistage stratified area probability cluster sampling design" for locating the final observations. Specifics from the source data are available elsewhere [92]. Table 1 displays the socioeconomic and demographic characteristics of the survey respondents.

5.2.2 Analysis

5.2.2.1 Assessment of mental health

Two distinct mental health estimations were employed in the study. One, the attribute 'HT009' is utilized for multimorbidity analysis of chronic mental health conditions such as depression, Alzheimer's or dementia, unipolar or bipolar disorders, convulsions, Parkinson's, etc. that have ever been diagnosed by a health professional. Second, the presence of psychological distress was assessed using the Centre for Epidemiologic Studies Depression (CES-D) scale15 so that links between socioeconomic status, discrimination, and family network structure and mental health in the elderly could be explored. The Centre for Epidemiologic Studies on Depression (CESD) is a "short self-reported scale designed" to screen for depressed symptoms in the general population. [93].

5.2.2.2 Multimorbidity Network analysis

Medical professionals use the term "multimorbidity" to describe the presence of two or more coexisting chronic conditions in a single patient. It is well known that as one ages, their likelihood of contracting several ailments also rises. Having many diseases increases a person's risk of death and accelerates the decline of their physical and functional skills. Bayesian network analysis [94] [95] [67] was used in this research to visually examine the comorbidity pattern among the elderly. We included conditions like cancer or malignant tumor, thyroid disorder, gastrointestinal problem, and cardiovascular-disease CVD (including one or more from hypertension HT or high blood pressure, stroke, and chronic heart diseases). We also included diabetes mellitus, high cholesterol, and chronic lung diseases like asthma, COPD, and bronchitis. And then, I discovered how to do a visual assessment of multimorbidity by learning the joint probabilistic link between mental illnesses and Bayesian networks.

5.2.2.3 Association of Socioeconomic status Factors with Mental-Health

Important correlations between SES and psychological health have been uncovered via research. It has been observed that this connection is especially robust among the elderly. There is a correlation between socioeconomic status and mental health, and researchers think that a number of variables contribute to this link. Especially for underprivileged communities, it is crucial to include socioeconomic factors when designing interventions to promote mental health and reduce the burden of mental illness. Data included a wide range of socioeconomic and demographic characteristics: information on: age, sex, marital status, education, languages spoken at work, employment history,

retirement plans, and pension benefits; General health ("Excellent/Very good/good" or "fair/poor"), chronic ailments (cardiovascular, chronic heart, and vascular diseases, stroke, cancer, urogenital, and gynecological disorders, etc.), and mental health; Physical and mental impairment, mobility, ADLs, IADLs, and aids; family and social network; living arrangements; social and family connectedness; social support; autonomy in decision-making, social participation, and life satisfaction; ageism and discrimination; life satisfaction. Learning how different socioeconomic circumstances affect people with varying degrees of mental health (never, sometimes, frequently, most of the time) was the major result of interest. We employed data-driven Bayesian network conclusions to examine the association between socioeconomic status and psychological well-being.

5.2.2.4 Age based stratification analysis for SES and multimorbidity

We conducted a "age-based stratification analysis" after learning about multimorbidity and the correlation between socioeconomic position and psychological well-being. Our findings indicate that both the number of co-morbidities associated with mental health and the impact of socioeconomic status (SES) on mental health vary with age. When it comes to mental health issues and co-morbidities, certain age groups are more vulnerable than others to the impact of the same socioeconomic variables. For instance, the likelihood of being exploited as a result of a lack of social support is lowest for seniors between the ages of 45 and 64. As a result, studies examining the link between socioeconomic position, mental health, and the occurrence of various disorders

must account for age-based stratification. When these connections are made clear, age-specific interventions and policies can be created to mitigate the effects of mental health issues and multiple diseases.

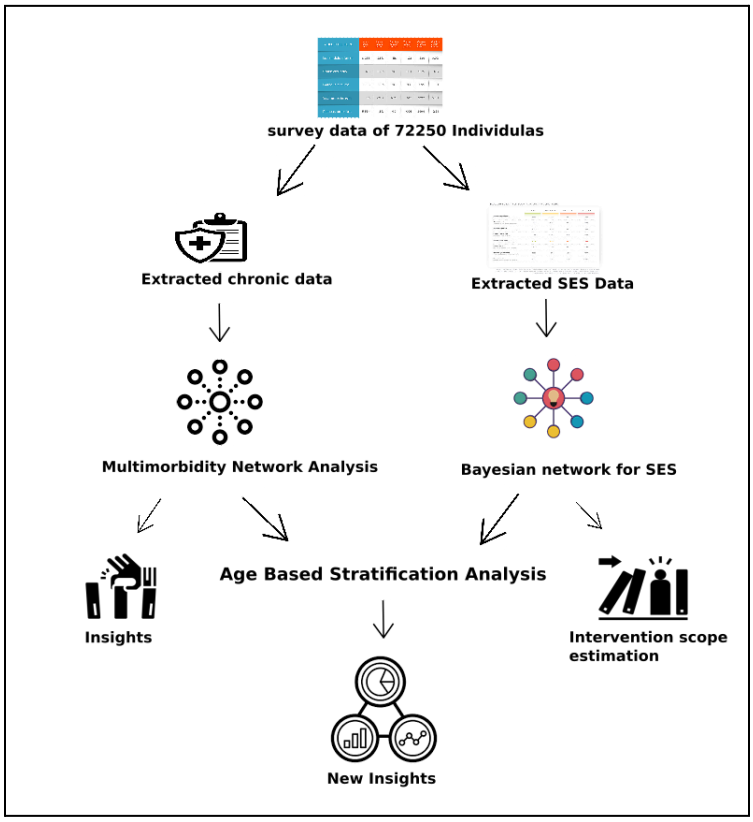

Fig: Schematic Diagram

5.3 Results

In order to have reliable insight, we started with exploratory analysis followed by estimation of direct association between different ailments (comorbidity) among elders. Further we performed an analysis to estimate the effect of social and economic factors on the mental health of the elderly population in India.

5.3.1 Exploratory Analysis

We analysed the responses of 72250 seniors (42% males, 57% females) in great detail. Researchers discovered very minimal change in the frequency of mental health across variables such as gender, socioeconomic position (MPSE_Quintile), and geography (rural vs. urban).

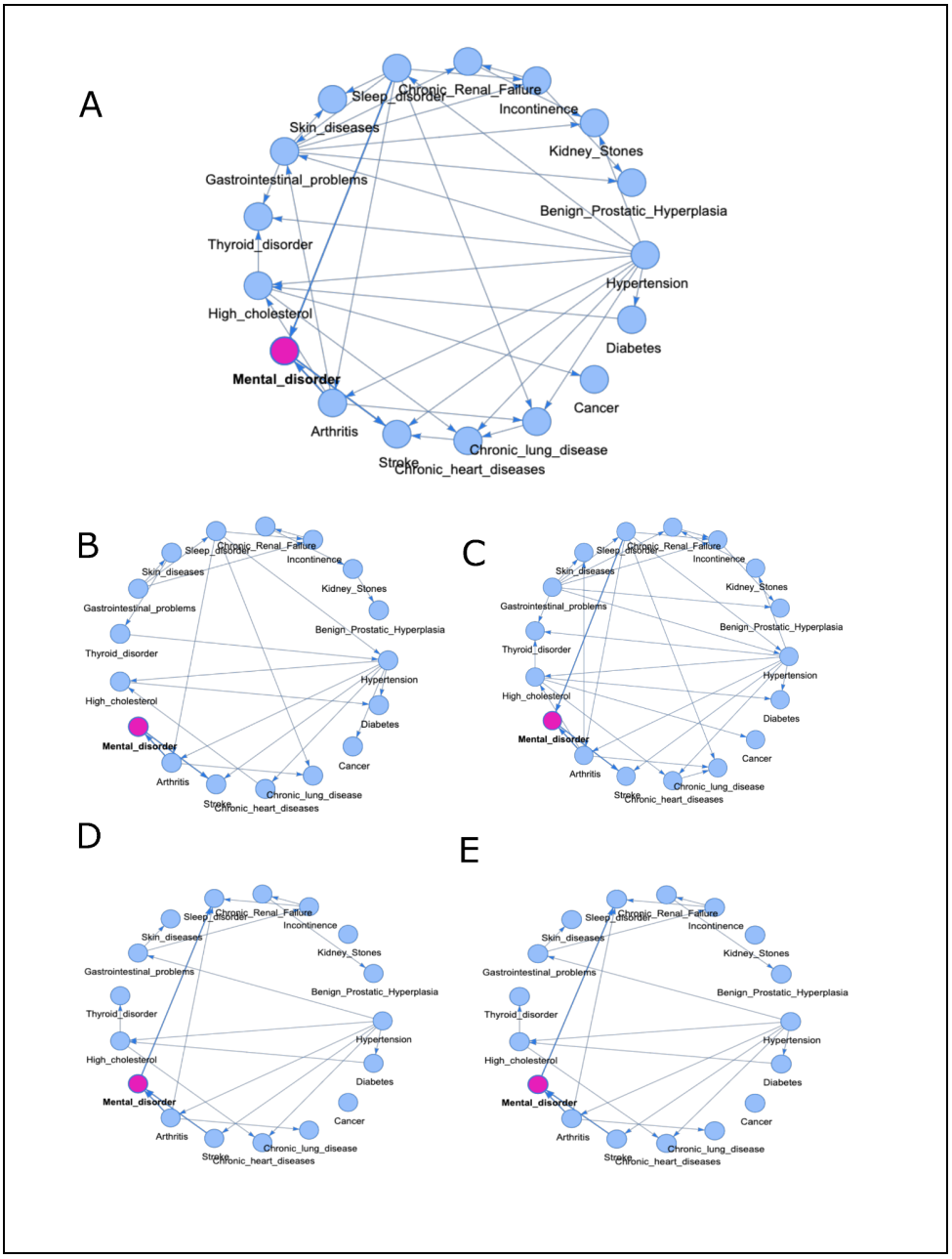

Figure 2: "Multimorbidity Analysis: (A) A Bayesian network of all 17 diseases together, consisting of the complete dataset. The goal was to find out which chronic diseases are also linked to mental health problems. (B) Bayesian network for individuals having ages less than or equal to 45. It is evident that mental health is not associated with sleeping disorders here. (C) Bayesian network for people between the ages of 45 and 64. Here, one additional linkage associated with mental health is found with sleep

disorders. (D) Bayesian network for people between the ages of 65 and 74. Here, one additional linkage associated with mental health is found with sleep disorders. (E) A Bayesian network for people over the age of 75. Here also one additional linkage associated with mental health is found with sleep disorders."

5.3.2 Multimorbidity Network analysis: Chronic Diseases and Mental Health

The Bayesian network inference demonstrated a direct relationship between mental health difficulties and rheumatoid arthritis, sleeplessness, and cardiovascular disease. For arthritis, the exact inference produces a conditional probability difference of 20% (95% confidence interval [CI] 5%), for sleep disorders, it yields 18% (95% CI 5%), for stroke, it yields >6% (95% CI 3%), and for hypertension, it yields >5.5% (95% CI 5%).

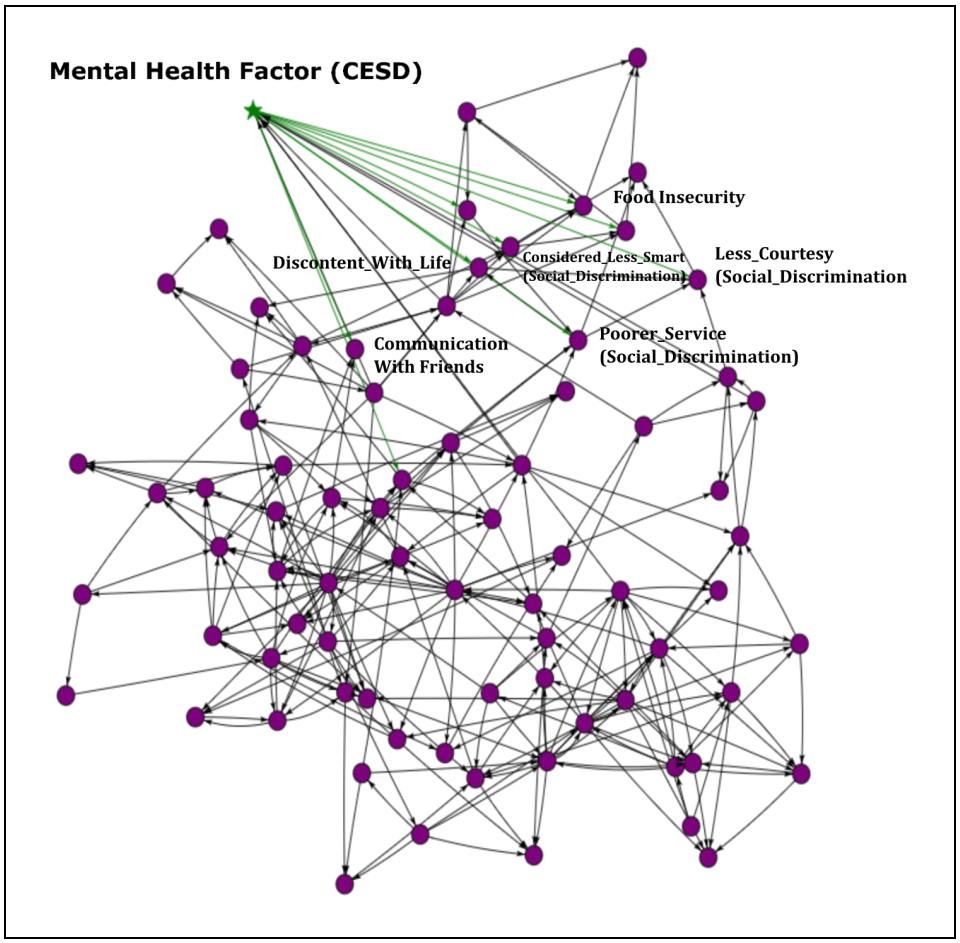

Figure 3: "Representation of some of the association of socioeconomic factors with mental health. We learned a Bayesian Network from the LASI data, where the Starred green node represents mental health, in terms of CESD"

5.3.3 Age based stratification analysis: Multimorbidity

The concept of age-based stratification pertains to the uneven allocation of multimorbidity among distinct age cohorts. The stratification of individuals based on age in the realm of mental health has significant ramifications for the planning and provision of healthcare services. Our findings indicate that there is

a notable disparity in the occurrence of mental health multimorbidity between older and younger adults. According to the Bayesian network, there exists an association between mental health and arthritis as well as stroke among individuals aged 45 years or younger. Conversely, the Bayesian network pertaining to individuals aged 45-64, 65-75, and over 75 years indicates that their mental well-being is linked not solely to arthritis and stroke, but also to sleep disorders [as depicted in Figure 2]. Consequently, it can be inferred that the process of ageing plays a role in the heightened correlation between sleep disorders and mental health.

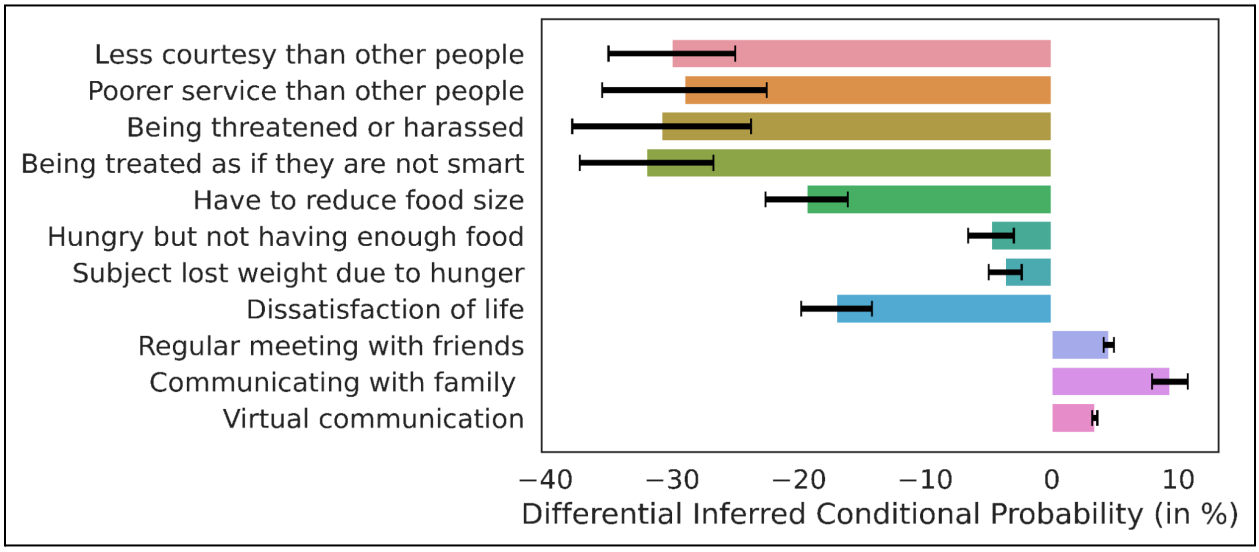

Figure 4: Socio-economic factors analysis with Mental illness

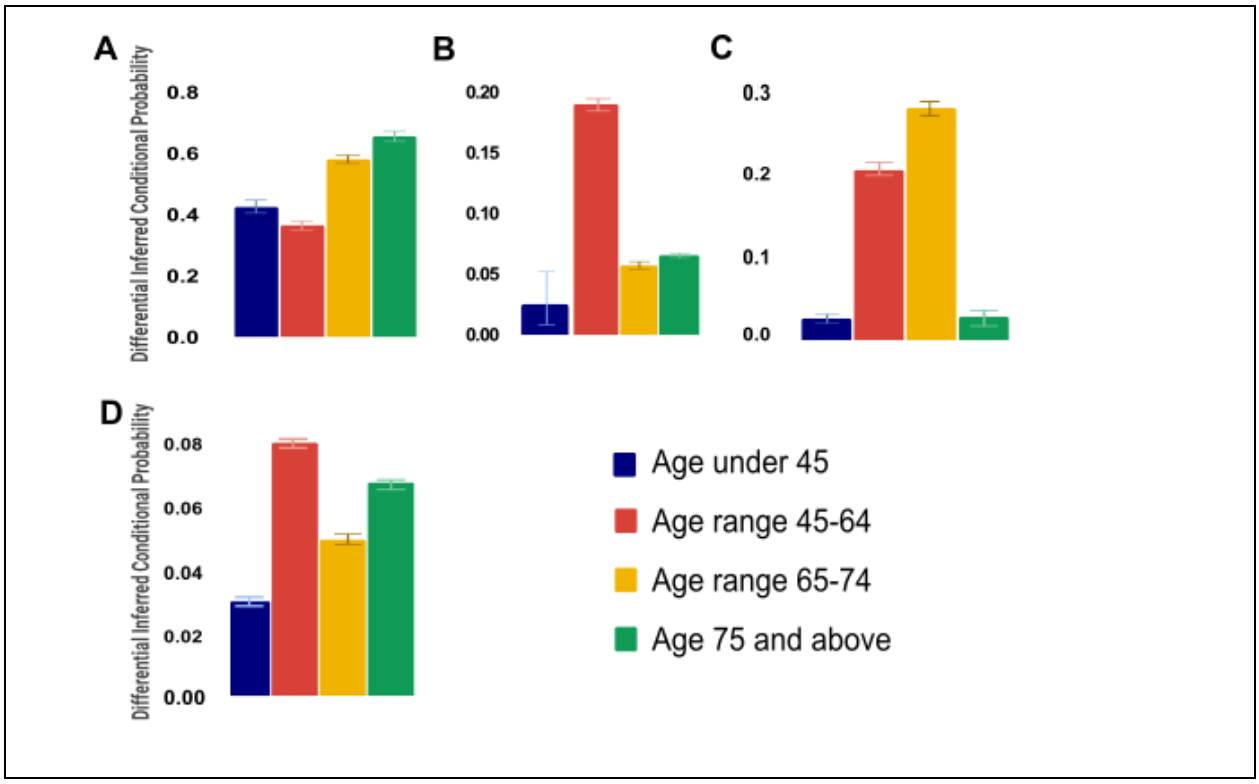

Figure 5: Age based stratification analysis for (A) Social Factor - Social Discrimination (B) Economical Factor - Food Insecurity (C) Psychological Factor - Discontentment about Life (D) Family and friends connectedness Factor - Communication

5.3.4 Data-Driven Bayesian Network Analysis

Our analysis to recognise social factors and economical factors affecting mental health of elderly, proved to be insightful as explained below.

5.3.1.1 Social Factor (Social Discrimination)

Ageism is a prevalent and concerning issue in contemporary society. The impact is observed in both developed and developing nations. Older individuals frequently report experiencing various forms of social discrimination in their

daily lives. Differential interference of conditional probability can be observed in the associations related to the mental health of senior citizens. The elderly population is experiencing social discrimination in diverse ways, including but not limited to receiving less courteous or respectful treatment compared to other individuals (with a confidence interval of approximately 5% and a mean of 30%), receiving poorer service at public establishments such as restaurants or stores (with a mean of 29% and a confidence interval of approximately 6.5%), being subjected to threats or harassment (with a mean of 30.8% and a confidence interval of approximately 7.06%), and being perceived as intellectually inferior (with a mean of 34% and a confidence interval of approximately 5.27%).

5.3.1.2 Economical Factor (Food Insecurity)

The results of our analysis indicate a discernible correlation between food security and an individual's mental well-being. The issue of food security has been found to have a significant correlation with depression, as evidenced by a reduction in meal size among subjects due to insufficient food (with a confidence interval of 95% and a percentage of over 19%), as well as instances of hunger without food consumption due to a lack of available resources (with a confidence interval of 95% and a percentage of 4.8%). Nonetheless, research has indicated that insufficient food intake and hunger resulted in a reduction in body weight, which subsequently had an impact on the participant's psychological well-being (with a 95% confidence interval of 1.3%, the effect size was 3.68%).

5.3.1.3 Psychological Factor (Discontentment of Life)

The psychological well-being of older people is significantly influenced by a prevalent feeling of dissatisfaction with one's life. Older individuals who have experienced general dissatisfaction with life have a higher conditional probability of depression, with a 95% confidence interval of approximately 2.8% lower and 17% higher.

5.3.1.4 Family and friends connectedness Factor (Communication)

The positive impact of social connectedness with family and friends on the mental health of elderly individuals has been identified as a significant factor. The study found that older individuals who engage in regular socialisation with friends have a differential inferred conditional probability of 4.5% [95% CI: 0.4%]. In contrast, those who talk about and discuss personal matters with their partner/spouse have a higher differential inferred conditional probability of 9.34% [95% CI: 1.41%]. It is noteworthy that virtual communication through means such as phone and email has a comparatively lower impact on mental health, while also exhibiting a positive effect on it, with an increase of 3.4% and a 95% confidence interval of 0.2%.

5.3.1.5 Age based stratification analysis: association of SES with mental health

An "age-based stratification analysis" was conducted to investigate the correlation between various factors and mental health, as depicted in Figure 3.

Our research indicates that social discrimination has a negligible impact on the mental health of individuals aged 45-64, who are considered the most industrious segment of the elderly population. It is noteworthy that the likelihood of experiencing depression as a result of social discrimination increases with advancing age. Likewise, our findings indicate that the age group ranging from 46 to 60 years old, which represents the most industrious segment of the elderly population, experiences the greatest impact on mental health as a result of food insecurity. As per a particular interpretation of these results, the effect of life dissatisfaction on the psychological well-being of elderly individuals is negligible for those below the age of 45, but gradually intensifies as they advance towards the oldest age group. Individuals within the age group of 75 years and above exhibit limited influence on their mental well-being owing to a feeling of discontentment with their life circumstances. The present study indicates that the aforementioned factor of connectedness exhibits the greatest enhancement in mental health outcomes for individuals aged 65 years and above, encompassing the youngest-old, middle-old, and older-old age groups.

5.4 Discussion

Our study examines the probabilistic graphical models which generate a list of valuable interventions and assistance programmes for reaching the senior citizens who are most at risk for mental health problems. The study's objective was to examine the prevalence of mental health problems among older persons as well as their relationships to various socioeconomic as well as psychological variables. According to our analysis, older persons' psychological characteristics, social discrimination, and poor food security are all strongly correlated with

poor mental health. Additionally, the "age-based stratification analysis" showed that some particular age groups are more vulnerable than others and the likelihood of multimorbidity of mental health issues rises with age. The policies governing geriatric well-being both in India and overseas may be affected by these findings.

A major global health issue is the incidence of geriatric mental health problems. In this study, we found a connection between older adults' food insecurity and mental health problems. Many elderly individuals lack access to adequate food to consume in both developing and industrialised nations. The Indian government has undertaken numerous initiatives throughout the years to lower food insecurity and enhance its residents' access to food. Integrated Child Development Services Scheme (ICDS), the Public Distribution System (PDS), the Mid-day Meal Programmes (MDM), the Special Nutrition Programmes (SNP), the Wheat Based Nutrition Programmes (WNP), the Applied Nutrition Programmes (ANP), and the Antyodaya Anna Yojana (AAY) are a few noteworthy initiatives. These programmes have improved the access of the most disadvantaged members of society to food and reduced food insecurity. However, issues like leaks in the PDS system and poor scheme implementation continue to be an issue. Additionally, a few food insecurity programmes are accessible to senior citizens in India. The National Old Age Pension Scheme (NOAPS) and other pension plans do not cover senior individuals, hence the Ministry of Rural Development launched the Annapurna Scheme in 2000 to address this issue. Those who qualify for this initiative receive ten kilogrammes of free food grains each month. However, a number of factors prevented the desired goals from

being achieved. In order to improve mental health outcomes for older persons who are food insecure, our study highlights the necessity for focused treatments.

Another element associated with geriatric mental wellness is social prejudice. Our research revealed that older persons experience social discrimination in a variety of ways, including being treated with less decency or respect, experiencing subpar service in public settings like stores or restaurants, receiving threats or harassment, and being made to feel stupid. Social prejudice can lead to feelings of isolation, depression, and loneliness, all of which are harmful to one's mental health. The Indian government has launched a number of programmes to encourage social inclusion and combat prejudice towards older persons. The National Council of Senior Citizens, which was reorganised in 2012 as part of the National Policy on Senior Persons, is one such effort that was first established in 1999.

Providing older persons with chances for full participation in society and ensuring their well-being were the objectives. Health care, social security, and abuse and neglect prevention are among the topics on which the strategy primarily focuses. The Rashtriya Vayoshri Yojana (RVY), a programme that gives older adults living in poverty with assistive devices like hearing aids, walking canes, and eyeglasses, was introduced by the government in 2017. The program's goal is to help older people live better lives and maintain their independence. The National Council of Senior Citizens was also established by the government to provide guidance on issues pertaining to the welfare of senior citizens and to oversee the execution of policies and programmes. Along with

these programmes, the government has also taken action to encourage intergenerational relationships and lessen social isolation among the elderly. One such initiative is the Anubhav programme, which encourages schoolchildren to interact with senior citizens and gain knowledge from their wisdom. The Senior Citizen Savings Scheme, which offers financial security to seniors and assists them in meeting their financial demands, was also introduced by the government. Even with these efforts, social isolation and prejudice against elderly persons in India remain major obstacles. To address these problems and guarantee that senior citizens may live in their communities with respect and dignity, more research is required, as are legislative initiatives. As a result, one effective method for enhancing mental health outcomes is to encourage social involvement and combat social discrimination against older individuals.

It has been discovered that psychological elements including despair, anxiety, and stress are closely related to mental health problems in older persons. Our study emphasises how crucial it is to address psychological issues in order to improve mental health outcomes. Age-related mental health outcomes can be improved by psychological interventions such cognitive behavioural therapy and mindfulness-based stress reduction, according to research [96]. The Indian government has launched a number of measures to encourage older people's psychological health. All individuals, especially the elderly, must have access to mental health services, and primary healthcare must be linked with mental health services, according to the National Mental Health Policy of 2014. In order to lessen stigma and discrimination, the policy also emphasises the importance of encouraging mental health literacy among senior people and their families.

The 2010-launched National Programme for the Health Care of the Elderly (NPHCE) intends to offer elderly individuals complete medical care, including mental health treatments. In medical centres and hospitals across the nation, specialised clinics for elderly patients have been established as part of this programme. These clinics offer older patients mental health services and counselling. The programme also emphasises raising awareness among elderly people and the people who care for them about mental health issues. The government has also started the "National Initiative for the Care of Elderly" (NICE) programme, which aims to give older people a continuum of care, including psychological and social support. The programme calls for the creation of senior daycare facilities, counselling services, and support groups for both seniors and the persons who care for them. The government has also launched a number of programmes and schemes to help the elderly. These include the National Old Age Pension Scheme and the Indira Gandhi National Old Age Pension Scheme, which give monthly pensions to elderly persons who are living in poverty. The psychological strain brought on by financial instability in old life may be lessened with the aid of this financial assistance. Through a variety of policies and programmes, the Indian government has generally made significant efforts to promote older people's psychological well-being. To ensure that all older people have access to mental health services, to combat the stigma attached to mental illness, and to offer comprehensive assistance to older persons for their psychological well-being, much work remains to be done. In order to effectively treat mental health difficulties among older persons, it may be wise to incorporate psychological therapies into mental health policies and practices.

It was discovered through "age-based stratification analysis" that the multimorbidity of mental health conditions rises with advancing years. Our results are in line with earlier research that emphasises the multimorbidity burden rises with age. The growing prevalence of multimorbidity among older adults highlights the need for integrated care strategies that treat several health issues at once. Improvements in health outcomes for older adults with various chronic illnesses have been proven to be possible using integrated care approaches, such as the Chronic Care Model. As a result, integrating care models into mental health policies and practices can be a successful tactic to deal with the growing burden of multimorbidity among older individuals.

Finally, it appears that older adults' mental health issues are frequently disregarded and improperly diagnosed. In order to address mental health difficulties among elderly persons in India, our study highlights the necessity for focused interventions. The Indian government launched the National Programme for Health Care of the Elderly (NPHCE) to address the medical requirements of senior citizens. The program's goal is to offer senior citizens full medical services, including mental health care (Gov. of India, 2011). But implementing the programme will be difficult due to a lack of funding, a shortage of qualified medical staff, and a lack of knowledge about mental health issues among seniors. In order to effectively address mental health difficulties among elderly people in India, these issues must be addressed, and the NPHCE programme must be strengthened.

A serious public health challenge around the world is older people's mental health. Our findings highlight the critical need for better mental health services

for the elderly population, even in the Indian environment where the population of elderly people is expected to grow quickly in the future decades. Due in part to the stigma associated with mental illness, elderly people's mental health issues are frequently disregarded and undertreated in India [97]. Our study underlines the need for focused actions to enhance older Indians' mental health consequences, including enhanced recognition of mental health services and initiatives to lessen social prejudice and food insecurity.

In summary, the research we conducted emphasises the significance of addressing social determinants of health, implementing "age-based stratification analysis," and creating focused interventions to enhance the outcomes of older people's mental health. To address the complex and multidimensional nature of mental health concerns in this population, more study is required to identify successful therapies.

Chapter 6

Conclusions

6.1 Overview

With the increase in volume and complexity of healthcare and social care datasets, there is a need for unorthodox feature learning strategies to address present issues. The analysis of health and biological datasets allows for the evaluation of established computational methods and leads to the development of novel algorithms and methods that can be generalized to challenges from other domains as well. Using these notions, we have not only developed novel algorithms, but also conducted rigorous analyses to address issues related to healthcare and social care. It is to be mentioned that the methodologies and analysis process may be adapted to additional datasets with varied data types and formats as well.

Big data analysis includes classification, regression, and anomaly detection. Reducing dimensionality speeds these operations. Classification may be altered by numerous elements due to biomedical data set variety. For most datasets, visualising methods like t-SNE may not improve analysis performance. Distance may help distinguish data points during dimension reduction. By balancing local-topology with distance preservation, TPDS reduced "MDS-cost" (distance stress), improved visualisation, and boosted "clustering purity" for analysed data sets. This equilibrium may assist other analysis approaches.

TPDS scaling distances maintains local topological information to facilitate "MDS-like" convergence. Higher data point separation may cause outliers that prevent convergence. MDS cost is reduced by weighting samples closer together. This approach preserves local topological information for visualisation. Outliers

substantially influence other "distance-preserving" approaches. Data point collapse and longer convergence computation times arise from its greatness. TPDS outperforms distance-preserving because it weights outliers less. TPDS shows dense "co-localization among data points" of the same class without location collapse and "separability among data points" of different classes. DBSCAN density-based clustering with tight co-localization improves TPDS classification purity. "Density-based clustering approach" is possible with TPDS. TPDS can handle outliers and enhance classification results with time complexity. We're accelerating "SOM" estimated "manifold" here. We studied "TPDS" performance and SOM shape. TPDS' rectangular geometry features four neighbours per node. SOM's hexagonal form was analysed. SOM hexagonal geometry produced roughly equal results to default mode. The protein expression data set showed that "proximity likelihood" averaging made data points from the same class closer in "TPDS." Noise or sample-specific artefacts prohibited t-SNE from estimating the "likelihood of proximity for data points" in "protein expression data".

TPDS visualisation demands an appropriate choice of SOM clusters (grid size), which may have the least impact on subsequent analysis processes like supervised classification and regression. t-SNE and Sammon mapping require parameters. "Sammon mapping" optimises using lambda, while t-SNE depends on perplexity. TPDS follows this rule since one or two criteria determine it. The present distance preservation-based dimension reduction method preserves "heterogeneity among data points" and improves categorisation. We proposed a strategy to overcome these concerns.

NIRD infers regulatory networks from gene expression data with precision. Matrix factorization reduces data dimensionality, thereby enhancing the efficacy of tree-based ensembles. NIRD faced competition from GENIE3, GrnBoost2, and others. The NIRD is superior to these methodologies.
The strategy determines the regulation of complex biological systems. Gene expression data are utilised by systems biology to infer regulatory networks. Using gene expression data to infer regulatory networks can help us treat complex diseases and comprehend fundamental biological processes.

Certainly. NIRD-based "gene regulatory network (GRN)" inference is another area. NIRD outperforms GENIE3 and grnBoost2 in terms of time and golden set edge overlaps. It appears promising to infer "gene regulatory network (GRN)"s from high-dimensional gene expression data. Current research benefits from NIRD's ability to analyse single-cell expression data despite noise and dropout.
The use of NIRD in personalised medicine is possible. The inference of "gene regulatory network (GRN)"s from patient-specific expression data may disclose dysregulated regulatory pathways, allowing for more targeted and effective treatments. Performance of NIRD on single-cell expression data may reveal cell-specific regulatory networks that explain cellular differentiation and progression of disease.

Large datasets increase the computational complexity of NIRD. Matrix factorization reduces data dimensionality and enhances computational performance, but it is difficult to implement on large datasets. It may not be suitable for intricate organisms or tissues.

There are numerous research avenues in this field. First, the method can incorporate information regarding epigenetic and protein-protein interactions. Biological data integration reveals intricate regulatory networks. Second, the recommended method can be evaluated against diverse data sets to demonstrate its utility and resilience. With the proposed method, Bayesian networks or causal inference can increase accuracy and reveal regulatory mechanisms.
Although advantageous, NIRD has disadvantages. First, outliers and disturbance can reduce the precision of matrix factorization. The hyperparameters of a tree-based ensemble approach affect algorithm performance. After hyperparameter tuning, assess NIRD performance. Lastly, the approach is constrained by the quality and quantity of input gene expression data. This research should resolve these flaws and improve the approach's precision and robustness.

The use of NIRD to infer "gene regulatory network (GRN)"s from high-dimensional gene expression data is promising. NIRD effectively manages both bulk and single-cell expression data despite noise and dropout. Consider the method's assumptions and the computational challenges posed by vast datasets. More experimentation is required by NIRD in the areas of personalised medicine, cellular differentiation, and disease development.

One more aspect with reference to the developed novel algorithm was to understand the notion of perplexity. In t-Distributed Stochastic Neighbor Embedding (t-SNE), perplexity is a crucial hyperparameter that balances attention between local and global aspects of the data. It essentially determines how to consider the neighbors of each point, influencing the formation of

clusters in the reduced-dimensional space. The choice of perplexity can significantly affect the results of t-SNE.

In the context of TPDS, spread factor (ρ) is somewhat analogous to perplexity in t-SNE. It determines the scaling of distances between data points, thus impacting how local topology is preserved. Adjusting the spread factor can influence whether the model prioritizes local or global structures in the data. Lagrange Multiplier (λ) helps balance between distance-stress (preservation of global structure) and local-topology preservation. Tuning this parameter can be seen as a way to manage the focus on local versus global features, akin to adjusting perplexity in t-SNE.

In the context of NIRD, matrix factorization parameters influence how much of the data's variance (and thus, structure) is captured. While not directly analogous to perplexity, tuning these parameters affects the balance between capturing detailed local interactions and broader patterns. Regularization Parameters indirectly influence how the model handles local versus global structures by controlling the complexity of the model and its sensitivity to noise, similar to how perplexity affects cluster formation in t-SNE.

In summary, while the concept of perplexity in t-SNE doesn't have a direct equivalent in TPDS or NIRD, similar effects can be achieved by tuning other relevant parameters in these methods. The goal remains to find an optimal balance between capturing the intricacies of local interactions and preserving the overall structure of the data.

Public health includes mental health. Mental health and suicidal behaviours have increased across all demographic cohorts and sexes over the past decade

[18,19]. Public spaces, schools, restaurants, and factories closed due to the rapid spread of coronavirus. Remote work, schooling, and social isolation are common today. Employment has dropped. This event raised global anxiety and depression. No empirical research has used models to predict and explain how life circumstances affect mental health. Our study used "explainable" "probabilistic graphical modelling" with "bootstraps and exact inference methods". We resiliently captured many consequences using this method. Our study found that pre-existing mental disorders increase COVID-19 pandemic mental health risks. Thus, mental health monitoring and treatment should be national regulations. Integrated mental health is financially viable. "Social Security or Supplemental Social Security" reduces anxiety disorders in our model. This is the first proof of these initiatives' success. These measures can be estimated by modelling in different regions with different support structures.

Our US study may inspire cultural and social research in countries with similar or different social structures. In India, where communal living and extended families persist, face-to-face versus online communication may have different effects. US neighbours communicate better than digitally. This shows the fundamental mental health effects of these two factors and the need for more organised research. Evolution may explain the above variations. Isolated primates are depressed [19,20]. Parenthood and "neuropeptide hormones" may explain why having many children lowers anxiety disorder. This matches prior research [21]. COVID-19 provides a unique opportunity to study community psychological responses to public health pandemics.

As the world confronts the pandemic, more research may be needed to understand its life cycle. Our research suggests psychological well-being can be

affected quickly, especially in adolescents. Longitudinal research with the same participants is needed to understand community psychology.

The study shows technology cannot replace human interaction. Thus, more effective and compassionate scientific devices are needed to create a community that discourages seclusion and detachment while maintaining social distance and disease prevention to limit transmission. Customising and adapting these assessments is crucial because prior psychological issues may have a disproportionate impact.

The findings indicate mental health disruption susceptibility. Compared to those with no anxiety or less than a day of stress per week, the system worked best for those with psychological strain for over 5 days. The method may identify vulnerable groups like frontline medical professionals under unfair stress.
"Clinical and public health" models require "transparency and explainability" in complex relationships. Expect "confounding variables, mediation, and inter-causal dependency" and psychological complexity. "Data-driven structure learning of Bayesian networks" expanded associative analysis. This method was preferred over "black-box machine learning" and statistical modelling for several reasons. "Structure learning" can identify and represent "confounding factors" openly, while black-box ML models like "random forests and gradient boosted machines" are not optimised for "transparent" justification. Statisticians struggle to model thousands of variables. "Structure learning" analyses "mediation, confounding, and inter-causal effects" by breaking down relationships. An "ensemble" method that selects the "ensemble's voted structure" from 101 "Bayesian networks" addresses erroneous learning. This

study proposes a "AI" approach to complex psychological wellness issues, similar to public health research. [20,21]

This study's limitations. "Confounding" makes "causal inference" in "cross-sectional data" difficult. The "structure learning approach" was chosen to identify and explain "confounding factors". 101 "Bayesian Networks (BNs)" addressed erroneous learning in this study. "Ensemble approach" selected a model from "bootstrapped structures" majority vote. The "bootstrapped structures" used made this method robust. "Probabilistic reasoning" reveals psychological wellness factors. It estimates COVID-19-related psychological mortality. Unknown reasons may lower anxiety in cancer and hypertension patients. Reducing workplace stress or increasing family interaction may explain this. The dataset lacks detail. Unknown duration. US-only findings. The model's validity is increased by the large number of respondents and multiple ethnic groups. Our research advances explainable artificial intelligence to forecast population-wide mental health outcomes using survey data, increasing its potential for widespread adoption. Noisy survey data. Our 80%-accurate method balances information discovery and predictability. New scenarios have baselines. Algorithms can screen for help, and adding measurements and characteristics may improve prognostications. Predictive models for screening and assessing COVID-19's mental health effects are crucial for proactive treatment and prevention.

Models based on probabilistic graphs identify the most mentally vulnerable seniors and provide intervention and support. Socioeconomic and psychological factors affecting the mental health of older adults were investigated. Food

insecurity, social discrimination, and psychological factors impact the mental health of seniors significantly. Age-based stratification revealed that multimorbidity in mental health increases with age and that certain age groups are more susceptible. These findings have implications for Indian and global geriatric health policies.

There are several limitation outlined in the thesis, which may be considered as areas of improvement and further research:

Algorithm Optimization and Efficiency: One of the primary limitations of the current study is the computational efficiency of the algorithms, especially when dealing with very large datasets. The author plans to optimize these algorithms, potentially by exploring more efficient data structures, parallel processing, or leveraging advances in GPU computing. This would allow for faster processing times and the ability to handle larger datasets more effectively.

Validation on Diverse Datasets: The current study's methodologies, such as TPDS and NIRD, have been tested on specific datasets. The author plans to validate these methods on a more diverse range of datasets from various domains. This broader testing can help fine-tune the algorithms and demonstrate their applicability and robustness across different types of data.

Enhancing Interpretability: While the current methods provide improved interpretability of high-dimensional data, there is still room for enhancement. The author intends to incorporate additional techniques or metrics that can provide deeper insights into the data, particularly in understanding the biological or healthcare implications of the findings.

Addressing Overfitting and Generalization: The risk of overfitting in high-dimensional data analysis is a concern. The author plans to develop and integrate techniques to better assess the generalizability of the models and to ensure that the findings are not just artifacts of the specific data sets used but are truly representative of broader trends and patterns.

Incorporating Feedback from Domain Experts: Collaborating with experts in fields like genomics, healthcare, and computational biology can provide valuable insights into how these methods can be refined and applied more effectively. This interdisciplinary collaboration can also help in identifying new challenges and opportunities for applying these techniques.

Longitudinal and Dynamic Data Analysis: Another potential area of improvement is the application of these methods to longitudinal and dynamic datasets, where the data change over time. This requires adapting the algorithms to account for temporal aspects, which can provide richer insights, especially in healthcare and genomics.

In summary, the author's plan to overcome the limitations of the current study involves a combination of technical optimizations, broader validation, enhanced interpretability, addressing overfitting concerns, interdisciplinary collaboration, and exploring new applications and extensions of the developed methodologies.

Global geriatric mental health concerns exist. This study found that older food insecure individuals had mental health issues. In both developing and developed nations, the elderly are malnourished. The Indian government has enhanced food security and accessibility. Important are PDS, ICDS, MDM, SNP, WNP, ANP, and AAY. These programmes have decreased food insecurity and increased

access for vulnerable groups. Persistent are PDS leaks and scheme implementation issues. India has advanced food insecurity initiatives. The 2000 Annapurna Scheme of the Ministry of Rural Development reduces food insecurity among senior citizens not covered by NOAPS or other pension programmes. Beneficiaries receive 10 kg of free food grains per month. However, numerous factors hindered outcomes. Our study emphasises targeted food insecurity interventions for the elderly to improve mental health.

Social discrimination impacts the mental health of the elderly. Our research revealed that older individuals are treated with less courtesy, receive inferior service in restaurants and stores, are threatened or harassed, and are viewed as less intelligent. Isolation, depression, and social discrimination can all be detrimental to mental health. The government of India has taken numerous steps to combat ageism and social exclusion. The 1999 National Policy on Senior Citizens evolved into the 2012 National Council on Ageing. The objectives were senior health and full social participation. The coverage includes health, social security, and abuse and neglect. The 2017 Rashtriya Vayoshri Yojana (RVY) provided hearing aids, walking sticks, and glasses to low-income senior citizens. The programme enhances the quality of life and autonomy of seniors. The National Council of Senior Citizens provides advice regarding the welfare of senior citizens. The government promotes intergenerational relationships and reduces social isolation among the elderly. Anubhav advises students to gain knowledge from their elders. The Senior Citizen Savings Scheme assists senior citizens. Despite these efforts, older Indians continue to face discrimination and social exclusion. There is a need for more research and policy interventions to help older people live with dignity in their communities. Thus, social inclusion

and addressing social discrimination can improve the mental health of older individuals.

Depression, anxiety, and stress were symptoms of mental illness in the elderly. We discovered that psychological factors enhance mental health. Cognitive-behavioral therapy and mindfulness-based stress reduction are advantageous for older adults. Multiple Indian government initiatives have promoted the psychological health of senior citizens. The National Mental Health Policy for 2014 emphasizes the integration of mental health services into primary care for all citizens, including the elderly. Mental health education is provided to seniors and their families to reduce stigma and discrimination. The 2010 National Programme for the Health Care of the Elderly (NPHCE) addresses senior mental health. Under this programme, mental health clinics for elderly patients have opened nationwide. Seniors and carers are instructed in mental health. The "National Initiative for the Care of the Elderly" (NICE) provides psychological and social support to the elderly. The programme provides counseling, caregiving support, and adult day care. The National and Indira Gandhi National Old Age Pension Schemes provide seniors living below the poverty line with monthly pensions. Financial insecurity in old age can strain the mind. Through policies and programmes, the Indian government has promoted the psychological well-being of senior citizens. There is still much work to be done to provide all older adults with mental health services, eliminate the stigma associated with mental illness, and promote their psychological well-being. Thus, psychological interventions in mental health policies and practises can benefit mentally ill older adults. Multimorbidity in mental health by age increased. Our findings support multimorbidity with age. As the

multimorbidity of the elderly increases, they require integrated care. The Chronic Care Model benefits older adults with multiple chronic conditions. Consequently, integrating mental health policies and practises reduces the multimorbidity of older adults.

The methodologies developed in this thesis can enhance machine learning models in healthcare AI by providing improved feature learning and dimensionality reduction techniques. This is especially important in developing predictive models for disease diagnosis and treatment, ensuring these models are both accurate and interpretable. The use of probabilistic graphical models, as discussed in the thesis, can be instrumental in public health for discovering causal mechanisms of disease outcomes and estimating health concerns such as mental health prevalence. This approach can lead to more effective public health strategies and interventions.

The methodologies can also be applied to social data analysis, where understanding trends and patterns in human behavior and sentiment is crucial. The ability to handle high-dimensional data effectively can lead to more insightful analyses of social phenomena. In the broader field of data science, the techniques developed in this thesis can be applied to various types of high-dimensional data, enabling more effective data analysis, visualization, and interpretation across multiple domains.

Analyzing the US dataset in the context of global changes, particularly in regions like India or other under-developed countries, can have significant implications on the results obtained. Several factors need to be considered:

Genetic Diversity and Population Heterogeneity: The genetic makeup and diversity in a population like India, which is significantly different from that of the US, can affect the results. Diseases may manifest differently due to genetic variations, and certain genetic markers or traits prevalent in one population may not be as significant in another.

Environmental and Lifestyle Factors: Under-developed countries often have different environmental exposures and lifestyle patterns compared to developed countries like the US. Factors such as pollution levels, dietary habits, prevalence of infectious diseases, and healthcare access can significantly influence health outcomes and must be considered when analyzing healthcare data.

Disease Prevalence and Health Priorities: The prevalence of certain diseases and health priorities can vary greatly between developed and under-developed countries. For instance, infectious diseases and malnutrition-related conditions might be more prevalent in under-developed countries, whereas lifestyle-related diseases might be more common in the US. This variation can affect the relevance and applicability of findings from the US dataset to other regions.

Socioeconomic Factors: Socioeconomic conditions such as poverty, education, and healthcare infrastructure significantly impact health outcomes. Under-developed countries may face challenges like limited access to healthcare, which can skew data analysis and interpretation if the model developed based on the US dataset does not account for these disparities.

Data Representation and Bias: If the US dataset does not adequately represent the global diversity, applying the findings to regions like India might lead to

biased or inaccurate conclusions. It's essential to ensure that the data and the models are representative and inclusive of diverse populations.

Cultural and Ethical Considerations: Cultural differences can influence health behaviors, treatment adherence, and disease perception. Ethical considerations, especially in terms of data collection and privacy, can also vary between regions, affecting the data's availability and quality.

Global Health Trends and Emerging Challenges: Changes in global health trends, such as the rise of antimicrobial resistance or the impact of climate change on health, might affect different regions in distinct ways. Understanding these trends and their implications on various populations is crucial for accurate analysis and application of the results.

In conclusion, while the methodologies developed in the thesis can provide valuable insights into high-dimensional healthcare data, their application to datasets from regions like India or other under-developed countries requires careful consideration of the above factors. Tailoring the analysis to account for these regional differences is essential for ensuring the accuracy and relevance of the results in a global context.

Lastly, the mental health issues of older adults appear undiagnosed. Our study suggests older Indians receive mental health interventions. NPHCE serves the health needs of older Indians. Comprehensive mental health care is provided to the elderly (Government of India, 2011). However, lack of funding, health care professionals, and awareness of older people's mental health make programme implementation challenging. These obstacles and NPHCE programme enhancements can assist older Indians with mental health problems. Mental

health in the elderly is a global public health concern. Our findings indicate that the rapidly ageing population of India requires improved mental health services. Due to stigma, older Indians with mental illness are neglected [32]. Our study emphasises the need for targeted interventions to improve the mental health of older Indians, such as increased access to mental health services and decreased social discrimination and food insecurity. Our research focuses on social determinants of health, age-based stratification analysis, and targeted interventions to improve older adults' mental health. Research is required to treat the complex mental health issues of this population.

References

1. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: With applications in R. 2013th ed. New York, NY: Springer; 2013. Available: https://play.google.com/store/books/details?id=qcI_AAAAQBAJ

2. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.

3. Brunton SL, Nathan Kutz J. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press; 2022.

4. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17: 147–154.

5. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26: 303–304.

6. Blouvshtein L, Cohen-Or D. Outlier Detection for Robust Multi-Dimensional Scaling. IEEE Trans Pattern Anal Mach Intell. 2019;41: 2273–2279.

7. Sammon JW. A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers. 1969. pp. 401–409. doi:10.1109/t-c.1969.222678

8. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9. Available: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl

9. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290: 2323–2326.

10. Lee JA, Verleysen M. Nonlinear Dimensionality Reduction. Springer New York; 2010.

11. Bunte K, Haase S, Biehl M, Villmann T. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. Neurocomputing. 2012;90: 23–45.

12. Gorban AN, Kégl B, Wunsch DC, Zinovyev A. Principal Manifolds for Data Visualization and Dimension Reduction. Springer Science & Business Media; 2007.

13. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. 2008. J Mach Learn Res.

14. Dzemyda G, Kurasova O, Žilinskas J. Multidimensional Data Visualization. Springer New York; 2013.

15. De Silva V, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. [cited 14 Nov 2022]. Available: https://cecas.clemson.edu/~stb/ece904/fall2004/silva-nips2003-global-local.pdf

16. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20: 273–282.

17. Su Y, Shi Q, Wei W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. Proteomics. 2017;17. doi:10.1002/pmic.201600267

18. Pang Y, Zhang L, Liu Z, Yu N, Li H. Neighborhood Preserving Projections (NPP): A Novel Linear Dimension Reduction Method. Advances in Intelligent Computing. Springer Berlin Heidelberg; 2005. pp. 117–125.

19. Website. Available: Khosla, K.; Jha, I.P.; Kumar, A.; Kumar, V. Local-Topology-Based Scaling for Distance Preserving Dimension Reduction Method to Improve Classification of Biomedical Data-Sets. Algorithms 2020, 13, 192. https://doi.org/10.3390/a13080192

20. Naranjo L, Pérez CJ, Martín J, Campos-Roca Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording

replications. Comput Methods Programs Biomed. 2017;142: 147–156.

21. Zarchi MS, Fatemi Bushehri SMM, Dehghanizadeh M. SCADI: A standard dataset for self-care problems classification of children with physical and motor disability. Int J Med Inform. 2018;114: 81–87.

22. Higuera C, Gardiner KJ, Cios KJ. Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. PLoS One. 2015;10: e0129126.

23. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49: 708–718.

24. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans Database Syst. 2017;42: 1–21.

25. Website. Available: Khosla K, Jha IP, Kumar A, Kumar V. Local-Topology-Based Scaling for Distance Preserving Dimension Reduction Method to Improve Classification of Biomedical Data-Sets. Algorithms. 2020; 13(8):192. https://doi.org/10.3390/a13080192

26. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML]. 2018. Available: http://arxiv.org/abs/1802.03426

27. Saxena A, Gupta A, Mukerjee A. Non-linear Dimensionality Reduction by Locally Linear Isomaps. Neural Information Processing. Springer Berlin Heidelberg; 2004. pp. 1038–1043.

28. Wang S, Karikomi M, MacLean AL, Nie Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. 2019;47: e66.

29. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics. 2018;19: 232.

30. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14: 1083–1086.

31. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5. doi:10.1371/journal.pone.0012776

32. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35: 2159–2161.

33. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33: 2314–2321.

34. Chan TE, Stumpf MPH, Babtie AC. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Syst. 2017;5: 251–267.e3.

35. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65: 631–643.e4.

36. Zhou Q, Chipperfield H, Melton DA, Wong WH. A gene regulatory network in mouse embryonic stem cells. Proc Natl Acad Sci U S A. 2007;104: 16438–16443.

37. Enge M, Arda HE, Mignardi M, Beausang J, Bottino R, Kim SK, et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell. 2017;171: 321–330.e14.

38. Jolliffe IT. Principal Component Analysis New York Springer-Verlag. Inc; 2002.

39. Turk M, Pentland A. Eigenfaces for recognition. J Cogn Neurosci. 1991;3: 71–86.

40. Lee D, Seung HS. Unsupervised learning by convex and conic coding. Adv Neural Inf Process Syst. 1996;9. Available: https://proceedings.neurips.cc/paper/1996/hash/2de5d16682c3c35007e4e92982f1a2ba

-Abstract.html

41. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401: 788–791.

42. Field DJ. What is the goal of sensory coding? Neural Comput. 1994;6: 559–601.

43. Földiák P, Young MP. The handbook of brain theory and neural networks, chapter Sparse coding in the primate cortex. MIT Press, Cambridge; 1995.

44. Paatero P. Least squares formulation of robust non-negative factor analysis. Chemometrics Intellig Lab Syst. 1997;37: 23–35.

45. Lee D, Seung HS. Algorithms for non-negative matrix factorization. Adv Neural Inf Process Syst. 2000;13. Available: https://proceedings.neurips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization

46. Welcome to Nimfa — Nimfa 1.3.1 documentation. [cited 18 May 2023]. Available: https://nimfa.biolab.si/index.html#

47. As jobs crisis deepens, ILO warns of uncertain and incomplete labour market recovery. 2020 [cited 14 Nov 2022]. Available: https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_749398/lang--en/index.htm

48. Mrklas K, Shalaby R, Hrabok M, Gusnowski A. Prevalence of perceived stress, anxiety, depression, and obsessive-compulsive symptoms in health care workers and other workers in Alberta during the …. JMIR Mental. 2020. Available: https://mental.jmir.org/2020/9/e22408/

49. Psychiatry TL, The Lancet Psychiatry. Mental health and COVID-19: change the conversation. The Lancet Psychiatry. 2020. p. 463. doi:10.1016/s2215-0366(20)30194-2

50. Pereira-Sanchez V, Adiukwu F, El Hayek S, Bytyçi DG, Gonzalez-Diaz JM, Kundadak GK, et al. COVID-19 effect on mental health: patients and workforce. Lancet Psychiatry. 2020;7: e29–e30.

51. Chen Q, Liang M, Li Y, Guo J, Fei D, Wang L, et al. Mental health care for medical

staff in China during the COVID-19 outbreak. Lancet Psychiatry. 2020;7: e15–e16.

52. Mahase E. Covid-19: Mental health consequences of pandemic need urgent research, paper advises. BMJ. 2020;369: m1515.

53. Seabrook EM, Kern ML, Rickard NS. Social Networking Sites, Depression, and Anxiety: A Systematic Review. JMIR Ment Health. 2016;3: e50.

54. Bruen AJ, Wall A, Haines-Delmont A, Perkins E. Exploring Suicidal Ideation Using an Innovative Mobile App-Strength Within Me: The Usability and Acceptability of Setting up a Trial Involving Mobile Technology and Mental Health Service Users. JMIR Mental Health. 2020;7: e18407.

55. Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. Mental health problems and social media exposure during COVID-19 outbreak. PLoS One. 2020;15: e0231924.

56. Homepage. In: Untitled [Internet]. [cited 14 Nov 2022]. Available: https://www.covid-impact.org

57. COVID impact survey. In: Untitled [Internet]. [cited 14 Nov 2022]. Available: https://www.covid-impact.org/about-the-survey-questionnaire

58. Revelle W, Revelle MW. Procedures for psychological, psychometric, and personality. R package Bpsychˆ, Version. 2017. Available: https://www.yumpu.com/en/document/view/37947343/package-psych-the-personality-project

59. Gámez JA, Mateo JL, Puerta JM. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min Knowl Discov. 2011;22: 106–148.

60. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987;52: 345–370.

61. Yedidia JS, Freeman W, Weiss Y. Generalized Belief Propagation. Adv Neural Inf Process Syst. 2000;13. Available: https://proceedings.neurips.cc/paper/2000/hash/61b1fb3f59e28c67f3925f3c79be81a1-Abstract.html

62. Sethi T, WiseR MS. A Shiny Application for End-to-End Bayesian Decision Network Analysis and Web-Deployment.; 2018.

63. Parsons S. Probabilistic Graphical Models: Principles and Techniques by Daphne Koller and Nir Friedman, MIT Press, 1231 pp., $95.00, ISBN 0-262-01319-3. Knowl Eng Rev. 2011;26: 237–238.

64. Bilsen J. Suicide and Youth: Risk Factors. Front Psychiatry. 2018;9: 540.

65. Hennessy MB, McCowan B, Jiang J, Capitanio JP. Depressive-like behavioral response of adult male rhesus monkeys during routine animal husbandry procedure. Front Behav Neurosci. 2014;8: 309.

66. Sethi T, Mittal A, Maheshwari S, Chugh S. Learning to Address Health Inequality in the United States with a Bayesian Decision Network. AAAI. 2019;33: 710–717.

67. Awasthi R, Patel P, Joshi V, Karkal S, Sethi T. Learning Explainable Interventions to Mitigate HIV Transmission in Sex Workers Across Five States in India. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2012.01930

68. Luppa M, Luck T, Brähler E, König H-H, Riedel-Heller SG. Prediction of Institutionalisation in Dementia. Dement Geriatr Cogn Disord. 2008;26: 65–78.

69. Weyerer S, Eifflaender-Gorfer S, Köhler L, Jessen F, Maier W, Fuchs A, et al. Prevalence and risk factors for depression in non-demented primary care attenders aged 75 years and older. J Affect Disord. 2008;111: 153–163.

70. United Nations. World Population Ageing 2019 Highlights. United Nations; 2019.

71. Nations U. World population ageing 2019 highlights. U N Disarm Yearb.

72. Ruan Y, Guo Y, Zheng Y, Huang Z, Sun S, Kowal P, et al. Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1. BMC Public Health. 2018;18: 778.

73. Fried LP, Ferrucci L, Darer J, Williamson JD, Anderson G. Untangling the concepts of disability, frailty, and comorbidity: implications for improved targeting and care. J Gerontol A Biol Sci Med Sci. 2004;59: 255–263.

74. Ferrucci L, Guralnik JM, Studenski S, Fried LP, Cutler GB Jr, Walston JD, et al. Designing randomized, controlled trials aimed at preventing or delaying functional decline and disability in frail, older persons: a consensus report. J Am Geriatr Soc. 2004;52: 625–634.

75. Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. Alzheimers Dement. 2013;9: 63–75.e2.

76. Skoog I. Psychiatric Disorders in the Elderly. The Canadian Journal of Psychiatry. 2011. pp. 387–397. doi:10.1177/070674371105600702

77. Skoog I. Why should general psychiatrists learn more about mental disorders in the elderly? Canadian journal of psychiatry. Revue canadienne de psychiatrie. 2011. pp. 385–386.

78. Waring EM. Psychiatric illness in the elderly. Am Fam Physician. 1980;21: 109–112.

79. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T. Global Burden of Mental, Neurological, and Substance Use Disorders: An Analysis from the Global Burden of Disease Study 2010. In: Patel V, Chisholm D, Dua T, Laxminarayan R, Medina-Mora ME, editors. Mental, Neurological, and Substance Use Disorders: Disease Control Priorities, Third Edition (Volume 4). Washington (DC): The International Bank for Reconstruction and Development / The World Bank;

80. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T. The global burden of mental, neurological and substance use disorders: an analysis from the Global Burden of Disease Study 2010. PLoS One. 2015;10: e0116820.

81. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386: 743–800.

82. Stuart H. Reducing the stigma of mental illness. Glob Ment Health (Camb). 2016;3: e17.

83. Gaiha SM, Taylor Salisbury T, Koschorke M, Raman U, Petticrew M. Stigma

associated with mental health problems among young people in India: a systematic review of magnitude, manifestations and recommendations. BMC Psychiatry. 2020;20: 538.

84. [No title]. [cited 24 Feb 2023]. Available: https://www.apa.org/pi/ses/resources/publications/age

85. Mushtaq R. Relationship Between Loneliness, Psychiatric Disorders and Physical Health ? A Review on the Psychological Aspects of Loneliness. JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH. 2014. doi:10.7860/jcdr/2014/10077.4828

86. Wu B. Social isolation and loneliness among older adults in the context of COVID-19: a global challenge. Glob Health Res Policy. 2020;5: 27.

87. Reiss F, Meyrose A-K, Otto C, Lampert T, Klasen F, Ravens-Sieberer U. Socioeconomic status, stressful life situations and mental health problems in children and adolescents: Results of the German BELLA cohort-study. PLoS One. 2019;14: e0213700.

88. Warner LM, Ziegelmann JP, Schüz B, Wurm S, Tesch-Römer C, Schwarzer R. Maintaining autonomy despite multimorbidity: self-efficacy and the two faces of social support. Eur J Ageing. 2011;8: 3–12.

89. Ahmedani BK. Mental Health Stigma: Society, Individuals, and the Profession. J Soc Work Values Ethics. 2011;8: 41–416.

90. Longitudinal Ageing Study in India (LASI). [cited 24 Feb 2023]. Available: https://www.iipsindia.ac.in/lasi

91. Lee J, McGovern ME, Bloom DE, Arokiasamy P, Risbud A, O'Brien J, et al. Education, gender, and state-level disparities in the health of older Indians: Evidence from biomarker data. Econ Hum Biol. 2015;19: 145–156.

92. Website. Available: https://www.iipsindia.ac.in/lasi.

93. Radloff LS. The CES-D Scale. Applied Psychological Measurement. 1977. pp. 385–401. doi:10.1177/014662167700100306

94. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012.

95. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press; 2009.

96. Weidman AC, Cheng JT, Tracy JL. The psychological structure of humility. Journal of Personality and Social Psychology. 2018. pp. 153–178. doi:10.1037/pspp0000112

97. Grover S, Malhotra N. Depression in elderly: A review of Indian research. Journal of Geriatric Mental Health. 2015;2: 4.