Cloud based Network Analysis Model for Predicting Disease-Diet Associations

*A **Thesis***

submitted in partial fulfillment of the requirements for the award of degree of

Doctor of Philosophy

by

Rashmeet Toor

(901603013)

under the guidance of

Dr. Inderveer Chana

Professor

Computer Science and Engineering Department

Dean (Student Affairs)

Thapar Institute of Engineering and Technology

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

Patiala-147004, INDIA

July 2023

Contents

List of Figures

List of Tables

List of Abbreviations

ICU	Intensive Care Unit
EHR	Electronic Health Records
CDSS	Clinical Decision Support Systems
WWW	World Wide Web
SaaS	Software as a Service
PaaS	Platform-as-a-Service
IaaS	Infrastructure-as-a-Service
IBD	Inflammatory Bowel Disease
AWS	Amazon Web Services
DASH	Dietary Approaches to Stop Hypertension
miRNA	Micro Ribonucleic acid
HMDD	Human MicroRNA Disease Database
DO	Disease Ontology
MeSH	Medical Subject Headings
OMIM	Online Mendelian Inheritance in Man
HPO	Human Phenotype Ontology
BMI	Body Mass Index
HITS	Hyperlink Induced Topic Search
IoT	Internet of Things
SIoT	Social Internet of Things
PPI	Protein–Protein Interaction
NP	Network Propagation
RWR	Random Walk with Restart
SP	Shortest Path
SVM	Support Vector Machine
SNA	Social Network Analysis
AUC	Area Under Curve
ASLIP	Age-series based link prediction
SMILES	Simplified Molecular-Input Line-Entry System

SIDER	Side Effect Resource
CTD	Comparative Taxicogenomics Database
MSigDB	Molecular Signatures Database
HMDAD	Human Microbe Disease Association Database
WHO	World Health Organization
ORA	Organization Risk Analyzer
DNA	Dynamic Network Analysis
ADNI	Alzheimer's Disease Neuroimaging Initiative
SR	SimRank
CN	Common Neighbours
AA	Adamic Adar
CC	Clustering Coefficient
D	Diameter
BC	Betweenness Centrality
ATUS	American Time Use Survey
HEI	Healthy Eating Index
HR	Heart Rate
SBP	Systolic Blood Pressure
SD	Steps per Day
EC2	Elastic Cloud Compute
OCR	Optical Character Recognition
ECG	ElectroCardioGram
DIDACE	DIsease Diet Association database Curator and Explorer
NLM	National Library of Medicine
NCBI	National Center for Biotechnology Information
ANN	Artificial Neural Network
MLP	Multi-Layer Perceptron
TF-IDF	Term Frequency-Inverse Document Frequency
NIH	National Institutes of Health
CAN	Center for Applied Nutrition
CQL	Cypher Query Language
NAFLD	Non Alcoholic Fatty Liver Disease
T2DM	Diabetes Mellitus Type 2
UC	Ulcerative Colitis
CD	Crohns Disease

PR	Page Rank
KNN	K Nearest Neighbors
CART	Classification And Regression Trees
GB	Gradient Boosting
NB	Naïve Bayes
SE	Stacking Ensemble
UCM	University of Central Missouri
LA	Louvain Algorithm
MD	Mediterranean Diet
DAPNEML	Disease-diet Association Prediction in NEtwork using Machine Learning
SMOTE	Synthetic Minority Oversampling Technique
IAM	Identity and Access Management
HVM	Hardware assisted Virtual Machine
API	Application Programming Interface
URL	Uniform Resource Locator

Certificate

I hereby certify that the work which is being presented in this thesis entitled "**Cloud based Network Analysis Model for Predicting Disease-Diet Associations**", in partial fulfillment of the requirements for the award of degree of "**Doctor of Philosophy**" submitted in Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed University), Patiala, India, is an authentic record of my own work carried out under the supervision of **Dr. Inderveer Chana** and refers other research works which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for the award of any other degree of this or any other university.

(**Rashmeet Toor**)

Regn. No. 901603013

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(**Dr. Inderveer Chana**)

Professor

Computer Science and Engineering Department

Dean (Student Affairs)

Thapar Institute of Engineering and Technology (Deemed University)

Patiala, 147004

Punjab, INDIA.

Acknowledgements

I am grateful to the Almighty for granting countless blessings, opportunities, knowledge and strength, without which I could not have accomplished this work.

I am indebted to my supervisor, Dr. Inderveer Chana, Professor and Dean (Student Affairs), Computer Science and Engineering Department), Thapar Institute of Engineering and Technology (Deemed University), Patiala (India), for her continued guidance and support. I have benefitted greatly from her wealth of knowledge and meticulous editing. The time and energy she devoted towards this work has immensely improved the quality of this research. Her faith in me and encouragement has helped me stay focused over the years. Meeting her in tough times filled me up with hope and motivation. Her unassuming approach towards research and life is a source of inspiration, which I hope to carry forward throughout my career.

I offer my sincere gratitude to our director Prof. Prakash Gopalan, Dr. Rafat Siddique, Dean (Research Sponsored Projects) and Dr. Shalini Batra, Professor and Head, Computer Science and Engineering Department for providing the necessary academic and administrative assistance in completion of this work. I am thankful to my Doctoral committee members- Dr. Singara Singh, Associate Professor and Dr. Nitin Saxena, Assistant Professor for ensuring the progress of my research work. Special appreciation for Dr. N. Tejo Prakash, Professor, School of Energy and Environment for their constructive comments and useful suggestions. I am also thankful to PhD. Coordinator Dr. Sushma Jain and all the faculty and staff members of Computer Science and Engineering Department for always helping me and being a source of inspiration. I wish to further extend my thanks to all the lab mates and peers for the constant support and making this journey memorable.

It would not have been possible without my mothers' unconditional love and belief in me. Deepest thanks to my dear mother, father and sister for encouraging me to start this journey and providing me every possible support throughout. I would always be grateful to them for their patience and compassion in every sphere of my life. Immense gratitude to my loving husband Dilpreet Singh for the sacrifices he made in order for me to pursue this degree. His unconditional support and eternal confidence

in me has always been an inspiration. I would always be indebted to him for his love and commitment. Special thanks to my in-laws family for believing in me and being a pillar of support and encouragement. I gratefully acknowledge their cooperation and blessings for me. My warm and heartfelt thanks to all my friends for the hope and encouragement they have given me which motivated me throughout this degree. Their kind words have always been re-energizing.

September, 2022 (Rashmeet Toor)

Patiala

Abstract

Predictive analytics in healthcare is an integration of computational technologies and healthcare domain for retrieval, storage and analysis of medical data. With the immense progress in computational techniques and technologies, healthcare domain has witnessed unparalleled achievements since the last decade. Comprehending the relationship between health and diet is another such area which presents numerous opportunities for predictive healthcare. Disease-diet associations pose an arduous problem in computational domain because of the evident complex interdependencies. The intertwined relations among disease, diet and their subtypes along with the varying nature of their associations (harmful or helpful) adds to the complexity. Thus, the associations need to be explored with a close integration of significant computational techologies. Predictive analysis of such associations would be an aid for healthcare professionals to foresee the risk of occurrence or progression of a disease on the basis of diet and thus make informed decisions.

This work aims to efficiently and effectively predict unknown disease-diet associations using integrated computational technologies. To achieve this, initially, a review of the work done in exploring the relation of disease and diet has been undertaken. It is evident from the review that several studies aim to explore the associations, but they have been designed for a specific disease and diet combination. It is also realized that while some disease-diet associations are well established since ages, there are others which have found acknowledgement only in literature. This presents an opportunity for bringing together such studies and exploiting them on a large scale. Further, a survey of the existing services and techniques designed for understanding relation of disease and other factors like drugs, symptoms etc. has been done. It highlights a plethoric use of an upcoming technology Network Analysis for representing and analyzing complex relationships. Thus, a further investigation of Network Analysis and its role in predictive and healthcare applications has been conducted. Various challenges that are posed while exploiting Network Analysis for healthcare along with measures that might be adopted for overcoming the challenges have also been discussed. Several promising applications of Network Analysis in healthcare domain have been proposed.

As an outcome of the survey, Network Analysis is deemed to be a significant technology for exploring disease-diet associations. Considering the complexity of computations involved in this task, there is also a need of a platform which assists effective analysis and does not compromise with performance in case of higher load. Another propitious technology Cloud computing is found to be suitable for this work, given its extensive application in healthcare domain, which has also been reviewed. As a consequence, amalgamation of Network Analysis and Cloud Computing are established as a great fit for exploring disease-diet associations.

Data corresponding to disease-diet associations is not available as such, thereby it becomes necessary to extract it from the literature. It is also recognized that visualization of known disease-diet associations in the form of a graph and its quantification offers opportunities for advanced learning. Thus, a Network Model DID-NEM has been proposed and developed for extracting, visualizing and modelling disease-diet associations. Firstly, a custom-made automatic technique DIDACE is utilized to extract and quantify the associations using literature mining. This eliminates the drawbacks of manual curation and assist in fast and efficient extraction. 2,74,131 records containing 1917 different diseases and 143 diet terms have been extracted using this technique. Further, nature of a subset of associations are predicted by performing sentiment analysis using a MLP with an accuracy of 86%. The associations are then transformed into a graph to be readily available for analysis. DID-NEM is novel and can be utilized by domain researchers for extraction of associations between entities other than disease and diet. It also contributes a novel disease-diet associations database to the research community for further study.

A prediction approach PredNEM has been proposed to accomplish manipulation and analysis of the curated database. PredNEM aims to firstly quantify and integrate different networks like disease-diet, disease-disease or diet-diet by utilizing different resources including curated database, pattern mining and semantic similarity. Further, two different learning methods, TBM and TFM have been engineered from a combination of network algorithms/parameters and machine learning for prediction of unknown associations. The first method starts by finding communities in the network

followed by ranking of nodes for finding most similar nodes, while the second method crafts network algorithms/parameters as features, compares different machine learning algorithms and select the best performing for prediction. Validation of PredNEM and its two learning methods have been demonstrated through two different case studies corresponding to Covid-19 and Inflammatory Bowel Disease (IBD) respectively. Out of top 20 diets predicted for Covid-19, some enthralling associations have been validated through literature including kefir, carrot and strawberry. In the second case study, nature of 16 out of 21 associations has been correctly predicted as per dietician and medical literature for Crohn's disease using Naive Bayes classifier with ROC AUC value of 82.7%. The predictions enhance traditonal know-how of domain experts and help them to stay updated. These are also beneficial for researchers in further study.
Cloud platform has been introduced for provisioning the network based analysis as an efficient and adaptable service to the concerned stakeholders. EC2 and Neo4j instances have been created for deploying the case study of IBD over cloud and connecting to graph database respectively. Results of prediction are provided through an accessible cloud service named CloudMenu. Optimum values of CPU utilization and throughput suggest good performance and better resource utilization.
This work contributes significantly by developing a Network model DID-NEM which involves automatic curation of disease-diet associations through DIDACE and its visualization as a network. PredNEM further utilizes network algorithms/parameters and machine learning for advanced analysis. This network based analysis is deployed over cloud, making the service CloudMenu easy to use, flexible and economical.

Chapter 1

Introduction

Healthcare domain has witnessed immense progress in the last decade owing to novel inventions and outstanding research. This has led to generation of data which is vast and complex but at the same time it is now more transparent, searchable and reusable. The formulation of technologies and computational techniques to handle the ever-increasing healthcare data has led to evolution of predictive analytics in healthcare.

Understanding the relationship between health and diet is one such domain which presents numerous opportunities for predictive healthcare. There are many unfamiliar and complex interdependencies between diseases and diets which are evident in literature but have not been fully explored. Analysis of such complex associations is extremely beneficial for domain experts to understand the interdependencies. Prediction of previously unknown associations might aid medical researchers to discover the risk of occurrence or progression of a disease and thus suggesting methods to improve dietary habits.

This Chapter provides an overview of predictive analytics in healthcare along with discussion of the current state of research in understanding disease-diet associations. It further unveils closely related computational technologies Network Analysis and Cloud Computing. Based on the existing scenario, problem statement is formulated and motivation of this research is outlined. The Chapter concludes with the contribution of this research work and its organization.

1.1 Predictive Analytics in Healthcare: An Overview

Predictive analytics in healthcare is a branch of Health Informatics and Computer Science which deals with retrieval, analysis, storage and processing of healthcare data with the help of computational techniques [2, 3]. It dates back from the realization of importance of healthcare data and its storage. This data was initially just stored so as to use it for gaining insights. With better computational techniques, refinement of raw data became possible which led to the term "information". This information was further used to understand patterns using traditional rule based inferences or data mining. "Knowledge" in the form of patterns is currently being analysed using different advanced analysis techniques like machine learning, text mining etc. Such kind of analysis forms the basis for development of predictive models. The next step realized in Predictive healthcare is to develop predictive models which can utilize data from past and predict future decisions [2]. An efficiently designed predictive model helps clinicians and doctors in making informed decisions. Thus, many organizations have already adopted predictive models for tasks like providing home care to at-risk patients, detecting patient deterioration in Intensive Care Unit (ICU) or early decision making in administrative tasks. The concepts like personalised healthcare are possible with innovation of such powerful predictive applications.

1.1.1 Evolution of Computational Techniques and Technologies in Predictive Healthcare

As technology evolves, there is always a need to cope up with the data it generates. Since this data is the basis for predictive models, it requires equivalent computational techniques. Due to this, the evolution of technology and computational techniques is almost always parallel. The technological innovations for healthcare triggered around 1960s with the advent of Electronic Health Records (EHR) [4]. These records were electronic versions of patient's medical records stored on computers. This offered ad-

vantage of remote access allowing people to access and share their records at anytime and anywhere. The linking of EHR with clinical labs improved data storage as it acted as a central repository for medical history, lab tests and prescriptions. This data served as a reference and eased the interactions between patients and doctors. With the accumulation of diverse medical data like scanned images, text and handwritten prescriptions, there was a growing need of algorithms and techniques to analyze these records so as to provide predictive and quality solutions. EHRs gradually became the basis for formulation of next stage systems, Clinical Decision Support Systems (CDSS). These facilitated decision making by minimizing medical errors and providing doctors with alerts for specific situations. This became possible by utilizing the knowledge base extracted from EHRs or by applying machine learning techniques to it [5]. Using these methods, associations or if–then rules were generated to provide future solutions for specific medical decisions. In around early 1970s, extensive research was carried out for the application of artificial intelligence to CDSS so as to design expert systems. Expert systems were designed to imitate human intelligence without human interference. The system learns from the knowledge base using intelligent techniques and updates it as per need. Examples of expert systems include MYCIN [6] which was the first inference engine in medical domain for the treatment of blood borne diseases. It was gradually realized that analysis of the accumulated data using computational techniques like machine learning and artificial intelligence can improve accuracy and hence quality of diagnosis and prognosis decisions [7]. Fig.1.1 depicts the progression of technological innovations in healthcare from EHR, CDSS to Expert systems aided by evolving computational techniques like machine learning and artificial intelligence. Such transition is evident in different stages of emerging trends in medical informatics as analysed by Mihalas *et al.* [8].

To aid efficient computations, techniques and technologies such as Cloud Computing, Internet of Things and Network Analysis came later into the picture. They are currently emerging as powerful techniques for integration, visualization, storage and analysis of real-world data. They are being used for healthcare solutions which involve predictions based on complex and heterogeneous data.

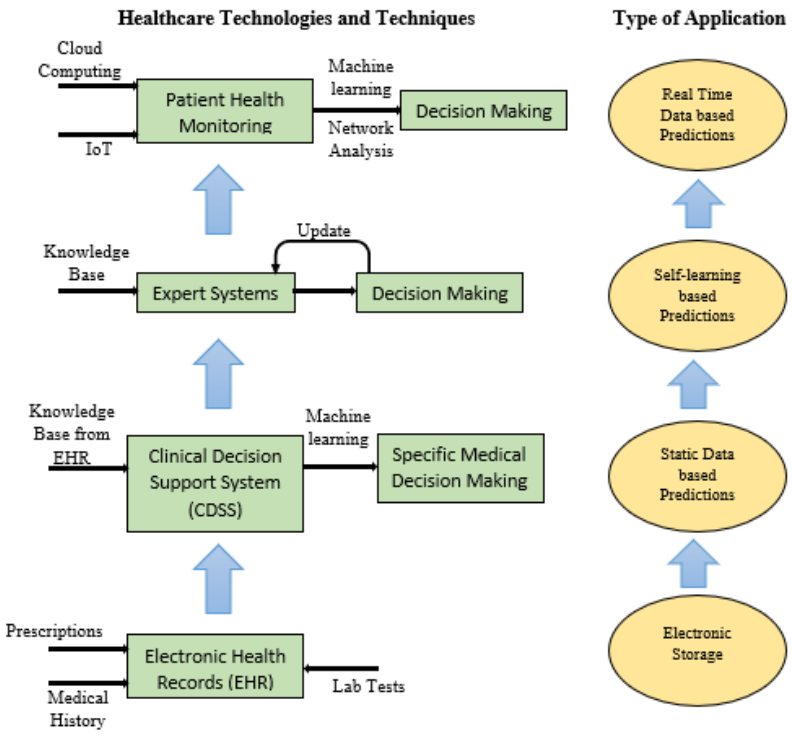


Figure 1.1: Progression in Predictive Healthcare Technologies through Computational Techniques

1.1.2 Emerging Trends for Predictive Healthcare

Predictive analytics requires interplay of multiple techniques and technologies to develop a framework for modelling. A general framework covers different phases like extraction of data, storage of data, computations and analysis, which requires numerous technologies and computational techniques as is evident in literature. Computational intelligence techniques such as Artificial Neural Network, Fuzzy methods, Support Vector Machines, Metaheuristic optimization algorithms, Ensemble learning approaches, Bayesian approaches and Markov models have been employed for diagnosis of prostate cancer as surveyed by Cosma *et al.* [9]. Various machine learning techniques devised for assessing heart failure, predicting its presence, estimating the subtype and other aspects of heart failure management have been discussed at length by Tripoliti *et al.* [10]. Different applications of machine learning technologies in healthcare field have been explored by Yoo *et al.* [11] and Fang *et al.* [12]. Herland *et al.* [13] provided a multi-dimensional view of healthcare big data analytics by exploring data at mul-

tiple scales such as population, patient, tissue and molecular levels. Similarly, W. Raghupathi *et al.* [14] outlined big data analytics in healthcare by reviewing general architecture frameworks, methodology, its advantages and challenges. Costa [15] reviewed major institutions and breakthroughs accountable for application of clinical data in personalized medicine.

It is evident from these reviews that computational techniques have experienced a steep development in healthcare perspective in a very short span of time. The techniques have evolved by incorporating and improvising underlying tasks according to medical data needs. Tasks like handling noisy data, validation techniques, feature selection and extraction have undergone significant improvements. Meanwhile, the technological improvements were taking place simultaneously. The idea of body area networks and other wearable sensors has grown gradually to monitor healthcare data and communicate it to healthcare providers. Similarly, after realising the potential of mobile technologies for healthcare, researchers have begun utilizing it for varied medical tasks. It has also been observed that the current practices for Predictive Healthcare are more inclined towards Internet of Things, Cloud technologies, Image Analysis, Big Data analytics and Mobile applications [16, 17, 18]. Within the framework of these technologies, techniques like data mining, machine learning, network analysis and text mining are being used for efficient processing of healthcare data. This section discusses the role of these techniques in predictive modelling particularly for healthcare applications. The various emerging trends are discussed as follows:

- *Cloud Computing for Predictive Healthcare*

  Cloud Computing is one such technology which is increasingly being used to accelerate the computational efficiency of predictive medical tasks [19]. It has proved to be a cost-effective solution for various medical problems like monitoring applications. The treatment of chronic diseases is otherwise quite costly and if a patient waits a long time for hospital visits, the disease might become more intense. Hence, it was realised that monitoring of health parameters in real time could facilitate early prediction. Various health parameters like heart rate, blood pressure and ECG needed to be monitored for a large number of patients

[20]. Various healthcare monitoring applications were developed to monitor and store this data which usually amounted to petabytes of data per year, requiring an effective storage and manipulation. With the adoption of wireless and body sensors, large amount of heterogeneous data is accumulated and delivered to Cloud through internet [21]. Cloud provides a pay-as-you-use model for delivering services on demand, thus becoming a good choice for a reliable and cost-effective solution.

- *Data Mining for Predictive Healthcare*

  Data mining is a computational technique which supports decision making by discovering unknown and useful patterns from massive data [22]. Data mining has been applied in numerous healthcare applications [11, 23] for example; data of important healthcare parameters like BP, heart rate and cholesterol is collected for patients in a home monitoring system so that it can be mined to predict dangerous clinical events in near future. Similarly, classification or clustering of epidemic data is done using machine learning algorithms to predict risk of reoccurrence or outbreak of diseases like diabetes, hepatitis or heart disease. Electronic Health Records are also mined in order to explore unknown patterns from population data.

- *Text Mining for Predictive Healthcare*

  Text mining, which is a branch of data mining, is also used extensively for healthcare applications. It is applied to clinical data like radiology reports which contain unstructured data. These are mined in order to extract constructive data from clinical imaging information, which helps radiologists and clinicians in decision making [24, 25]. It has also been applied to online social data for predictions to better understand the healthcare systems. For example, online reviews by patients in countries like China and U.S. were extracted from websites like RateMDs.com [26]. This data was mined to extract differences of experiences across nations. Major topics for positive and negative reviews were text mined to better understand the healthcare systems in different countries. Patients in

China focussed on bedside manners while Americans reviewed doctors. This information helps in developing a patient centric healthcare environment. Duggal *et al.* [27] used clinical data for predictions. Clinical notes and past records of diabetic patients were used to extract parameters like number of surgeries and predict the chances of their readmission in hospital using text mining.

- *Network Analysis for Predictive Healthcare*

  Medical datasets are often depicted as network visualisations due to their interdependent nature, making Network Analysis a suitable technology for their analysis [28]. Network Analysis is currently an element of various medical applications [29] for example, it is employed for predicting associations among different entities such as drugs, diseases, genes based on factors like age or symptoms. These unknown associations aid in healthcare solutions like drug repositioning and personalized medicine. It is also utilized for finding disease associated modules from gene or protein interaction networks. Thus, Network Analysis is an imperative component of computational biology. Apart from using it for biological predictions, networks have also been used to analyze hospital data so as to improve quality of medical care. Social networks realized from such resources are useful for healthcare predictions.

1.2 Predictive Healthcare for Understanding Relation between Health and Diet

Diet/food ("food" and "diet" terms have been used interchangeably in this work) is one such component which works with our body systems for optimal function. The diet we consume has a powerful impact on our immune system, brain chemistry, hormones, gene expression and microbiome. It acts as information for our body either to build a healthy individual or promote illness. Due to this, the concept of diet as a medicine is rapidly evolving. Various international projects like Personalized Nutrition Project [30] and Hundred Person Wellness Project [31] have been undertaken in order to discover

progression of a disease based on its relationship with diet. Similarly, Food4me is one such project which aims to understand relationship of genetic information and food for designing diets for individuals [32] [33]. Such projects collected data of physical and other parameters of individuals in order to explore the changes in health based on diet. Thus, understanding the intricate links between diet and health is essential but a challenging task in the computational domain. This is because of the complexities of associations between diseases and diets. For example, a particular diet might be helpful for someone who suffers from a disease, but harmful for someone who suffers from another. Diets influence the interaction of multiple diseases and their subtypes in an organism and moreover, they are dependent on one another as well. Infact different forms of a diet or different combinations of diet alter patients' health differently. As evident in a study by Bhattacharyya *et al.* [1], intricate dependencies are found in disease-diet associations, also shown in Fig.1.2. Thus, different dimensions of research can be derived from the complex relations between diseases and diet.

For useful analysis of such complex associations, construction of a comprehensive dataset is required as the first step. Data representation and inference of meaningful relations from the data are successive steps required to realize meaningful associations. Combination of right computational tools, techniques and a proper methodology can help in solving such kinds of problems efficiently. Thus, development of predictive applications is crucial for expansion in this domain. Infact, there are numerous research works evident in literature which tend to culminate predictive analytics of disease associations based on factors other than diet (including symptoms, drugs etc.). Similarly, predictive healthcare for disease-diet associations will aid extraction of new forms of knowledge which would be of great importance for different stakeholders like doctors, medical researchers, patients and dieticians. Two prominent technologies emerging in predictive healthcare namely Network Analysis and Cloud computing are best suited for storage, integration, visualization and computation of heterogeneous and distributed datasets. If endeavoured in the right manner, these technologies can prove to be an asset for this domain and thereby for the healthcare industry as well.

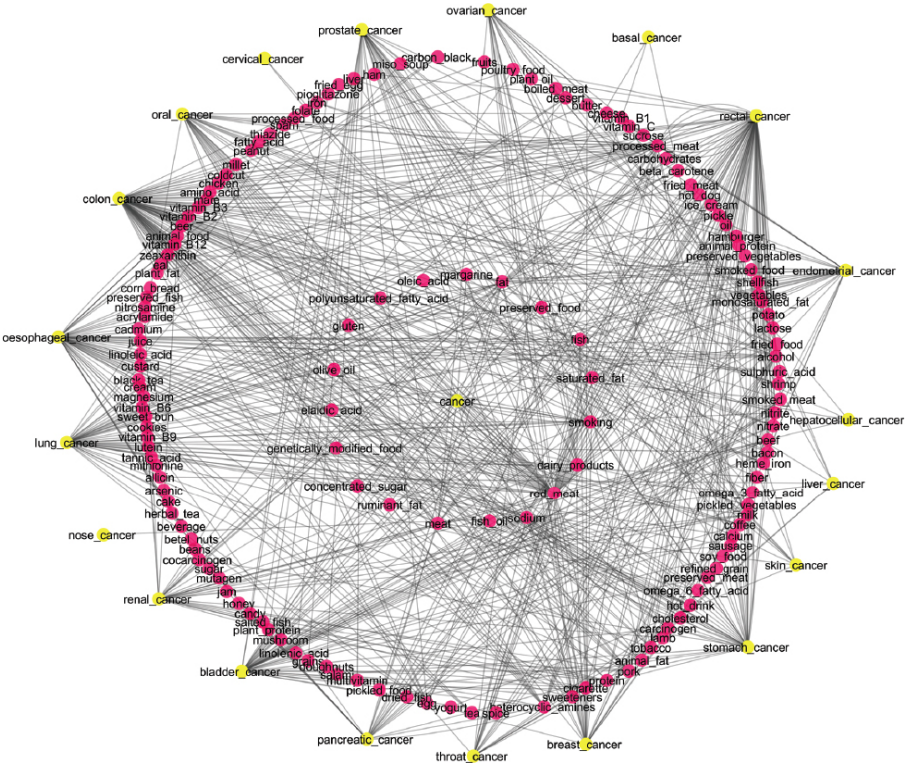


Figure 1.2: The Cancer-Food Association Network containing Foods having Harmful Associations with Cancer [1]

1.3 Network Analysis: An Overview

Networks are a ubiquitous part of real world. Different systems in nature can be depicted by a network model like food webs, biological networks of proteins or hierarchies in an organization. The examples are infinite because numerous physical and non-physical systems or processes can be portrayed as networks. Networks can be assorted into different types and corresponding definitions. C. Shi *et al.* [34] define an information network informally as: "An information network represents an abstraction of the real world, focusing on the objects and the interactions among these objects." In terms of social network analysis, environments are expressed as patterns in relationships among interacting units [35]. D. Chambers *et al.* [36] define Social Network Analysis as "It maps and measures formal and informal relationships to understand what facilitates or impedes the knowledge flows that bind interacting units, viz., who knows whom, and who shares what information and knowledge with whom by what communication media." It has been observed since the last decade that the networks

in real life are not simply random, but follow a structure in their evolution. These are termed as complex networks and are used to model many distinct real life associations [35, 37]. Complex networks came into picture with the ideas of small world and scale free networks [38, 39]. World Wide Web (WWW) is such a kind of network or collection of web pages [40]. Similarly, publications of different authors and their collaborations can be represented by co-authorship networks [41]. Social networks are also complex networks representing the links between people who might be an acquaintance, friends or family. These networks facilitate the study of different patterns prevalent in different interactions. The network structure and interactions among the nodes of a network have led to important research problems. For example, analysis of navigation data of users on the WWW is carried out so as to rank the pages to improve searching or for development of recommendation systems [42]. Co-authorship networks predict the future possibility of co-authors for collaboration [41]. Social networks help to comprehend the spread of ideas and innovations, which in turn help in prediction of future links. Network Analysis is rapidly emerging as a valuable technique for efficient analysis because it has the capability to scrutinise diverse and large datasets to solve multitude of real life problems. It enables more comprehensive study by providing features like interactive visualization and integrative data analysis of heterogeneous datasets. With the essence of vital features, Network Analysis exhibits manifold real-world applications. The various kinds of real-world networks used for Network Analysis applications are presented in Fig 1.3. This section provides a brief introduction of Network Analysis by exploring its basics and its applications in predictive healthcare domain.

1.3.1 Basic Characteristics of Networks

Networks or graphs consist of nodes which are connected by edges [43, 44]. The edges represent relationship between the nodes, for example in journal citation networks, the edge between two journals represent that an article in one of the journals cites an article in another journal. A network consisting of one particular set of actors is called as a one-mode network for example, friendships among residents of a neighbourhood. Network which contains two sets of actors is referred to as a two-mode network. In

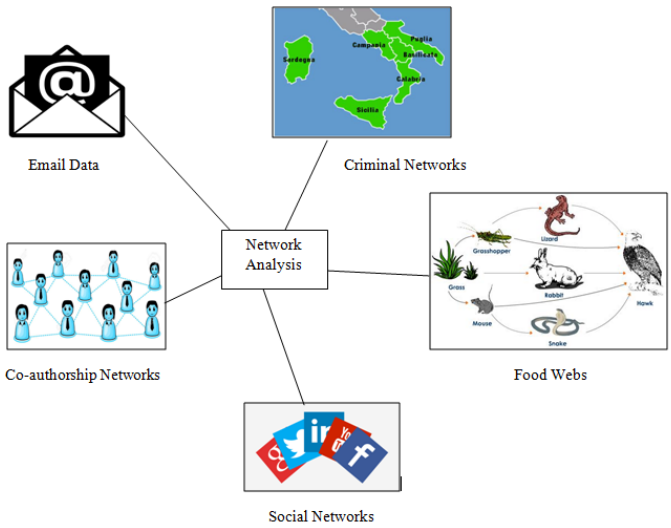


Figure 1.3: Various Networks used for Network Analysis

this, actors from one set have interactions with those in another set [43]. The major characteristics of complex networks include the following [45]:

1. Complex networks have a feature of following power law and thus are also termed as scale-free networks. In power law distribution, a very few number of nodes cover quite large number of links and vice versa.

2. The networks are called small world networks because any node in such a network has a small diameter, following the six degrees of separation principle.

3. The networks have high clustering coefficient or transitivity.

1.3.2 Topological Features

The topological features which are important for the understanding of networks are mainly of two types: Global and Local [45, 46, 47] Global features focus on the overall pattern of the network while local features focus on the neighbour's influence. There are numerous topological features of networks, some of which are described as follows:

1. *Degree Centrality* It is defined as the number of relations of a node with other nodes. It can be represented as *dc(i)* and defined as:

$$dc(i) = \sum_j e_{ij} \tag{1.1}$$

where e_{ij} represents a connection between i and j which is either 1 (link present) or 0 (link not present).

2. *Closeness Centrality* This measure defines how close a node is to other nodes. The normalized closeness centrality is represented as:

$$c(j) = \frac{(n-1)}{\sum_i d_{ij}} \tag{1.2}$$

where d_{ij} is the number of connections from node i to j and n is the total number of nodes.

3. *Betweenness Centrality* It is defined as the number of shortest paths that pass through a given node. It is represented as:

$$b(i) = \sum_{i,j} \frac{s_{i,j}(k)}{s_{i,j}} \tag{1.3}$$

where $s_{i,j}(k)$ represents number of shortest paths between i and j passing through k and $s_{i,j}$ represents number of shortest paths between i and j.

4. *Diameter* Diameter of a graph is the maximum distance needed to be traversed on the shortest path to reach from one node to another. It can be represented as $d(i, j)$ where i, j are nodes in the graph.

5. *Clustering Coefficient* It defines the degree to which neighbours of a node are connected. Clustering coefficient of node v can be represented as:

$$C(v) = \frac{2e_v}{n_v(n_v - 1)} \tag{1.4}$$

6. *Network Motifs* It is a local property of networks and can be defined as small subgraphs in a network that are recurrent and statistically significant.

7. *Clique* A clique is a complete subset of a graph which implies that its vertices are subset of the vertices in graph and it covers all the edges present among those vertices.

1.3.3 Need of Network Analysis for Predictive Healthcare

The convergence of accurate computational methods like Network Analysis and efficient tools yield predictive analytic models for healthcare data. The benefits of these models are manifold. Firstly, such analytic models facilitate personalized and preventive inferences which are the foundation for developing customized healthcare solutions. Such strategies ultimately lead to P4 medicine, which refers to medical solutions which are Preventive, Predictive, Participatory and Personalized [48]. Its focus is wellness for each individual owing to the strategies devised for predicting disease onset or its progression. Secondly, apart from the medical benefits these technology-oriented solutions are cost effective, accurate and time saving [49]. Their use has reduced cost of lab experiments and setup as well as minimized medical errors. The solutions are useful for different stakeholders like patients, pharmacists, doctors, medical companies, medical researchers and institutes. There is a rapid growth in the use of Network Analysis for predictive analytics in healthcare because of the following reasons:

- Network visualization provides an intuitive view of large real-world problems.
- Analysis of network structure and patterns is significant for comprehension of future predictions.
- It handles diverse, heterogeneous and dynamic real-time datasets effectively.
- Network Analysis merged with other techniques is vital for developing effective computational models.

1.4 Cloud Computing: An Overview

Cloud Computing offers computing resources to its customers through a remote data center on demand by means of internet [50, 51, 52]. The resources include storage infrastructure, servers, development platforms, database and networking services. In this manner, it eliminates the need to setup resource and rather rent it to be utilized on the go. Cloud Computing as defined by Buyya *et al.* in [52] is described as:
"A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers."
Provisioning of an demand model is beneficial because the customers have to pay only for the resource that have been utilized. Enterprises have benefitted drastically from Cloud Computing services as now they can quick start their business with reduced operating costs. Moreover, the resources on the platform can be extended or cut down as per requirement. This supports a flexible environment for leveraging high performance or dynamic load applications [52, 53, 54]. The various models available to cloud users are discussed in the next section.

1.4.1 Cloud Models and Services

There are several methods that can be employed for deploying Cloud application. The various deployment models are [50, 51, 55]:

- *Public Cloud:* As apparent from its name, third-party service providers manage all the resources which can be accessed by customers over the internet. For example, Microsoft Azure is a public cloud which offers storage, data analysis, application development, monitoring and many other solutions.

- *Private Cloud:* When the resources are owned by a single organization and available through a private network, it is known as a private cloud. The cloud might be located in the organization's datacenter or it might be present at a third-party service provider.

- *Hybrid Cloud:* A mix of public and private clouds is known as a hybrid cloud. Such kind of model offers the ability to switch between the two use cases as needed. Moreover, the applications can be shared between the use cases as well. This makes hybrid cloud a flexible and optimized solution.

Apart from the deployment models, most common computing services available to cloud users are as follows [50, 51, 56]:

- *Software-as-a-Service (SaaS)* When the complete application is managed and hosted using cloud services, it implies that a SaaS service model has been used. For example, a web email application is a SaaS service which is used by customers through internet irrespective of the underlying infrastructure.

- *Platform-as-a-Service (PaaS)* This model provides with the necessary hardware, software and development tools on the go so that the users need not take care of required platform, rather focus on development. For example, AWS Elastic Beanstalk is a PaaS where one can upload the code and other deployment tasks are taken care of.

- *Infrastructure-as-a-Service (IaaS)* Compute resources including servers, storage and networking are offered to users on-demand via IaaS. It facilitates scaling of infrastructure as per need, thus reducing costs and improving efficiency of solutions. AWS is one such example of IaaS.

1.4.2 Need of Cloud Collaboration

In the past decade, a significant number of enterprises have adopted Cloud models for their services. Navale [57] discussed the abundant utilities designed using Cloud Computing in biomedical science. A recent example of employing Cloud services in healthcare is a hospital information service developed for grassroot instituitions [58]. In this application, resource pooling is performed to serve multiple grassroot instituitions with dynamic resource allocation of resources as per need. This not only acts as a cost effective solution, but also enhances the performance of systems. As new grassroot

instituitions are added in the organization, the resources can be scaled up without much efforts. Similarly, another application where Cloud Computing has proved its mettle is the Cloud library [59]. With the ever increasing knowledge, many books and e-books are bought every new session. Generally, libraries manage their resources on their own and predict future load beforehand. It is also a fact that many different libraries contain the same material. A lot of space and cost will be saved if the libraries share content, which can be facilitated using Cloud Computing. This ensures seamless management, reduced costs and effortless sharing. HathiTrust Digital Library [60] is a notable example of Cloud Library which consists of academic and research material available to multiple instituitions around the globe. Deployment of such libraries also support the concept of Green library as it reduces the carbon emission. Thus, numerous advantages of being flexible, scalable, cost-cutting and adaptable have been witnessed by its users. The reasons Cloud is readily adopted in predictive healthcare are as follows:

- The adoption of Cloud Computing encourages the notion of everything as a utility, with users paying for what they use. This makes it a cost effective solution.
- Since, a cloud service provider manages the involved infrastructure, there is no hassle in deploying the service and sharing it as a deliverable with the concerned stakeholders.
- Cloud Computing assists in resource management and scaling as per need which is an appropriate solution for growing datasets. The infrastructure will be expanded as per the load requirements.
- As new data is added every other day in the medical repositories, the underlying database in predictive applications change with time. This change can simply be incorporated in the service if it is deployed over the cloud, making it up-to-date.

1.5 Problem Statement

Diet has been found to be closely related to diseases, due to which understanding their relationship is essential for a designing a healthy lifestyle. A comprehensive analysis of diet and disease data can determine disease-diet associations, which further act as a blueprint for predicting risk of disease onset. Study of such associations is a challenging task due to the ingrained interdependencies, which can be resolved with the help of predictive analytics by providing appropriate methods for representation, modelling and analysis. Visualization of disease-diet associations in the form of a network can simplify this task as well as promote learning of intricate details. This is because transformation of associations as a network offers better representation leveraging various parameters for analysis. Further, instead of using traditional measures for inferring relations, advanced and dedicated measures need to be designed for quantification of associations. Analysis of these measures can be made more accurate and robust by incorporating machine learning algorithms. There is also a need of an adaptive platform which can provide sufficient resources and makes the predictive model more accessible.

This work aims to propose and develop an efficient analysis model for apprehension of unknown disease and diet associations from diverse datasets using computational technologies including Network analysis, machine learning and Cloud computing. The model is developed as a Cloud service to be provided to stakeholders like doctors, dieticians or healthcare professionals for helping them in making informed decisions. The model derived from the combination of these technologies will help to better understand the correlation between disease progression and dietary habits and thus aid in progressive healthcare solutions.

1.6 Research Motivation

The motivation behind this work is to help the healthcare community in understanding importance of diet for a disease. The main ideology behind this research work is enumerated as:

- Health professionals and dieticians rely on their own knowledge gained through study in the current practice. This work is designed to assist the stakeholders for making informed decisions by providing them analyzed associations. In this manner, it is an attempt to improve medical or nutritional care by introducing predictive analytics to the health and diet domain.

- With newer research adding everyday to the healthcare repository, novel associations between diseases and diets are also realized. The healthcare professionals and dieticians do not research through the updated knowledge and use their traditional know-how for decision making. This leads to lack of futuristic outcomes. This work aims to provide a solution which can be tuned to include updated and modern day research, which will ultimately enhance traditional diet and healthcare solutions.

- This work has been developed not only for healthcare providers and dieticians but it is also an attempt for the research community. It tends to maximize the insight from already known information so that the predictions can be utilized for further study. Experiments and case studies can be performed on some interesting associations found out during this research.

1.7 Objectives

The objectives of the proposed work are as follows:

1. To analyse existing tools and techniques for realizing different associations of diseases based on factors like symptoms, drugs, food etc.

2. To propose and develop a network model for depicting disease-diet associations.

3. To analyse this disease-diet network on the basis of network parameters and algorithms to predict the risk of disease occurrence and progression pertaining to dietary habits.

4. To implement the proposed model over the Cloud so that the network based analysis can be developed as a service.

To achieve the first objective, study of literature has been performed to understand the current scenario of analytics for disease based associations. Various techniques have been realized of importance, but Network Analysis is found to be the most practiced as well as beneficial technology. Therefore, an in-depth survey of role of Network Analysis in predictive and healthcare applications have been undertaken. Being extensively used in exploration of disease based associations along with other techniques, a section has been devoted to depict the role of Network Analysis in this subject. Being a novel technology, there are certain challenges posed to Network Analysis for predictive healthcare which are also realized in this work along with its possible usage. Based on the extensive survey, need of Network Analysis for disease-diet associations is realized which led to the problem formulation in this work.

To attain the second objective, a network model DID-NEM is proposed and developed for depicting disease-diet associations. A network model aims to represent known associations between different diseases and diets in the form of a network/graph and further quantify their relationships using appropriate measure. This is done so that the model is readily available for advanced analysis. The most crucial requirement for modelling is a dataset consisting of disease-diet associations and a measure to quantify the associations. Since, such kind of database is not available by itself, the associations are extracted from literature to develop a comprehensive dataset. Thus, disease-diet associations have been automatically curated and quantified from medical literature using a custom-made application. Further, the curated and quantified disease-diet associations are transformed into a network to be manipulated using data integrations and analysis. This is a first of its kind model designed specifically for exploring relations between diet and disease.

To realize the third objective, the designed network model is manipulated and analyzed using data integrations, network parameters/algorithms and machine learning. A prediction approach PredNEM has been proposed to utilize these measures to unveil profound knowledge from network and imply possible disease-diet associations. For

this purpose, two different case studies have been undertaken corresponding to diseases Covid-19 and Inflammatory Bowel Disease (IBD). Network model is customized using data integrations to be suitable for each case study. Next, prediction frameworks are designed for performing analysis using amalgamation of network parameters/algorithms and machine learning algorithms. Prediction of some unique associations and their validation through a certified dietician exhibits that the proposed approach is another step in the right direction.

To accomplish the fourth objective, the proposed network based analysis is deployed on the cloud platform in order to design a scalable and adaptive disease-diet prediction service named CloudMenu. Such a deployment ensures that the service is up and running even if the load changes by scaling its compute resources. Moreover, this makes the service easily accessible to the stakeholders. Thus, a cloud based service is designed to provide robust results with a good performance using elastic compute resources from Amazon Web Services (AWS).

1.8 Thesis Contributions

This research work contributes through the following ways:

- A detailed survey of predictive analytics in healthcare specifically through Network Analysis has been performed. This emerging computational technology has revolutionized healthcare domain and its applications. Thus, this work contributes in providing a systematic review of the current status of Network Analysis in medical domain.

- One of the most important contributions of this work is development of a curation technique DIDACE for extracting disease-diet associations. The technique can be used to obtain associations between other entities apart from disease and diet. Moreover, this technique develops a robust disease-diet association database, which is also another contribution. The novel database can be utilized by other researchers for applying their own computational methods or techniques.

- Another major contribution of this work is the proposition of a network model DID-NEM. A network model is necessary for realizing complex interdependencies present within the database which are not easily observable in relational data. Thus, a network model has been proposed for visualization and modelling of complex associations.

- This work also contributes significantly by designing a prediction approach Pred-NEM incorporating a combination of machine learning and network parameters and/or algorithms for advanced analysis of disease and diet relations. The frameworks have been designed for two different case studies, demonstrating different dimensions of predictive solutions.

- Another contribution of this work lies in the deployment of the proposed model in the Cloud environment. This makes the proposed model easy to use, flexible, economical and efficient in terms of resources. The cloud service is available globally to dieticians and healthcare professionals for decision making.

1.9 Thesis Organization

After the Introduction presented in Chapter 1, rest of the thesis has been organized into following chapters:

Chapter 2: Literature Review

This chapter begins by introducing the current scenario of research in diet and health domain. It briefly describes some already designed services or case studies for exploring this domain. It further provides a detailed survey of a technology which is the core of this research work namely Network Analysis. Different dimensions have been covered including role of Network Analysis in healthcare, its challenges and proposed solutions. The discussion is then extended by incorporating another important technology Cloud Computing. It is found to have valuable contributions for predictive healthcare. Due to this, it has immense potential to act as platform for designing the proposed model as a service as discussed in the section.

- R Toor and I Chana. Network Analysis as a Computational Technique and Its Benefaction for Predictive Analysis of Healthcare Data: A Systematic Review. Archives of Computational Methods in Engineering. 2021 May;28(3):1689-711. [Impact Factor:8.171]

- R Toor and I Chana. Network Analysis of Disease-Diet Associations: A Healthcare Perspective. International Journal of Biology Today's World, Special Issue On: Role of IT in Bioinformatics and Neuroinformatics, 2017 [Scopus Indexed]

Chapter 3: DID-NEM: Proposed Network Model

This chapter discusses development of the proposed network model DID-NEM which involves curation of required database and its subsequent network construction. It firstly describes the methods, tools and techniques for extraction of already known disease-diet associations from literature. In this section, it not only covers the method for extraction of disease-diet data automatically but also suggests a methodology for automation of labelling the associations. Later, this chapter describes the avenue for construction of a network using curated database. This Chapter has been derived from:

- R, Chana I. DIDACE: Literature Mining and Exploration of Disease-Diet Associations. Journal of Information Science and Engineering. 2022. [Impact Factor:0.541]

Chapter 4: PredNEM: Proposed Prediction Approach using Network Model

This Chapter is a thorough description of the essentials required to design a prediction approach PredNEM for manipulation and analysis of the developed network model. Combination of network algorithms, network parameters and machine learning are utilized as learning methods for prediction. Thus, this Chapter discusses two different frameworks designed using learning methods for predicting diet associations in two case studies corresponding to Covid-19 and IBD. The Chapter is accompanied by discussion of its results and validations in real life. Research papers which brought this work into existence include:

- R Toor and I Chana I Exploring diet associations with Covid-19 and other diseases: a Network Analysis–based approach. Medical & biological engineering & computing. 2022 Apr;60(4):991-1013. [Impact Factor:3.079]

- R Toor and I Chana. DAPNEML: Disease-Diet Associations Prediction in a NEtwork using a Machine Learning based approach. Information Systems Frontiers. 2022 [Under Review] [Impact Factor:5.261]

Chapter 5: Deployment as a Cloud Service

In this Chapter, specifications of the developed Cloud service CloudMenu are provided. Deployment of the proposed model over Cloud is discussed. Chapter 5 has been derived from:

- R Toor and I Chana. CloudMenu: Cloud based Network Analysis for Disease-Diet Associations and Recommendations. [Under Review] [Impact Factor:3.077]

Chapter 6: Conclusions and Future Scope

This chapter concludes this research work by discussing its major contributions and further suggests probable future directions.

Chapter 2

Literature Review

The previous Chapter discussed the evolution of predictive analytics in healthcare and its significance in understanding the relation between health and diet. It further describes the need of upcoming technologies in predictive healthcare, namely Network Analysis and Cloud Computing which might prove to be beneficial for exploring associations between diseases and diets. This leads to the proposal of amalgamation of these techniques for predicting future possibilities in health and diet domain.

This Chapter is designed to review the literature, beginning with understanding the current scenario in the domain of exploring disease and diet associations. Further, the chapter traverses through an extensive survey of Network Analysis, which is a core technique utilized in this work. The survey is an attempt to comprehend the role of Network Analysis in healthcare domain, along with its posed challenges and proposed solutions. Another objective of this Chapter is to uncover the opportunities of Cloud technology for predictive healthcare.

This Chapter begins with outlining the current scenario of understanding disease-diet associations in Section 2.1. This is followed by a regressive description of Network Analysis for predictive and healthcare applications in Section 2.2 and Section 2.3 respectively. Challenges of Network Analysis and its proposed usage methods are potrayed in Section 2.4 and 2.5 respectively. Further, existing role of Cloud computing technology in healthcare solutions is described in Section 2.6 and the Chapter concludes with discussion in Section 2.7.

2.1 Disease-Diet Associations

Diseases and diets are inextricable entities closely related to one another. While some disease-diet associations have been established since ages, some other associations have evidence in literature, but are not popularly known. For example, certain diets including ginseng, black tea and ginger have been found to be helpful in prevention or management of Lung cancer as evident from NutriChem 1.0 server [61]. NutriChem is one such database containing 6242 associations between plant based foods and diseases which were extracted by text mining millions of abstracts from biomedical literature of MEDLINE [62] [63]. The database provides references of evident research studies for verifying associations between plant based foods and diseases along with the type of association observed (preventive or harmful). A preventive one refers to the one which is helpful for prevention or management of disease whereas harmful refers to the one which helps in progression of disease. There are cases existing in NutriChem database in which the type of association of a food is different even for different subtypes of a disease, for example ginseng has no evidence of relation with skin cancer whereas it has a prominent connection with lung cancer [61]. Similarly, there are diets which are preventive for a disease but harmful for its subtypes.

Table 2.1 depicts various case-control and prospective studies done to understand disease-diet relationships. Such studies are designed to explore the relationship between a specific combination of disease and diet through surveys or experiments. Similarly, there are many works (as summarized in Table 2.2) which have used advanced techniques for analysis using medical literature, but in most of the studies, disease associations have been extracted based on parameters other than diet. Although these are sound approaches, but there is a need to explore evident complex relationships of diseases in terms of diets specifically on a global level.

Table 2.1: Studies Related to Exploring Disease-Diet Associations

Authors	Year	Type of Study	Approach	Methodology	Results
Lin *et al.* [64]	2022	Cross-sectional study	To study association between dietary patterns and stages of Chronic Kidney Disease	Logistic regression used	Vitamin, minerals, cholestrol and poly-unsaturated fatty acids found to be associated
Siddiqui *et al.* [65]	2022	Prospective cohort study	To study association between diet quality and cardiometabolic health in school children	Food-frequency questionnaire and multivariable linear regression used	A better diet quality and cardiometabolic health are associated driven by lower blood pressure
Sezaki *et al.* [66]	2022	Comparative study	To study association between mediterranean diet and life expectancy	Statistical analysis using linear mixed models on international data	Found a positive association between the two
Liu *et al.* [67]	2022	Umbrella review	To study association between diet and risk of gastric cancer	Statistical analysis using systematic reviews and meta-analysis	Findings support dietary recommendations of lower intake of alcohol and salt preserved food.
Roustazadeh *et al.* [68]	2021	Case-control study	To study association between gestational diabetes mellitus and fruits and dairy products in Iran	Food-frequency questionnaire and multivariable logistic regression used	Fruits and dairy products may reduce the risk
Salomé *et al.* [69]	2021	Cross-sectional study	To study association of unprocessed/minimally processed food with cardiometabolic risk and diet quality	Nova classification and health indices	Unprocessed/minimally processed food found to be associated with better diet quality and lower cardiometabolic risk in France
Heidari *et al.* [70]	2020	Case control study	To test association between DASH diet indexes and breast cancer risk in Iran	Food-frequency questionnaire and logistic regression used	2 out of 4 indexes have been found to be associated with reduced risk

Continued on next page

Table 2.1 – *(Continued)*

Authors	Year	Type of Study	Approach	Methodology	Results
Bondonno *et al.* [71]	2019	Cohort study	To test association between dietary inflammatory index and renal disease	Dietary data using food-frequency questionnaire used	Anti-inflammatory products may reduce the risk
Arab *et al.* [72]	2019	Review	To find association between mood and diet	Newcastle-Ottawa Scale and Jadad scale used for assessment	Diets including vegetable-based, DASH, ketogenic, Paleo and glycemic load-based might improve mood
Mendonça *et al.* [73]	2019	Cohort study	To explore association between polyphenols and cardiovascular disease	Cox proportional Hazard model used with food frequency questionnaire	Higher flavonoid consumption is found to be associated with less risk
Salari-Moghaddam *et al.* [74]	2019	Cross-sectional study	To test association between pro- inflammatory diet and irritable bowel syndrome	Food-frequency questionnaire along with dietary inflammatory index used	Pro-inflammatory diet might increase risk of irritable bowel syndrome in Iranian women
Shivappa *et al.* [75]	2018	Case control study	To test association between dietary inflammatory index and prostate cancer in Argentina	Food-frequency questionnaire and logistic regression used	Anti-inflammatory products may reduce the risk
Liu *et al.* [76]	2018	Cohort study	To test association between vegetable nitrate and cardiovascular disease	Food-frequency questionnaire, vegetable nitrate database and Cox proportional hazard regression used	Vegetable nitrates reduce risk of cardiovascular disease in older Australians
Namazi *et al.* [77]	2018	Meta-analysis	To explore association between dietary index and cancer	Index calculated from food frequency questionnaire and dietary recalls	Pro-inflammatory diets found to be associated with an increased risk of cancer
Rodríguez-Martin *et al.* [78]	2017	Cross-sectional multicenter study	To test association between diet quality index and cardiovascular risk	Food-frequency questionnaire and multiple regression analysis used	Diet quality index found to be associated to cardiovascular risk and arterial stiffness

Continued on next page

Table 2.1 – *(Continued)*

Authors	Year	Type of Study	Approach	Methodology	Results
Mertens *et al.* [79]	2017	Prospective study	To test association between healthy diet and cardiovascular disease	Different diet scores along with cox regression and linear regression used	Certain dietary scores were found to be associated with reduced risk for middle aged men
Rocha *et al.* [80]	2017	Review	To understand association between cardiometabolic risk and dietary pattern	Data from tests and food frequency questionnaire used for statistical analysis	Positive association found between cardiometabolic risk factors and unhealthy diet
Martínez-González *et al.* [81]	2014	Case control study	To test association between myocardial infarction and Mediterranean diet	Dietary data using food-frequency questionnaire used	Mediterranean diet helps in reducing the risk
Bernaud *et al.* [82]	2014	Cross-sectional study	To study association between dietary fiber and inflammation in Type 1 Diabetes patients	Different parameters collected and statistical analysis performed	A consumption of greater than 30 g per day might reduce inflammation
Otaegui-Arrazola *et al.* [83]	2013	Review	Role of diet in Alzheimer's disease risk	Survey of literature done from 2000 to 2013	Some studies suggest reduced risk with nutrients like omega-3 fatty acids and Mediterranean diet whereas other studies depict inconsistent results
Rohrmann [84]	2013	Prospective cohort study	To study association between processed meat and mortality	Cox proportional hazards regression used	Positive association observed between processed meat and mortality due to cardiovascular disease or cancer
Sun and Empie [85]	2007	Review	To understand lifestyle and obesity	National Health and Nutrition Examination surveys used	High fat consumption and different lifestyle factors associated with obesity

Table 2.2: Techniques Related to Identification of Disease based Associations

Authors	Year	Approach	Technique	Database	Association	Methodology	Limitations
Jiang *et al.* [86]	2022	Machine Learning	Random Forest Algorithm	Gold Standard Databases	Disease-Drug	Network Embedding	Needs to learn features again when new nodes are introduced
Ji *et al.* [87]	2020	Machine Learning	Random Forest Algorithm	Gold Standard Databases	Disease-MiRNA	Network Embedding	Performance can be improved
Yao *et al.* [88]	2019	Machine Learning	Random Forest Algorithm	Integration with HMDD	Disease-MiRNA	Similarity measures	Performance can be improved
Bhasuran and Natarajan [89]	2018	Machine Learning	Ensemble Support Vector Machine	Gold Standard Database	Disease- Gene	Semantic analysis	Not good for complex sentences
Gutiérrez-Sacristán [90]	2017	Text Mining	BeFree text mining tool	Medline	Disease-Gene	Curation and analysis	Not completely automatic, experts required
Khordad and Mercer [91]	2017	Machine Learning	Maximum Entropy Classifier	PubMed, Phenominer and their own previously developed database	Genotype-Phenotype	Semi-automated approach for self-training	Better performance and enlarged training set needed
Haslam and Perez-Breva [92]	2017	Co-occurrence based	Manual curation	Clinical trial, MeSH	Disease-Drug	Developed an improved MeSH vocabulary	Curation is cumbersome
Ma *et al.* [93]	2017	Co-occurrence based	Manual curation	Curated using literature search	Disease-Microbe	Annotated associations and other descriptions	Curation is cumbersome
Wang *et al.* [94]	2017	Semantic Analysis	Designed pipeline based approach	Medline	Disease- Treatment	Automated generation of disease based concepts	Single disease vocabulary used for semantic schema

Continued on next page

Table 2.2 – *(Continued)*

Authors	Year	Approach	Technique	Database	Association	Methodology	Limitation
Kim *et al.* [95]	2017	Semantic Analysis	DigSee tool	Medline	Disease- Gene	Tool based text mining	Specifically designed for disease and genes association extraction
Jang *et al.* [96]	2016	Machine Learning	Shallow Linguistic Kernel	Clinical,PubMed	Disease-Drug	Curation combined with association significance filters the data, then semantic analysis further refines it	Data was collected from 1950-2011, updated data required
Singhal *et al. et al.* [97]	2016	Machine Learning	Decision Tree	PubMed	Disease- Mutation	Machine learning to identify associations using manually curated data	Redundancy in database, More robust approach required
Lowe [98]	2016	Network Analysis	Random Walk	Wikipedia, DO, MeSH	Chemical- Disease	LeadMine tool used for text mining	Better performance can be achieved
Huang and Lu *et al.* [99]	2016	Semantic Analysis	Latent Semantic Analysis	PubMed	Chemical-Chemical and Chemical-Disease	Analysed semantic patterns	Performance can be further improved
Jiang [100]	2015	Network Analysis	Random Walk	OMIM records, Ensembl	Disease- Gene	BioMart tool for querying integrated data	-
Hoehndorf *et al.* [101]	2015	Semantic Analysis	Scoring and Ranking	DO, HPO, Medline	Disease- Phenotype	Aber-OWL repository used for semantic mining of medical ontologies	Designed for pre-defined medical ontologies
Sun *et al.* [102]	2014	Co-occurrence based	Similarity scores	BioGrid, OMIM, gene ontology, HuGene	Disease-Disease	Integrated multiple databases	Used already developed databases
Li and Agarwal [103]	2009	Co-occurrence based	Statistical	MeSH, Pub- Med	Disease-Gene	Co-occurrence based data curation	Manual curation

Numerous kinds of services are available for recommending diet in healthcare domain as reviewed by Trang Tran *et al.* [104] and Marshall *et al.* [105]. Table 2.3 discusses some most recent diet recommender services designed to be exploited by dieticians or health enthusiasts. It is observed that while some of the services recommend diets based on health conditions, most of them focus on a specific disease. Only a handful of research works by Toledo *et al.* [106], Agapito *et al.* [107], Iwendi *et al.* [108] and Ivaşscu *et al.* [109] focus on different health scenarios by considering disease related parameters like Body Mass Index (BMI), calorie intake etc. The prime emphasis of these works is based on chronic diseases like diabetes and obesity and not outlining a global scenario. One of the works by Ueta *et al.* [110] has attempted to include a global view by recommending recipes based on any specific health condition. This work utilizes co-occurrences of nutrients with health conditions in literature for recommendation. Considering only the co-occurrences evident in literature might lead to presence of publication biases. Inclusion of advanced analysis techniques in such databases would lead to reliable results.

The only study by Bhattacharyya *et al.* [111] which has carried out analysis of disease and diet associations focuses on developing their interaction networks. The networks have been statistically analyzed using network parameters to realize significant diets, disease complexity and similar diseases.

2.2 Network Analysis for Predictive Applications

Network Analysis has recently emerged as a computational technique. There are very few surveys presenting its usage. In one of the surveys, Aggarwal *et al.* [112] discussed Network Analysis in the context of evolutionary networks. They focused on the evolution analysis of dynamic graphs. It also summarizes evolutionary network analysis in different application domains. In another survey, Shi *et al.* [34] discussed Network Analysis for heterogeneous data. They presented different datasets and data mining techniques for analysing heterogeneous networks. A scoping review has been conducted by Chambers *et al.* [35] to evaluate use of Social Network Analysis in

Table 2.3: Some Recent Diet Recommender Services

Authors	Year	Application	Approach	Concerned with a disease
Starke *et al.* [113]	2021	To recommend recipes for managing cholesterol	A novel metric named "Cholesterol factor" based on guidelines by the Norwegian Directorate of Health introduced to be used in Collaborative Filtering	Yes
Shrimal *et al.* [114]	2021	To provide recommendations based on required calorie intake and other choices	Users' background, BMI and preferences information collected and diets are suggested using Collaborative Filtering and fuzzy logic	No
Mckensy-Sambola *et al.* [115]	2021	To recommend suitable diet for obese patients	BMI information collected and diet recommended based on ontology derived from reading medical literature	Yes
Toledo *et al.* [106]	2021	To suggest daily meal plans according to user characteristics	Meals suggested using AHPSort for decision analysis	Yes
Iwendi *et al.* [108]	2020	To provide diet suggestions to different patients	Machine learning and deep learning models used to predict suitable diet	Yes
Musto *et al.* [116]	2019	To suggest recipes based on users BMI and other information	Rules are generated and suggestions are given based on user profile	No
Ivaşcu *et al.* [109]	2018	To recommend suitable diet for individuals	Use of ontology and rule engine based reasoning done for suggestions	Yes
Agapito *et al.* [107]	2018	To provide diet suggestions for different chronic conditions	Questionnaires are filled and Calabrian diet suggested	Yes
Yusof *et al.* [117]	2017	To provide diet suggestions to patients having diabetes	Use of ontology and case-based reasoning done for suggestions	Yes
Chen *et al.* [110]	2011	To recommend recipes for any health condition	Co-occurrences of nutrients with health conditions in literature used for recommendation	Yes

healthcare. For this, networks were constructed using hospital data and influential doctors were identified from the survey. This helped in speeding up ordering of drugs in advance by specifying those that the doctors frequently specify. The subsequent sub sections provide a brief description of the major techniques of Network Analysis. It also provides a brief discussion of collaborations comprising of Network Analysis and other computational techniques and technologies useful for predictive applications.

2.2.1 Network Analysis Techniques

Various analysis techniques have been constructed to extrapolate relevant data from networks. The major techniques identified from literature can be assembled into four categories namely Link prediction, Community detection, Ranking and Subgraph de-

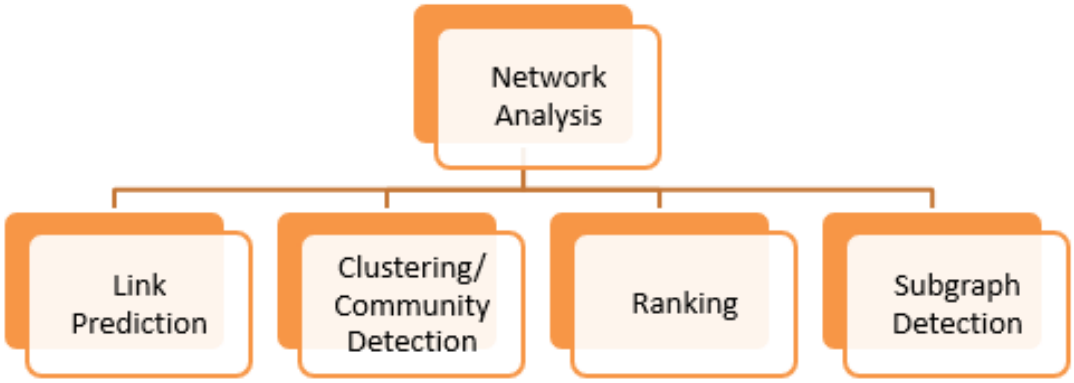


Figure 2.1: Major Network Analysis Techniques

tection as shown in Fig. 2.1 These algorithms have been used in diverse networks such as food webs, co-authorship graphs and social media to extract valuable knowledge with the help of their properties and structure.

- *Link Prediction:* Link prediction has been used in various networks to predict links which might occur in future or links which are not known yet [118]. Link prediction dates back to year 2000 with Markov Chain as its oldest technique. Applications of link prediction incorporate prediction of links in biological networks and saving a lot of time and cost, which would have otherwise incurred in performing the lab experiments. Link prediction is used for creating recommendation systems helpful for e-commerce websites too. It is also employed for predicting participation of actors in events like email or co-authorship. Link prediction provides the possibility of association between nodes in a graph even when new edges add up with time (also called dynamic graphs). Hence, it is valuable for predictions in online social networks [119]. In such networks, individuals in a group become vertices and their associations represent edges.

- *Community Detection or Clustering:* Graph clustering refers to vertices grouped into clusters such that the number of links in a cluster exceed greatly than the number of links between clusters. Various algorithms and similarity measures have been devised to be used for grouping similar or well-connected vertices together [120]. Community detection techniques have been used in literature for

various purposes like predicting future co-authorships, exploring criminal organizations' structure or for developing recommendation systems.

- *Ranking:* Ranking is referred to as categorization of objects in a network based on their similarity, influence, importance or distance [121]. The rank of an object is influenced by other objects in the network according to their proximity. Numerous algorithms and methods have been devised to rank objects like page rank algorithm, HITS algorithm or ranking based on similarity, centrality and prestige.

- *Subgraph Detection:* Dense components occurring frequently in a graph are beneficial to understand vital parts of a complex network. Finding these dense components or frequent subgraphs is called as subgraph detection. Subgraph discovery is classified into different categories derived from nature of input, search strategy and completeness of output [122].

2.2.2 Network Collaborations

It is essential to construct a framework consisting of interactive technologies and techniques for developing predictive models. A general framework entails technologies for data collection, storage, transfer and analysis. For designing a framework by means of Network Analysis, it requires other underpinning technologies like Cloud, Machine learning, Text mining and Internet of Things (IoT). Traditional Network Analysis software cannot handle graphs containing more than 1000 vertices resulting in poor performance and thus becoming a major hindrance in complex computation and analysis. Cloud Computing is rapidly becoming a preference for processing in such networks, like similarity between 4219 diseases was measured and 74,888,051 edges were obtained by Zhou *et al.* [123]. There have been very few recent works regarding the combination of IoT and Network Analysis. The union of social networks and IoT is also named as Social Internet of Things (SIoT) [124]. It can be applied in various real-world scenarios so as to benefit people through analysis of their networks formed via IoT devices as done by Guo *et al.* [125]. The combination of network analysis and text mining is used for

various healthcare predictions. Soliman *et al.* [126] constructed a glaucoma database and utilised text mining to mine genes and their associations. A network was generated from this and unknown gene interactions were predicted from its analysis. Similarly, Vyas *et al.* [127], generated a Protein–Protein Interaction (PPI) network from text mined proteins and its analysis assists in finding top most important proteins. Data mining techniques have been extensively employed in networks of biomedical data for retrieving useful associations between diseases and other factors. The related literature has been surveyed in the next section.

2.3 Network Analysis for Healthcare Applications

The complex and heterogeneous nature of medical data demands robust techniques so as to process it for inferring future predictions. Network Analysis facilitates comprehension of relationships between different datasets and role of individual nodes in the overall structure as seen in previous sections. Due to these reasons, Network Analysis is being increasingly used as an effective computational technique for healthcare predictions also. This section provides an outline beginning with the evolution of computational techniques and technologies for predictive analytics in healthcare, and then discussing the current role of Network Analysis as a computational technique for healthcare predictions. It was realized while searching for Network Analysis that it has a major contribution in literature for finding disease associations based on different factors like microbes, symptoms etc. This section also covers the role of Network Analysis for predicting disease associations based on different factors.

2.3.1 Applications of Network Analysis for Healthcare

Networks Analysis has been employed in various healthcare related predictive applications or research work. Network medicine, as discussed by Barabasi *et al.* [28] can be viewed in three layers ordered from molecular to social level. Fig.2.2 depicts the different network layers in healthcare and corresponding databases used in such applications.

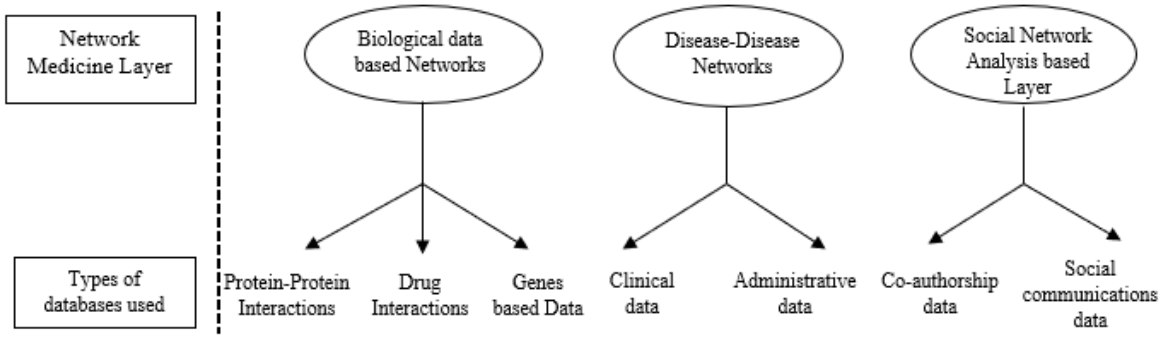


Figure 2.2: Various Network Layers in Predictive Healthcare

- *Biological Networks:* A major portion of research has been dedicated to Network Analysis at the molecular level. There are various biological networks which have been used for analysis at the molecular level as discussed by Carter *et al.* [128], Goh *et al.* [129] and Sah *et al.* [130]. Such networks aim to derive inferences from associations between biologically related nodes. Examples include PPI networks, genetic interactions, co-expression data *etc.* The different biological themes covered are as follows:

  (i) PPI networks have been extensively studied for exploring disease associations. There is one such example in which the interactome has been studied by Menche *et al.* [131] to understand relationships between diseases. Similarly, interactome and biological networks are analysed by Stevens *et al.* [132] for studying the pathways of endocrine diseases so as to relate it with disease progression. Relationship between diseases has been quantified based on the distance of their proteins and then compared with other datasets. PPI have also been investigated at different levels using network properties by Liu *et al.* [133], Alanis *et al.* [134], Prvzulj *et al.* [135]. Clustering a network of protein interactions has been improved by Hasan *et al.* [136] which helps in identifying functional modules.

  (ii) Many research works focus on exploring the drug target interactions so that it can be used for drug repurposing or drug discovery as done by Ibrahim *et al.* [137]. Drug-target interactions, disease genes and molecular pathway were used to identify drugs effects on disease genes in [138]. Similarly, new targets for drugs were predicted using a novel inference method from drug-target network by Cheng

et al. [139]. Relationship between cell lines and drugs were used to predict drug response for samples by Stanfield [140]. In this work, combination of molecular and network data has been depicted as a link prediction problem in the network and evaluated with 86accuracy. Supervised prediction method has been used in bipartite graphs to predict drug target links [141]. Drug target links were predicted using network substructures and scores in [142]. Many different network methods were explored for inferring drug targets like random walk with restart, Monte Carlo simulated annealing and similarity based inferences by Chen *et al.* [143]. Drug target predictions were performed through kernel based methods which also used network based similarity by Nascimento [144]. In another work by Fu *et al.* [145], a semantic network was generated by integrating biomedical data and topological features were derived using metapaths. Link prediction and ranking were performed using Random Forest algorithm to predict drug target interactions.

(*iii*) Genetic data has proved to be an aid for understanding disease and gene associations inferred using multiple network scenarios. Biological networks combined with gene expression data has been used to explore disease and gene associations by Ernst *et al.* [146]. Linghu *et al.* [147] constructed an association network which consisted genes along with their functional associations. Context sensitive networks generated using relationships between disease and phenotypes were integrated with genetic network to infer disease gene association using ranking by Chen *et al.* [148]. Similarly, biological pathways have been used to infer disease associations by Li *et al.* [103]. A network based approach has been utilized for prioritizing genes using differential gene expression data by Nitsch *et al.* [149]. They used machine learning kernel based techniques for analyzing network. Similarly, network topology has been explored using Network Propagation (NP), Random Walk with Restart (RWR) and Shortest Path (SP) algorithms to prioritize genes for diseases by Razaghi *et al.* [150]. Random walk approach has also been used by Jiang *et al.* [100]. Gene phenotype prediction was performed using network weights as features for Support Vector Machine (SVM) classifier by Guan *et al.*

[151]. The network was generated by integrating multiple biological databases and weights were defined using bayesian inference. These predictions might help in learning disease mechanisms. Eronen and Toivonen [152] proposed a system Biomine, which integrates biomedical data from various sources and used different network measures for predictions. Random walk with restart, supervised or unsupervised prediction methods have been used for disease-gene prioritization. Leiserson *et al.* [153] describe different algorithms and network tools to identify cause of a particular disease. This is possible by analysing the genes or proteins interaction networks so as to identify genetic variations which ultimately lead to a disease. Similarly, Ferrazzi *et al.* [154] studied heart disease development from gene network analysis.

- *Disease Networks:* The other layer evident in research is Disease network layer. The associations between diseases have been studied in literature to understand disease progression. There is one such example in which insurance claims data has been used by Kim *et al.* [155] to identify disease–disease associations. Clinical drug trial information has been used by Haslam *et al.* [92] to develop disease relationship network. Administrative healthcare data was obtained and trajectory of diseases in patients was modelled as network by Khan *et al.* [156]. Two comorbidity networks were constructed for patients with and without Type 2 diabetes, and explored to understand high risk diseases and patterns. Different similarity measures including standard methods and a proposed graphlet measure were used to predict associations between diseases based on biological network by Sun *et al.* [102]. Such visualizations and predictions are helpful in understanding diseases and aid in biological research.

- *Social Network Analysis:* Social Network Analysis is the third layer identified in research works. It has been used in literature for assessing the role of healthcare parameters with the help of human interactions [36], thus named as Social Networks. Apart from sole medical purposes, networks have allied with healthcare data for other tasks. A network is generated from medical expenditure data and

insights are derived from drug purchases by Belyi *et al.* [157]. It successfully derives drugs which are usually taken together despite having same compound. This would help in customising drug prescriptions and improving drug costs. Evans *et al.* [158], utilized multiple aspects of school, social networks and neighbourhoods of students for constructing networks to understand their role towards health. Various social network methods have been used to study co-authorship networks for health. Analysis of co-authorship graphs has been undertaken by Sampaio *et al.* [159] for identifying leading researchers in different domains of health research. Collaboration networks for influenza virus vaccine area were generated and Social Network Analysis was used to explore the research done in this field by Liu *et al.* [160]. Social network was used to understand interruptions in the role of people involved in an ICU by Mccurdie *et al.* [161]. They explored inter dependencies and factors for interventions from the network so that clinical workflow can be improved. Similarly, to improve the provision of care providers for patients suffering from life threatening disease, a social network was constructed by Moradianzadeh *et al.* [162] and optimal providers were selected with minimum cost using different network algorithms. A similar network was constructed by Steitz *et al.* [163]. Outbreak networks in hospitals were also studied by De *et al.* [164]. A network using clinical data was constructed to comprehend other co-occurring conditions for patients diagnosed with depression by Kim *et al.* [165]. They used statistical methods and network metrics to compare the networks and infer associations. Network parameters were studied by Mammone *et al.* [166] in brain networks to compare subjects which were healthy, or suffered from different stages of dementia. Social Network Analysis(SNA) was performed over research articles of nursing care by Choi *et al.* [167] to explore its role for delirium patients. Subgroups were identified and network structure was studied so as to understand the relations and improve the provision. There are many other works by Kjos *et al.* [168], Siden *et al.* [169] and Yao *et al.* [170] which have used SNA for improving medical care. Discharge of patients was planned by analysing communication patterns as networks in hospital sys-

tem by Prusaczyk *et al.* citeprusaczyk2019networks. The intricacies of networks are of great use for proposing medical solutions like predicting associations or finding disease associated modules. Table 2.4 presents some other applications of Network Analysis techniques for predictions in healthcare domain which have not been discussed above. It describes different Network Analysis techniques and their use for predictive healthcare solutions.

2.3.2 Network Analysis for Predicting Disease based Associations

Network Analysis is increasingly being used for studying disease based associations. This is possible using data mining and machine learning techniques for biological data. Network Analysis has been used extensively in combination with data mining so that associations between different healthcare parameters can be predicted. Various machine learning algorithms have been devised using features from network to predict future links based on different factors [1, 111]. Numerous works regarding this are discussed as follows:

- *Drug based Associations:* The development of new drugs for any disease is a costly and time-consuming task. Therefore, various computational techniques have been devised which utilize known disease-drug associations so as to predict possible unknown relations of drugs to diseases. This helps in repurposing of approved drugs, thereby assisting the process of drug discovery. Understanding the drug and disease relationships is important so as to gain knowledge about disease progression. The various works of drug based associations are depicted in Table 2.5. It describes the types of similarity measures, networks and databases used in different works. The Area Under Curve (AUC) parameter in the table depicts which model performs better with a value near 1 depicting better performance.

Table 2.4: Applications of Network Analysis in Healthcare

Authors	Year	Technique	Network layer	Application	Algorithm
Moscato *et al.* [171]	2022	Community Detection	Social Networks	Detect groups of users sharing same physical and dietary habits	Novel community detection algorithm
Pham *et al.* [172]	2020	Link Prediction	Biological Networks	Predicting drug–drug associations	Ensemble method using random walk, neighbor recommender method and matrix perturbation
Zhang *et al.* [173]	2017	Link Prediction	Biological Networks	Predicting drug–drug associations	Ensemble method using random walk, neighbor recommender method and matrix perturbation
Lu *et al.* [174]	2017	Link Prediction	Biological Networks	Predicting drug-target associations	Different similarity measures
Surian *et al.* [175]	2016	Community Detection	Social Networks	Detecting community structure from social data for opinion mining	Infomap and Louvain algorithm
Evans *et al.* [158]	2016	Community Detection	Social Networks	Study the effect of schools, social networks and neighborhood on Body Mass Index	Modularity Maximization and k-clique percolation method
Kaya and Poyraz [176]	2015	Link Prediction	Disease Networks	Finding disease connections based on age of patients	Age-series based link prediction method (ASLIP)
Narayanan *et al.* [177]	2014	Community Detection	Biological Networks	Functional enrichment and path analysis for PPI networks	Newtonian framework
Kaya and Poyraz *et al.* [178]	2014	Link Prediction	Disease Networks	Predicting relations between symptoms	Supervised link prediction method
Fakhraei *et al.* [179]	2014	Link Prediction	Biological Networks	Drug-target interaction prediction	Probabilistic Soft Logic

Continued on next page

Table 2.4 – *(Continued)*

Authors	**Year**	**Technique**	**Network Layer**	**Application**	**Algorithm**
Petrochilos *et al.* [180]	2013	Community Detection	Biological Networks	Finding cancer associated modules from gene expression data	Walktrap based on random walks
Singh-Blom *et al.* [181]	2013	Link Prediction	Biological Networks	Prediction of disease-gene associations	Katz and Catapult
Lei and Ruan [182]	2012	Link Prediction	Biological Networks	Reconstructing PPI networks	Random walk with resistance for topological similarity

Table 2.5: Drug based Disease Associations

Authors	Year	Similarity Measure	Type of Network	Method for finding Associations	Database	Performance
Jarada *et al.* [183]	2021	Similarity network fusion	Heterogenous with several benchmark datasets	Feed-forward Multilayer Perceptron	Many different sources including SND, Cdataset etc.	AUC 0.86
Wang *et al.* [184]	2019	Cosine similarity, Gaussian interaction profile kernel similarity	Heterogenous integration of drug-drug, disease-drug and disease-disease datasets	Bi-random walk	Many different sources including MeSH, PubChem etc.	AUC 0.96
Tian *et al.* [185]	2018	Semantic similarity	Heterogenous integration of drug-drug, disease-drug and disease-disease datasets	Scoring based on meta paths	Many different sources including OMIM	AUC 0.89
Liu *et al.* [186]	2016	Bipartite Projection Network and Jaccard similarity	Heterogeneous integrated drug–drug, disease–disease and drug–disease	Two pass Random Walk with restart	MimMiner, SMILES	AUC 0.93381
Yu *et al.* [138]	2016	Cosine Similarity	Heterogeneous with integration of networks, side effects of drugs and symptoms of disease	Clustering using ClusterONE and correlation analysis	Comparative Taxicogenomics Database (CTD), SIDER	Precision of prediction >0.8
Mullen *et al.* [187]	2016	Semantic distance measure	Heterogeneous with gene-disease and other datasets	Ranking and semantic subgraphs	Many different sources including literature, curated and predicted databases	AUC 0.75
Luo *et al.* [188]	2016	Tanimoto coefficient and clustering using ClusterONE	Bipartite graph	Bi-random walk algorithm	Gold standard dataset	AUC 0.917
Moghadam *et al.* [189]	2016	Jaccard and cosine scores, semantic similarity	Bi-partite drug-disease network	Support Vector Machine and Kernel fusion	DrugBank, OMIM	AUC 0.91
Oh *et al.* [190]	2014	Adjacency-based, module-distance based	Heterogeneous integration of protein and gene networks	Classifier using C4.5, Multilayer Perceptron and Random Forest	Online Predicted Human Interaction Database, DrugBank, CTD	AUC 0.917

Continued on next page

Table 2.5 – *(Continued)*

Authors	**Year**	**Similarity Measure**	**Type of Network**	**Method for finding Associations**	**Database**	**Performance**
Yang *et al.* [191]	2014	Transition weights similar to markov chain	Integration of gene, drug, target, pathway disease	Probabilistic Matrix Factorization Model	MeSH, CTD, Biocarta of MSigDB	AUC >0.95
Huang *et al.* [192]	2013	Tanimoto and pearson correlation coefficient	Integration of drug, genomic, disease and phenotypic data	Label propagation using Random Walk	OMIM, DrugBank	AUC 0.94
Zhao and Li [193]	2012	Shortest network distance and finding co-modules using Com- Cipher	Integration of five protein interaction databases	Monte Carlo Markov Chain	OMIM, DrugBank	AUC 0.9

- *MicroRNA based Associations:* A group of small non protein RNA molecules are termed as microRNA. They are essential part of many biological processes including cell growth and tissue development. Hence, discovery of associations of miRNAs with diseases will benefit the understanding of disease mechanisms and progression. Various computational approaches have been proposed to prioritize miRNA candidates which are depicted in Table 2.6.

- *Microbe based Associations:* The microorganisms comprising of bacteria, viruses, fungi etc. which reside in human body are called microbes. They form a healthy relationship with host organs and are infact significant for their physiology. For example, fermentation of food components is done by gut microbiota to help in digestion by the host. Microbes' functions include developing the immune system, maintaining drug metabolism and protecting from pathogens. Hence, the study of microbes and their relation to diseases would play a significant role for gaining a better perception of disease mechanisms and therapies. The traditional approaches which involved cultivation of microbes were time consuming and laborious, so computational approaches are being employed to minimize costs and time. The various works of microbe based associations are depicted in Table 2.7.

- *Phenotype based Associations:* Phenotypes include symptoms/side effects observed in a patient. The knowledge of relation of symptoms with molecular processes can aid in understanding personalized treatment. Also, similar drugs can be used for diseases having similar symptoms or side effects. Symptoms based disease associations can help in identifying targets of infectious diseases and prioritization of genes because it has been known that genes located in the close neighbourhood of targets in the PPI network also exhibit high symptom similarity with viral infections. Networks have been used by Borsboom *et al.* [194] to identify symptom dynamics in psychopathology using symptom networks. Symptom network based on diseases has been used by the author to extract useful information like finding the central symptom in a person. The various works of symptom or side-effect based associations are depicted in Table 2.8.

Table 2.6: MicroRNA based Disease Associations

Authors	Year	Similarity Measure	Type of Network	Method for finding Associations	Database	Performance
Dai *et al.* [195]	2021	Gaussian Interaction Profile kernel similarity	Heterogenous network	Logistic weighted Bi-random walk algorithm	HMDD	AUC 0.93
Niu *et al.* [196]	2019	Gaussian Interaction Profile kernel similarity and others	Heterogenous network	Random walk and binary regression	HMDD	AUC 0.80
Xie *et al.* [197]	2019	Gaussian Interaction Profile kernel similarity and semantic similarity	Heterogenous network	Bipartite Network algorithm	HMDD	AUC 0.93
Chen *et al.* [198]	2018	Gaussian Interaction Profile kernel similarity and semantic similarity	Heterogenous with lncRNA–miRNA association and lncRNA-disease network	Label propagation	HMDD and other databases	AUC 0.93
Luo and Xiao [199]	2017	Semantic similarities using MeSH directed acyclic graphs	Heterogeneous with disease similarity, miRNA functional similarity and miRNA-disease association	Unbalanced bi-random walk algorithm	Human miRNA-disease database and MeSH	AUC 0.846
You *et al.* [200]	2017	Semantic similarities using directed acyclic graphs, Gaussian interaction profile kernel and functional similarities	Heterogeneous with disease, miRNA and miRNA-disease similarities	Special Depth first search	Human disease-miRNA Database, dbDEMC, miR2Disease	AUC 0.9172
Chen *et al.* [201]	2016	Semantic similarities using directed acyclic graphs, miRNA functional similarities and Gaussian interaction profile Kernel	Heterogeneous with miRNA, disease and miRNA-disease similarities	Iterative algorithm	Human disease-miRNA Database	AUC 0.8781, 0.8077
Gu *et al.* [202]	2016	Different similarity scores were combined to develop miRNA-disease score	Heterogeneous with disease– disease similarity network, miRNA–miRNA similarity	Space projection scores	Predictive and benchmark datasets were used	AUC 0.9173

Continued on next page

Table 2.6 – *(Continued)*

Authors	Year	Similarity Measure	Type of Network	Method for finding Associations	Database	Performance
Xuan *et al.* [203]	2015	Semantic similarities using directed acyclic graphs	Heterogeneous with disease, miRNA and miRNA-disease similarities, bilayer network	Random Walk algorithm	Human miRNA-disease database	AUC ranges from 0.786 to 0.945
Chen *et al.* [204]	2015	Known miRNA-disease similarities	Homogenous with miRNAdisease associations	Restricted Boltzmann machine	Human miRNA-disease Database including miRNA-target interactions, circulation, epigenetics and genetics	AUC 0.8606
Chen and Yan [205]	2014	Semantic similarities using directed acyclic graphs and miRNA functional similarities	Heterogeneous with disease, miRNA and miRNA-disease similarities	Regularized least squares	Human disease-miRNA Database	AUC 0.8450 and 0.9511
Xuan *et al.* [206]	2013	Improved Wang's measurement for Semantic similarities using directed acyclic graphs	Homogenous with miRNAdisease associations	Weighted k most similar neighbours	Human disease-miRNA Database	AUC 0.825
Jiang *et al.* [207]	2010	Fisher's exact Test	Heterogenous with miRNA and phenome network	Length of the shortest path and hypergeometric distribution	miR2Disease and HMDD	AUC 0.758

Table 2.7: Microbe based Disease Associations

Authors	Year	Similarity Measure	Type of Network	Method for finding Associations	Database	Performance
Long *et al.* [208]	2019	Gaussian Interaction profile kernel and semantic similarities	Heterogeneous microbe–microbe, microbe–disease, disease–disease network	Meta-graph search algorithm	Human Microbe-Disease Association Dataset (HMDAD)	AUC 0.928
Huang *et al.* [209]	2017	Gaussian Interaction profile kernel	Heterogeneous microbe–microbe, microbe–disease, disease–disease network	Special depth first search algorithm with exponential decay function	HMDAD	AUC 0.9169
Shen *et al.* [210]	2016	Spearman Correlation	Heterogeneous integrated disease and microbe network	RWR	HMDAD	Higher specifity and sensitivity than for only random walk algorithm
Chen *et al.* [211]	2016	Community Detection	Social Networks	Study the effect of schools, social networks and neighborhood on Body Mass Index	HMDAD	AUC 0.8382
Roy and Filkov [212]	2009	Cosine similarity	Homogeneous with disease similarities	Frequency-inverse document frequency	HMDAD	P value for overlap of microbe-symptoms 0.4

There are numerous different methods utilized for finding disease based associations as evident through the survey. It is also realised that majority of associations have been found using random walk and markov chain methods. The next popular technique has been the use of traditional similarity measures. Few research works have focused on machine learning techniques of clustering, classification and semi-supervised learning. Other methods included depth-first search, ranking and probabilistic methods. The distribution of techniques has been depicted in Fig. 2.3. For extracting associations of diseases with different factors, different datasets needed to be used. Table 2.9 describes the various existing datasets which have been used in literature for extracting associations as evident in the research papers.

2.3.3 Dynamic Networks

Dynamic Networks are used to comprehend flow in a network as well as changes in topology over time as indicated by the name itself [214]. Exploring the network dynamics is an interesting and upcoming field. It is increasingly being used in biological networks to study disease progression. In one of the works by Hidalgo *et al.* [215], due

Table 2.8: Phenotype based Disease Associations

Authors	Year	Similarity Measure	Type of Network	Method for finding Associations	Database	Performance
Xu and Wang [213]	2017	Gaussian Interaction profile kernel	Heterogeneous microbe–microbe, microbe–disease, disease–disease network	Special depth first search algorithm with exponential decay function	HMDAD	AUC 0.9169
Hoehndorf *et al.* [101]	2016	Spearman Correlation	Heterogeneous integrated disease and microbe network	RWR	HMDAD	Higher specifity and sensitivity than for only random walk algorithm
Zhou *et al.* [123]	2009	Cosine similarity	Homogeneous with disease similarities	Frequency-inverse document frequency	HMDAD	P value for overlap of microbe-symptoms 0.4

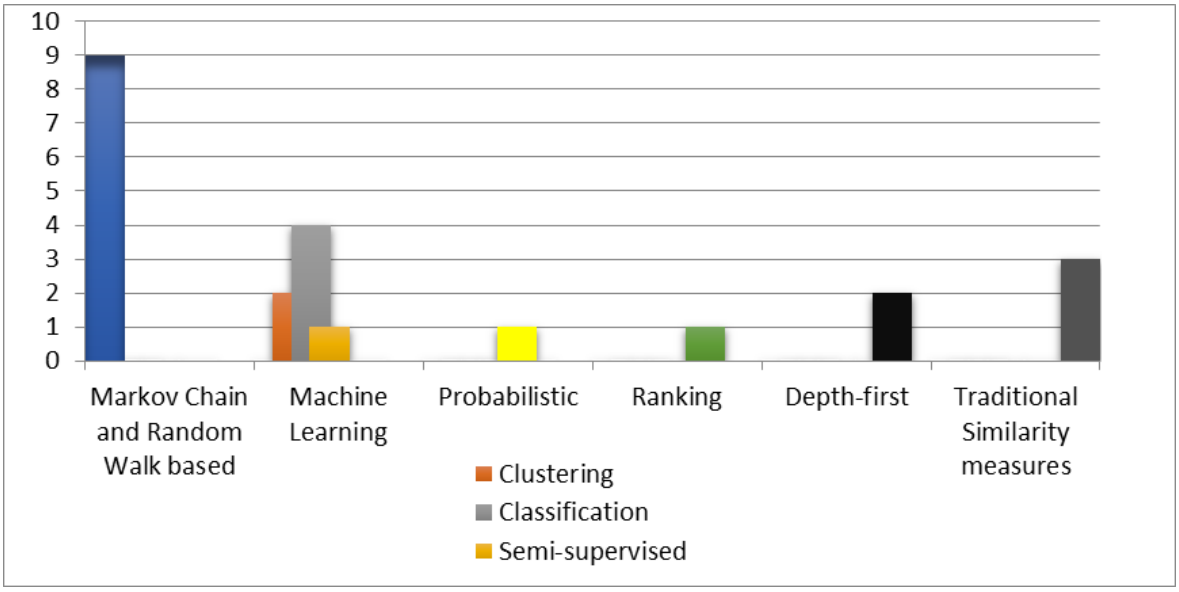


Figure 2.3: Distribution of Methods used for Extracting Association

Table 2.9: Existing Datasets

Database	URL	Purpose
KEGG disease	http://www.genome.jp/kegg/disease/	Disease entries with their drug and genes associations
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Obtaining biomedical literature citations like from MEDLINE
MeSH	https://www.ncbi.nlm.nih.gov/mesh	Vocabulary thesaurus for indexing PubMed articles
ICD10	http://apps.who.int/classifications/icd10/browse/2010/en	Diseases and related problems' classification list by World Health Organization (WHO)
Healing Food Reference	http://www.healingfoodreference.com/	Provides helpful foods for diseases
CTD	http://ctdbase.org/	Curated relations between genes, diseases, chemicals, proteins
NutriChem	http://www.cbs.dtu.dk/services/NutriChem/	Plant-disease associations

Table 2.10: Healthcare Analytics using Dynamic Networks

Authors	Year	Application	Technique	Tool/Dataset
Carroll [217]	2020	Early diagonosis of Alzheimer's disease by studying functional connectivity networks	Network similarity and machine learning algorithms used	ADNI
Carroll [217]	2019	Study hospital care networks	Network properties studied	SNA
Sood *et al.* [216]	2018	Identify risk areas for mosquito borne diseases	Temporal properties used for prediction	Gephi tool
Merrill *et al.* [218]	2015	Study of service deliveries	Transition probabilities and other novel techniques	ORA
Effken *et al.* [219]	2011	Study communication in patient care units	Network metrics studied	ORA
Hidalgo *et al.* [215]	2009	Understand disease progression	Correlations between different visits	Medicare claim data

to lack of data, a static network has been studied dynamically to understand disease progression. The correlation of diagnosis of diseases in first two visits is compared to that in next two visits, in order to understand the dynamics of the network. Comorbidity scores were calculated for quantifying the distance between diseases in network, and further used to compare patients of different ethnicities and gender. Temporal networks have been studied by Sood *et al.* [216] to detect the risk areas for spread of mosquito borne diseases. The network was created in Gephi at different time intervals and used to explore temporal properties. Six metrics including temporal correlation coefficient, path hops and betweenness centrality were calculated using network properties to predict the transmission of disease. Another work by Carroll *et al.* [217] discusses the importance of SNA for studying hospital care networks which originate from real time and continuously evolving data. Similarly, transition probabilities were evaluated by Merrill *et al.* [218] to explore transition networks of service deliveries so as to improve medical care. Instead of SNA, Dynamic Network Analysis (DNA) has been used by Effken *et al.* [219] to analyse communication networks in patient care units which change over time. Various network metrics were obtained using Organization Risk Analyzer (ORA) tool to understand relationship between communications and patient safety. Table 2.10 describes various works of healthcare analytics using dynamic networks.

It is realized from the survey that most of the medical applications capture real time data, which is ever evolving. Thus, it requires a dynamic approach to understand the progression in networks. Dynamic Network Analysis is a promising technique to study transitions in real time. Temporal properties and transition probabilities can be used to devise efficient algorithms for the study. Such approaches are still in their infancy but if designed properly, these will prove to be beneficial for developing predictive healthcare solutions.

2.4 Challenges in Application of Network Analysis for Healthcare Predictions

Although Network Analysis and its techniques have evolved since the 20th century, yet there are many challenges as evident from literature. This section presents the challenges in the application of Network Analysis for healthcare predictions. Some of the challenges and their possible solutions analyzed in this work are:

- Some techniques do not perform well with large or sparse medical graphs. Moreover, with the graphs growing gradually with time [216], there is a need to store the data effectively. One of the solutions can be the provision of a Cloud based framework. Cloud platform provides resources on the fly as per requirement which can be scaled up or down. This is a cost effective and time saving solution for large, heterogeneous and dynamic healthcare datasets. Network Analysis is already being used in conjunction with Cloud and IoT for monitoring healthcare applications for efficient computations in real time.

- One of the challenges arising is the failure to focus on dynamic networks [215], although most of the complex healthcare networks are dynamic in nature. For example, protein– protein or other biological networks of individuals can facilitate personalized medical solutions by understanding the patterns in their evolution over time which requires expertise in dynamic networks. Several novel techniques are being devised to handle such transient nature of networks. Link prediction along with machine learning techniques for temporal data would be a great solution in such scenarios.

- It has been observed that many Network Analysis techniques have been based on a particular case study or medical dataset [206]. This is another drawback because for efficient functioning of a technique, it needs to be validated for several case studies or for a comprehensive dataset. In case of heterogeneous dataset, a suitable Network Analysis technique must be opted for manipulation because networks are fit for interpreting integrative data. A combination of Network

Analysis technique with big data analysis would serve the purpose.

- In the domain of Network Analysis for extracting unknown associations, the major challenges are regarding the similarity measures and mining algorithms essential for analysis [174]. Although, there are numerous traditional similarity measures, but there is a strong need of dedicated similarity measures so that expedient networks can be constructed. The application of suitable machine learning algorithms over these networks will aid in automation of task of exploring unknown associations. Use of appropriate measures and algorithms will enrich the network based frameworks leading to better healthcare decisions for doctors.The global and local topological parameters of networks like betweenness, centrality etc. can be used as features to further apply machine learning algorithms so as to extract unknown associations or perform clustering.

- Although there have been advances in the other techniques and technologies used in conjunction with Network Analysis, yet their application is limited because they are still evolving. Moreover, use of these techniques specifically for healthcare applications is in its infancy. There is a need to comprehend suitable combinations of techniques and technologies with Network Analysis in order to be able to perform valuable analysis and computations.

- Real world complex networks not only consist of multiple kinds of nodes, but also multiple types of links. Bipartite networks are commonly observed in drug-target or recommender applications where two types of nodes are present. A new framework has been devised for predicting connections in such networks in [220] by developing local community based topological model for bipartite graphs. The research community should use and encourage such customized frameworks.

2.5 Proposed Usage of Network Analysis Techniques

Being a vast and novel area of research, Network Analysis can be utilized in multiple ways to extract significant disease and diet associations. The major computational approaches of Network Analysis which can be used for various diet related inferences are as follows:

- *Overlapping of Networks for prediction:* Network overlapping and comparisons are done to realize unknown associations. Two networks can be overlapped if they have same types of nodes. For example, two networks having same diseases were developed by Zhou *et al.* [123] where one was based on similar disease-symptoms and the other on similar disease-genes. The two graphs were integrated to create a single network of shared diseases and genes. The network parameters and analysis of this graph was useful for interpreting disease correlations and predict unknown associations. Thus, two networks of diseases, in which one is based on diets and other is based on a factor like symptoms, drugs or genes can also be overlapped to extract significant disease associations.

- *Dynamic graphs for understanding temporal relations:* Apart from overlapping, a network can be compared to itself as it evolves with time so that the progression in pattern can be observed and future pattern can be forecasted [217, 221]. There are numerous meta-analysis done by Aune *et al.* [222], Schwingshack *et al.* [223] and Ding *et al.* [224] in which consumption data of a specific food item by patients suffering from a disease is collected. Similarly, dietary patterns of patients can be noted for a period and represent it in the form of temporal graphs. The graphs can be used to study dynamic trends for understanding disease-diet interactions.

- *Ranking and Clustering for exploring similarity:* Ranking is exploring the most significant links from networks and order them accordingly [225]. It is used for fixing the priority of links according to their rank. It also helps to simplify a dataset by identifying vital links. Ranking for diseases has been done previously

based on their degree of association in the network by Bhattacharyya *et al.* [111]. Apart from using only the network parameters for ranking, the network can also be broken down into clusters and then ranking can be performed, so that the topmost significant associations are recognized. The inclusion of clustering will enhance the accuracy of ranking. The disease-diet network can be split into clusters and then ranking can be performed for each cluster. In this way, similar diets and disease links can be ranked.

- *Machine learning and Link prediction for effective analysis:* A machine learning framework can enhance traditional link prediction of networks. The features for this task can be extracted using various properties of a network and can be analyzed using machine learning techniques. The approaches like Clustering [226], Classification [189, 190, 227, 228] and Ranking [228, 229] can be applied to the database to predict future links. A framework generated using machine learning and link prediction as shown in Figure 2.4 can be used in forecasting the possibility of association of disease with another disease in a curated disease-disease network. Network parameters like SimRank (SR), Common Neighbours (CN), Adamic Adar (AA), Clustering Coefficient (CC), Diameter (D) and Betweenness Centrality(BC) [230, 231] can be extracted for different diseases like Diabetes Type 2, Diabetes Type 1, Lung cancer and Breast cancer and used as features as shown in the figure. Different machine learning approaches can then be applied to the retrieved dataset for inferring unknown associations.

- *Personalized Analysis:* Work done in literature has focused on generic predictions for diseases. With technological changes and realization of customized diets, the future research focus should shift to personalized diet analysis. This can be done by collecting data from patients and then developing a network. For example, Personalized Nutrition Project [30] and Hundred Person Wellness Project [31] have collected physical and other health parameters data for exploring disease progression. In case of a diet-based analysis, data of patients' dietary habits can be collected [30] or data of fit individuals can be accumulated using food

Figure 2.4: Framework for using Machine Learning Approach for Link Prediction in Networks by Extracting Features

questionnaires. Parameters like heart rate, blood pressure, Healthy Eating Index (HEI) and body mass index can be extracted and utilized as features. Using such data of different individuals, their similarities can be measured and further used to generate a network. Profiles of similar individuals can be clustered using appropriate clustering algorithm. Thus, a cluster can be predicted for a new individual based on its features and similarity. Based on the cluster an individual belongs to, diet of individual can be recommended (as shown in Figure 2.5). As shown in figure, features like Heart Rate (HR), Systolic Blood Pressure (SBP), EUE (for participation in exercise in last 7 days with 1 as yes and 2 as no), Body Mass Index (BMI), Steps per Day (SD) can be extracted. The data shown is sample data, but data for such analysis can be generated by monitoring patients as done in Personalized Nutrition Project or can be taken from resources like Eating and Health Module Dataset (American Time Use Survey) [232] as shown in Table 2.11. ATUS Eating and Health Module contains fields related to Body Mass Index (ERBMI), diet (EUDIETSODA for kind of soft drink with 1 for diet, 2 for regular and 3 for both/EUDRINK for drinking any beverage other than plain water with 1 for yes and 2 for no), Exercise (EUEXERCISE). Different approaches other than this, that can be embarked for personalized inferences are:

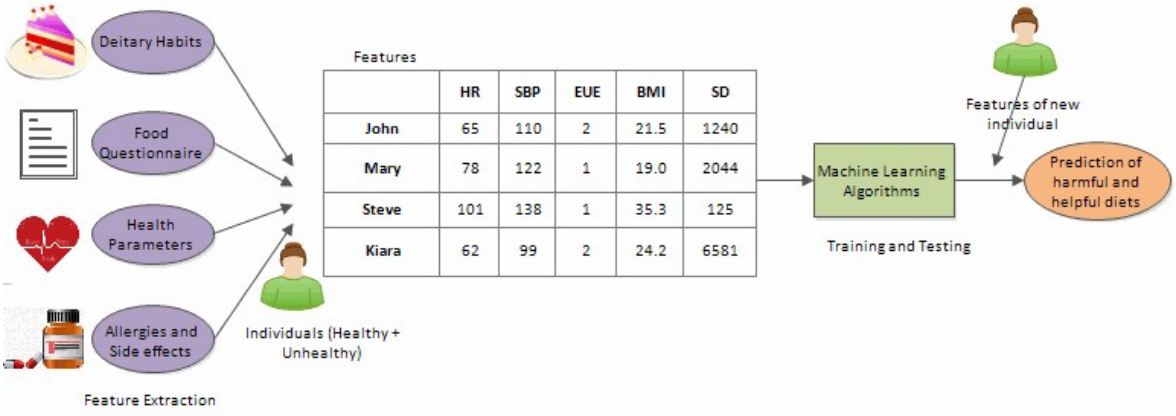


Figure 2.5: Framework for Personalized Dietary Predictions by Extracting Features

Table 2.11: ATUS Eating and Health Module Containing Different Fields

TUCASEID	EUEXERCISE	ERBMI	EUDIETSODA	EUDRINK
20160101160045	2	26.6	-1	1
20160101160066	1	44.3	-1	2
20160101160069	1	24.5	-1	2
20160101160083	1	21.2	-1	1

i. Understanding Diets: Diet oriented predictions can be performed by evaluating data from food questionnaires and dietary habits. Individuals with similar diets can be clustered in a network, and then their network and other parameters can be explored in order to perform ranking.

ii. Understanding temporal associations: Health parameters of an individual can also be collected at different points of time and dynamic graphs can be compared to understand the progression or digression of disease based on various parameters.

iii. Predicting possibility of disease: This can be done by training the dietary based dataset of healthy and unhealthy individuals suffering from a disease using different machine learning and network algorithms and creating a suitable prediction model.

2.6 Cloud Services for Predictive Healthcare

Diverse contributions have been made by researchers in terms of Cloud services for upgrading healthcare. Table 2.12 depicts several works in this regard which include different monitoring, management and prediction systems. Cloud platform has been used extensively in such applications for storage, analysis and development. It could have been customized for disease-diet prediction system in the similar manner, but there is only a single study in this regard as per literature. Rehman *et al.* [233] provides a cloud-based diet recommender service which suggests diets based on pathological tests of patients. The patients enter their test report values and diets are suggested using ant colony optimization. While this is a sound approach but it tends to suggest diets to end users. This is not a good option because sometimes the results might be ambiguous and a dietician intervention might be needed. Moreover, the results of pathological tests might vary over time leading to inconsistent diagnosis.

2.7 Conclusion

This Chapter reviews the current status of research in understanding relation between diet and health. It is realized that there is a dire need of advanced techniques for exploring disease-diet associations. Further, due to a close connection with disease based associations, this chapter presents a thorough investigation of emerging predictive technology Network Analysis, its techniques and collaborations as per the literature. The current picture of Network Analysis in healthcare domain has been discussed, along with its most prominent role in predicting disease based associations. Proposed usage of Network Analysis techniques has been provided, in addition to discussion of challenges posed to Network Analysis in healthcare applications. Existing Cloud based healthcare applications have also been summarized. This survey in itself exhibits the promising nature of these predictive technologies for exploring disease-diet associations.

The next chapter lays foundation for the exploration of disease-diet associations by describing the proposed and developed network model.

Table 2.12: Cloud Services in Healthcare

Authors	Year	Service Used	Application	Approach	Cloud Service Usage	Tool used
Rani and Kumar [234]	2021	IaaS	To predict mortality rate of babies and monitor rural pregnant women	Classification algorithms compared to predict mortality rate of new born based on pregnant women age and other factors	Pregnant women data collected in the cloud	Microsoft Azure
Rajput *et al.* [235]	2021	PaaS	To identify diabetes in the early stages	User statistics are stored and multiple prediction algorithms are compared	Data storage and analysis is performed using cloud	-
Motwani *et al.* [20]	2021	SaaS	A smart system for patient monitoring and recommendation	Data of patients collected and deep learning utilized to predict health status	Cloud used for storage, infrastructure as well as analytics	Microsoft Azure
Ahanger *et al.* [236]	2021	IaaS	To monitor and predict Covid19 spread	Data of persons is collected and temporal recurrent neural networks used for prediction	Data corresponding to various parameters is stored in the cloud	Amazon EC2
Balasubramanian *et al.* [237]	2021	SaaS	Monitoring and analysis of Covid-19 patients data	Near-Field Communication and Natural Language Processing for structured data	Cloud used for storage and natural language processing	Microsoft Azure OCR
Sood *et al.* [238]	2021	SaaS	Monitoring the spread of dengue virus	Social network analysis and Bayesian network used for monitoring and management	Cloud infrastructure used for storage, development and analysis	Amazon EC2
Al-Khafajiy *et al.* [21]	2019	SaaS	Health monitoring of elders	Physiological data of elders is stored and monitored	Cloud used for storage, infrastructure as well as analytics	-
Lin *et al.* [239]	2019	IaaS	To provide health management service to patients having cardiovascular disease	Various parameters like ECG, BP are monitored and suggestions are given accordingly	Data corresponding to various parameters is stored in the cloud	HealthMI solutions, Cmate solutions and several others
Pham *et al.* [240]	2016	SaaS	Home health monitoring using physiological data	Real time data collected and stored to be shared with caregivers	Cloud infrastructure used for storage	Private cloud built using OpenStack Juno

Continued on next page

Table 2.12 – *(Continued)*

Authors	Year	Service Used	Application	Approach	Cloud Service Usage	Tool Used
Pouladzadeh *et al.* [241]	2014	SaaS	To automatically categorize food in terms of calories	Classification of food images is performed	Cloud based SVM is used for distributed classification	Built on LibSVM using Hadoop implementation
Zhang *et al.* [242]	2014	SaaS	To recommend related drugs based on symptoms	Cloud assisted approach for recommendation	Tensor decomposition performed and data is stored in Cloud	-
Hussain *et al.* [243]	2013	IaaS	A smart clinical decision support system for chronic disease	Data of daily life activities collected and various analysis modules are run	A cloud infrastructure is used to deploy this application	Microsoft Azure

Chapter 3

DID-NEM: Proposed Network Model

In the previous Chapter, it is construed through literature review that Network Analysis and Cloud Computing are two compelling technologies which can assist in understanding food and health domain. While Network Analysis and its accomplices are promising for identifying significant disease-diet associations, Cloud platform is required for delivering the outcomes as a service. The current roles of these technologies and the future promises they hold for advancing healthcare research were also presented.

This Chapter proposes a network model DID-NEM for depicting disease-diet associations. Known disease-diet associations can be visualized in the form of a network and quantification of the associations is possible. Firstly, data pertaining to these associations is required to be curated from the available sources and literature. Further, a network is constructed to perform analysis. The proposed network model is novel and can be utilized for depicting any other similar medical associations as well.

The organization of this Chapter is as follows: Section 3.1 describes the fundamentals of the proposed network model, followed by Section 3.2 which provides an in-depth description of the curation of disease-diet associations. Further Section 3.3 describes the construction of a network utilizing the curated database, followed by conclusions in Section 3.4.

3.1 DID-NEM Description

Understanding the already known relations between different diseases and foods/dietary patterns aid in realizing the future prospects of diets. For example, if a food item is known to be related to a disease and further, the disease is known to be related to another disease, then an indirect relation between the food item and another disease can be inferred. The proposed DIsease-Diet Network Model (DID-NEM) aims to depict such associations in the form of a network. It upholds the idea that a network depiction has properties, which if analyzed using advanced techniques, can prove to be beneficial for uncovering future relations.

The first and foremost step towards developing such a model would be to collect required data, followed by identification of associations. Traditional method of collecting data is manual curation. There are several limitations due to which traditional methods of data collection and identification are not good enough:

- Information regarding relationships between diseases and diets is available from different resources for example, National Cancer Institute provides booklet for managing eating problems related to cancer treatment [244], but this data is unstructured and concentrates on a single disease. Another resource [245] which is a public education project has compiled data from food encyclopaedias and books by doctors and medical practitioners. It provides information regarding the helpful properties of foods along with references, but the data is in unstructured form and cannot be reutilized for analytics.

- Data of clinical trials undertaken in different countries is available [246] and might be used to provide information for different diseases, but such studies are based on population of a particular age, ethnicity or a specific disease. Due to this, the analysis retrieves associations which are very specific. These individual research works, if combined, are a great resource for exploring the associations on a global level.

- Extraction and development of a structured dataset is possible from the above

said available sources, but manual reading of individual research works is a tedious process, which leads to the need of an automatic system for curation.

- Many tools have already been devised for automatic mining of literature to extract different kinds of associations. Another limitation lies with the fact that these tools were designed for a specific data type, pre-defined ontologies or gold standard databases, thus they cannot be used for diet terms because it does not have a pre-defined database.

- The correlations identified in these case control studies were retrieved using traditional statistical analysis. Advanced techniques like co-occurrence retrieval, text analysis and machine learning if applied can promise improved results.

To eliminate the posed limitations and address the challenges of existing data formats, the disease-diet associations have been extracted using a curation technique. A curation technique is designed to extract associations from medical literature through a custom made application, as discussed in the next section.

3.2 Curation of Disease-Diet Associations

As described in the previous section, there is a dire need to automate the process of extraction of disease-diet associations from medical literature to further utilise it for analysis. Thus, this work aims to automatically extract disease-diet associations found in medical literature and further use the extracted database for inferring more refined relations. In this segment, a curation technique-DIsease Diet Association database Curator and Explorer (DIDACE) is proposed for automating the extraction process and inferring valuable associations for completing the dataset. The main contributions of this technique include:

- Design and development of an automatic technique for curation of associations between different diseases and diets from available medical literature.

- Development of a prediction model for predicting nature of association of a subset

of disease-diet pairs using sentiment analysis and machine learning in order to generate a complete disease-diet dataset.

The motivation behind this technique is to ease out the task of identifying relationships between diseases and diets. It aims to reduce the task of manual curation as well as achieve good accuracy. Once accurate data sets are available in right format, it can enhance accuracy and efficiency of further analysis.

3.2.1 Material Used

- *Medical Subject Headings*: (MeSH) is a vocabulary which annotates research articles by representing its main topics. A hierarchy of terms has been arranged in the form of tree structure with numbered notations as shown in Fig. 3.1. This was downloaded from MeSH website of National Library of Medicine (NLM) [247]. The MeSH tree contains 16 different categories, where category C has disease headings (or descriptors) while J02 has food descriptors. These categories have been used in this segment so that standard terms can be utilized for curation. Thus, 4758 disease terms (including subtypes) and 154 diet terms were taken from downloaded MeSH database and stored in csv files.

- *PubMed*: It contains citations of millions of biomedical articles from various journals. Moreover, MeSH thesaurus has indexed articles in PubMed. Thus, PubMed literature search is most suitable for extracting relevant papers containing both the disease and diet terms.

- *E − utilities*: Manual searching of large number of terms in such a vast literature is a tedious task. Thus, National Center for Biotechnology Information (NCBI) provides an API service named E-utilities which can be used to extract required data by posting URL queries through a software. The queries are sent as URL. For example, if one needs to search all the literature in PubMed in which both the terms Coffee and Diabetes occur together, then the following query should be posted as URL:

Figure 3.1: MeSH Tree Structure

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=coffee[mesh]+AND+diabetes[mesh]

The [mesh] term in URL depicts that the terms are taken from MeSH vocabulary. This returns an XML with number of research papers in which both terms occur together as count, along with PubMed ids of respective papers.

- Artificial Neural Network: Artificial Neural Network (ANN) is a concept replicating a human brain for solving complex problems using distributed and parallel computations. It is increasingly being used for many different applications like pattern recognition (speech, character or human face) and optimization problems. The network contains computational units called nodes/neurons connected via weighted edges. These nodes process information collected from previous nodes using an activation function. A Multi-Layer Perceptron (MLP) is a feed-forward neural network which comprises many layers of nodes with input, output and hidden layers [248] We used this because MLP is best suited for classification

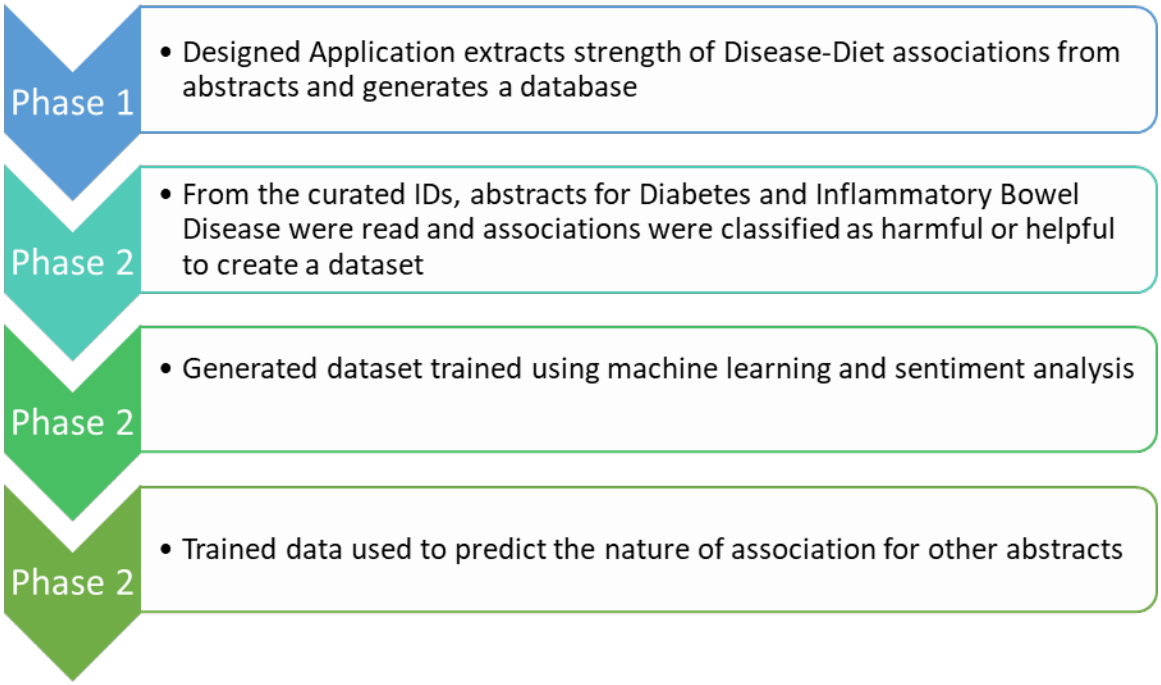


Figure 3.2: Different Phases of DIDACE

prediction problems that focus on tabular data.

3.2.2 Proposed Curation Technique: DIDACE

DIDACE is a two-phase approach proposed to design a technique for extracting disease diet associations automatically and further labelling the nature of extracted associations. In the first phase, a technique has been developed for extracting the count of medical abstracts in which both disease and diet terms occur together. This is done to quantify their strength of association. A portion of the extracted database is further used to design a prediction model for labelling the database in the next phase. Some of the abstracts retrieved in first phase (which include diet relations of diseases like IBD and Diabetes) were read and labelled as harmful or helpful. This data was further trained and used to predict harmful and helpful associations of diseases and diets in other abstracts. In this manner, labelling of database is also automated. The various phases of DIDACE have been depicted in Fig. 3.2. The two phases of the proposed approach are described as:

- *Phase 1*: In this phase, a technique has been developed for curating research

papers in which disease and diet terms occur together. It has been designed to automatically search PubMed database, extract the count of abstracts in which disease-diet terms occur together and normalize the count so as to develop a database portraying strength of association of different diseases and diets. The algorithm used to automate the extraction process has been constructed to first select all pairs of disease-diet terms from the MeSH vocabulary consecutively, then generate query URL for each pair and post it on Eutilities server, which further returns an XML page consisting the value of count of terms. The flowchart of this algorithm has been depicted in Fig. 3.3. The count (C) values thus retrieved were normalized using a formula based on Term Frequency-Inverse Document Frequency ($tf - idf$). It calculates the frequency of terms taking note of their significance across all papers. Thus, co-occurrence ($C_{i,j}$) of a disease term (i) and diet term (j) has been calculated using the formula:

$$C_{i,j} = C \times \log \frac{t}{n} \tag{3.1}$$

where t is total number of diseases, n is number of diseases in which diet term i occurs.

- *Phase* 2: In the second phase, a subset of abstracts pertaining to data extracted in the first phase have been used to predict the nature of association of disease-diet pairs. A bag-of-words representation was developed so that this data could be used for sentiment analysis. The steps of this phase are as follows:

    1. *Loading*: Abstracts were selected using IDs retrieved in XML and were read to classify them as harmful or helpful. These abstracts were loaded in Python and each abstract was cleaned for developing a vocabulary of tokens.
    2. *Cleaning*: Cleaning of abstracts involved removing punctuation, numerals and known stopwords from the documents. Tokens were converted into lower case and stemming was performed. Tokens with minimum occurrence greater than 10 were taken so as to further refine the vocabulary. The most common 50 tokens with their co-occurrences are depicted in Fig. 3.4.

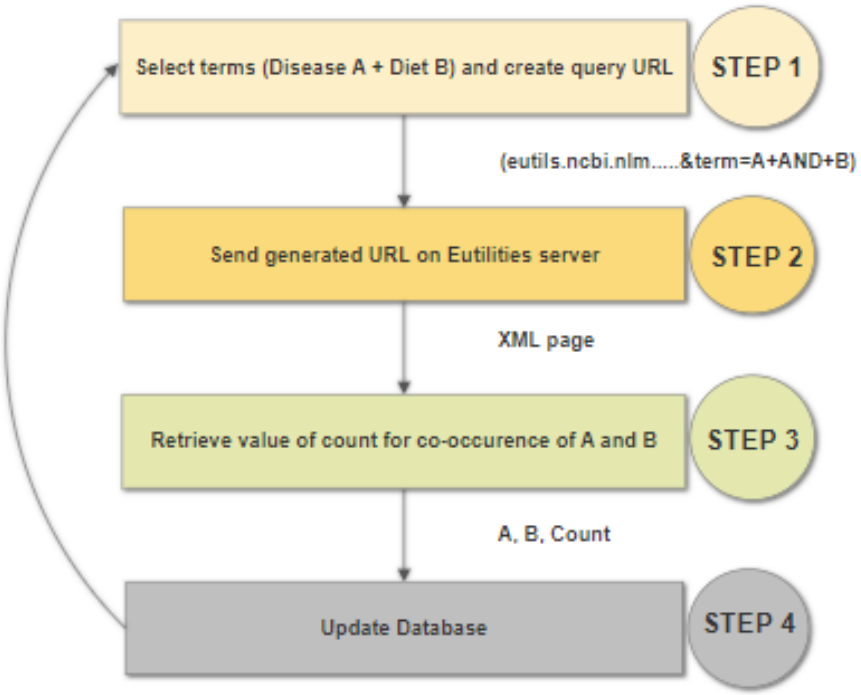


Figure 3.3: Flowchart of Proposed Algorithm

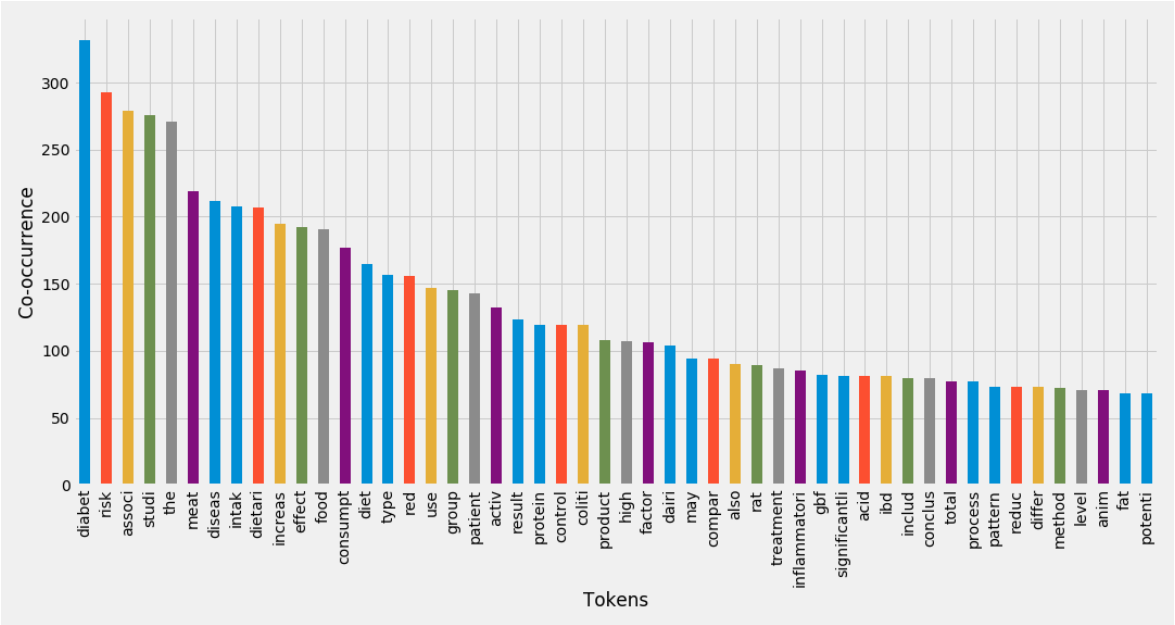


Figure 3.4: Distribution of Most Common (50) Tokens in Documents

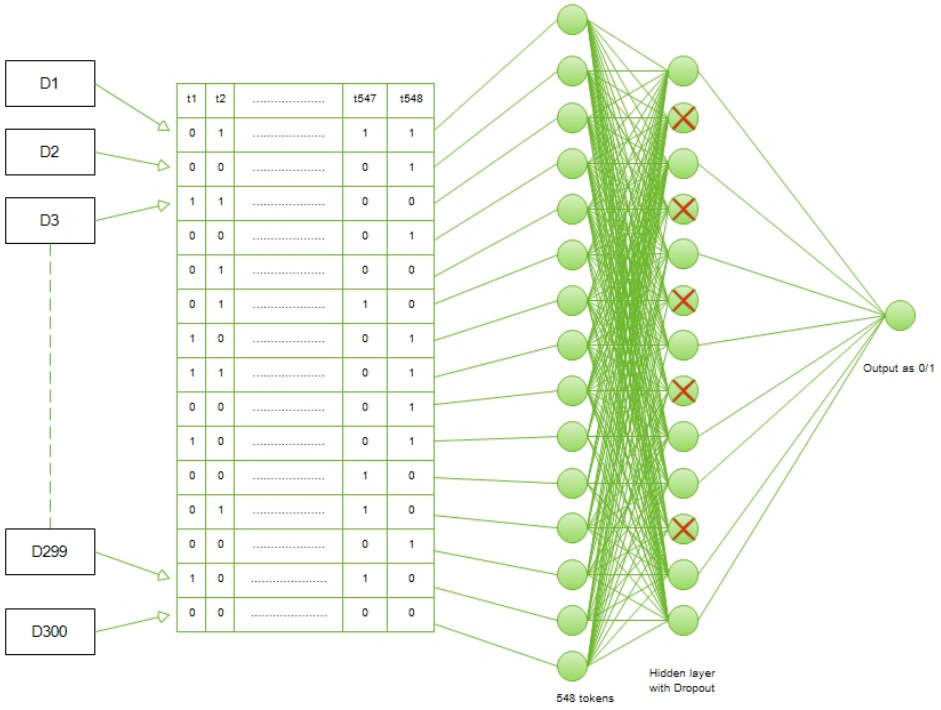


Figure 3.5: MLP based Neural Network Architecture

3. *Encoding*: This vocabulary was further used to convert the documents into encoded vectors. The tokens in each document were scored on the basis of three vectorization methods namely, binary, count and term frequency-inverse document frequency (tf-idf). Binary method simply marks the presence (*1*) or absence (*0*) of the token whereas count method outputs the number of occurrences of each token. For example, the first document (*D1*) with n tokens can be represented as follows:

   $t_1\ t_2\ t_3\ \ldots\ t_n$

   $D1 = (0\ 1\ 1\ \ldots\ 0)$

   where *t1* represents first token and a binary method of scoring is used.

4. *Prediction*: The encoding vectors were loaded in python for training a neural network so that prediction model can be developed. Due to tabular nature and low dimension of our data, a simple MLP has been applied. MLP performs supervised learning for predicting class (harmful or helpful) for new abstracts. It consists of many perceptrons taking input values of

Table 3.1: Number of Instances used for MLP Training

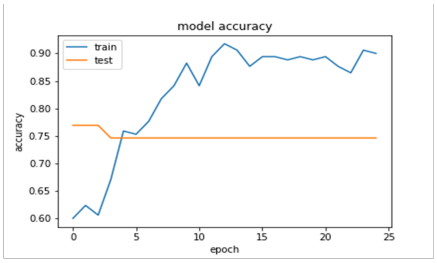
No. of Instances	**Negative**	**Positive**
Training	70	100
Testing	30	100

features (encoded vectors in this study), which are weighted and further summed up to be used as input to an activation function. This function aids in classification decision. There can be multiple layers in an MLP, but due to low dimension of data, we chose a single hidden layer for this task. The MLP based neural network for this model is represented in Fig. 3.5.

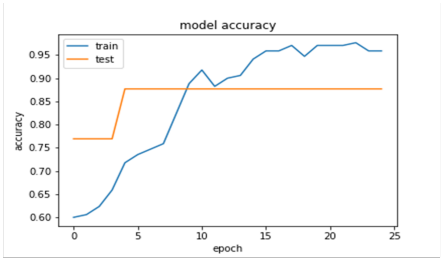
3.2.3 Experimental Details

- *Phase 1:* The proposed algorithm was executed as a java program. Around 3.5 lakh records were extracted, but data was filtered so as to remove records containing only 0's as correlation or other repetitions. Since this data has been curated using a program, more data could be collected in less time unlike the technique followed in [111].

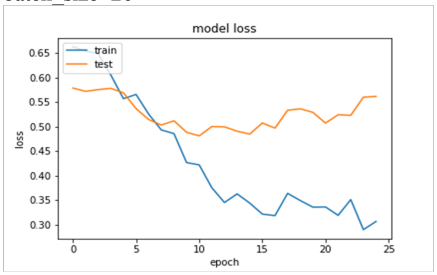
- *Phase 2:* Using the ids from XML in Phase 1, 300 abstracts for Diabetes and Inflammatory Bowel Disease were searched and read. 100 harmful and 200 helpful abstracts were identified. The cleaning, loading and encoding of abstracts for creating a vocabulary have been performed using Keras API in Python. 548 tokens were taken in vocabulary after cleaning, which were used to convert abstracts into encoded vectors. The ratio of instances used for training and testing are shown in Table 3.1. Since the negative samples (harmful abstracts) were less than the positive samples (helpful abstracts), we chose 70:30 ratio for splitting negative samples whereas 50:50 for positive samples. The run was repeated 10 times picking random samples for splitting in each turn so that average values can be considered. MLP with one hidden layer was trained using sigmoid activation function as the model achieves better accuracy with this function. The output layer consists of one neuron with sigmoid activation function. Adam optimizer



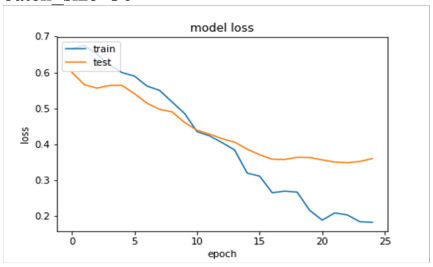
A) Graph of Epochs and Accuracy with batch_size=20



B) Graph of Epochs and Accuracy with batch_size=30



C) Graph of Epochs and Loss with batch_size=20



D) Graph of Epochs and Loss with batch_size=30

Figure 3.6: Comparison of Different Hyperparameters for Tuning

along with binary cross entropy loss function were chosen for training. Values of accuracy and loss were compared with different epochs and batch sizes so as to tune the parameters. Fig. 3.6 depicts the comparison of (i) Epochs and Accuracy, along with (ii) Epochs and Model loss, with different batch sizes. As can be seen in the figure, better accuracy has been achieved with a batch size of 30. Moreover, value of loss decreases more when this batch size is used. It reaches a minimum when the epoch value is between 20 and 25. We also introduced dropout in our MLP in order to randomly set nodes as 0 in the hidden layer. This helps in randomly selecting nodes, thus avoiding over-fitting of data. The dropout rate considered in our model is 0.2. The model also ensures that it does not suffer from exploding/vanishing gradients problems because it has only one single layer and moreover there are no large changes in loss on each update. The various hyperparameters set for this model after tuning are depicted in Table 3.2. The predictions were performed for new abstracts as 0 for harmful and 1 for helpful class.

Table 3.2: Parameters Tuned for Training MLP

Parameter	**Neurons**	**Dropout Rate**	**Learning Rate**	**Epoch**
Value	20	0.2	0.01	23

Table 3.3: Sample Associations Extracted in Phase 1

Diet	**Disease**	**Reference**
Avocado	Joint Disease	[249]
Tea	Arthiritis	[250]
Tea	Liver Disease	[251]

3.2.4 Results

- *Phase 1:* The final database extracted contains a total of 2,74,131 records containing 1917 different diseases and 143 diet terms. Some distinct associations were realized from this database as shown in Table 3.3. References for validation of extracted relations are also mentioned in the table.

- *Phase 2:* For prediction task, 73 new PubMed abstracts pertaining to Cardiovascular and Inflammatory Bowel Diseases were chosen. Due to this, the dataset constitutes of two different diseases, thus achieving the aim of predicting for varied diseases and diets. The model predicts 0 for harmful and 1 for helpful associations. The validation of results is performed using accuracy. The algorithm achieved different accuracies when different vectorization methods (namely binary, count and tf-idf) were used for encoding as shown in Table 3.4. A boxplot for 3 types of vectorization methods has also been presented in Figure 3.7. The boxplot depicts minimum to maximum accuracies achieved for different vectorization methods. The best accuracy of 85% is achieved when tf-idf is used. Another method of validation used in this study involves measure of precision and recall. Precision is a measure used to depict the accuracy of predicted positives. A false positive in our research indicates an association which is predicted to be helpful although it is not helpful. Due to this, a false positive is quite unfavorable for our research. Less number of false positives implies high precision, which in turn indicates a better model. The confusion matrix for this model has been

Table 3.4: Accuracies for Different Vectorization Methods

	binary	**count**	**tf-idf**
count	25.0	25.0	25.0
mean	0.804941	0.745412	0.857176
std	0.153381	0.150762	0.139858
min	0.458824	0.411765	0.517647
25%	0.664706	0.635294	0.758824
50%	0.835294	0.788235	0.923529
75%	0.947059	0.858824	0.958824
max	0.976471	0.935294	0.982353

Table 3.5: Confusion Matrix for Prediction Model

		Precision	
	N=73	Negative	Positive
Actual	Negative	9	6
	Positive	11	47

depicted in Table 3.5 which is used to evaluate precision and recall. The model has a good performance as it achieves precision 88.7%, recall 81% and F1 score 84.7% as shown in Table 3.6. It is also important to look at these measures to identify class imbalance problem which might have occurred in our dataset due to different number of positive and negative samples. Among these predicted associations, some test cases were taken to be validated in real life as shown in Table 3.7. Red meat is one such test case found to be a harmful component for cardiovascular diseases. A blog by National Institutes of Health (NIH) confirms that daily consumption of red meat triples a chemical related to heart diseases [252]. Apart from this, some helpful associations have also been predicted. Yoghurt and soy are found to be beneficial in case of Inflammatory Bowel Disease. Center for Applied Nutrition (CAN) of University of Massachusetts Medical School provides various recipes and dietary recommendations for IBD, and it recommends yoghurt and soy products for the same [253].

Table 3.6: Parameters for Validation of Prediction Model

Parameter	Value
True Positives	47
False Positives	6
True Negatives	9
False Negatives	11
Precision	88.7%
Recall	81%
F1-score	84.7%

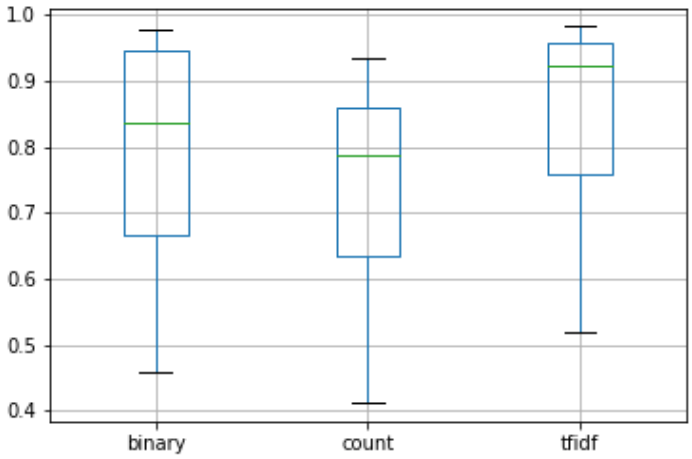


Figure 3.7: Boxplot for Accuracies in 3 Different Vectorization Methods

Table 3.7: Sample Associations Predicted in Phase 2

Diet	Disease	Association	Reference
Red Meat	Cardiovascular Disease	Harmful	[252]
Yoghurt	IBD	Helpful	[253]
Soy	IBD	Helpful	[253]

3.3 Construction of a Disease-Diet Network

Networks have been used to represent, organize, mine and predict unknown links from already existing complex data. They are extensively used for applications such as search engines, fraud detection and recommendation systems. Networks are suitable for interdependent and complex data because they can represent it effectively and thus, can infer new knowledge from the dependencies. The complex interdependencies generated in disease-diet associations can be explored using Networks so as to predict future progression. The various steps involved in the construction of a disease-diet network include:

- *Database Validation:* Before proceeding to construction of network, an important step of validating the database cannot be missed. Validation by a certified professional or PubMed publications ensures that only reliable information is passed on for further construction and analysis. Nevertheless, validation of such a large dataset containing 274131 records is not feasible. Thus, only subsets of the database have been validated as per requirement in the two case studies. However, the overall distribution of co-occurrences for curated disease-diet associations has been observed as shown in Figure 3.8. It is evident from the figure that most of the associations have low co-occurrences while a very few associations have high values. This is similar to a power law distribution which is a property of complex networks [35]. Thus, the curated database have been found to be a good fit for developing a network.

- *Database Uploading:* In order to create disease-diet network, first we need to upload the curated and validated database generated in the first segment. There are many graph database softwares which can be used to store and represent graphs including Neo4j, OrientDB and others [254]. In this work, the proposed network has been constructed in one of the benchmark platforms for graph database management, named Neo4j. It is a native graph database thereby meaning to include direct pointers in a node for all its connected nodes [255]. This implies that relationships can be fetched directly which saves lookup time and improves

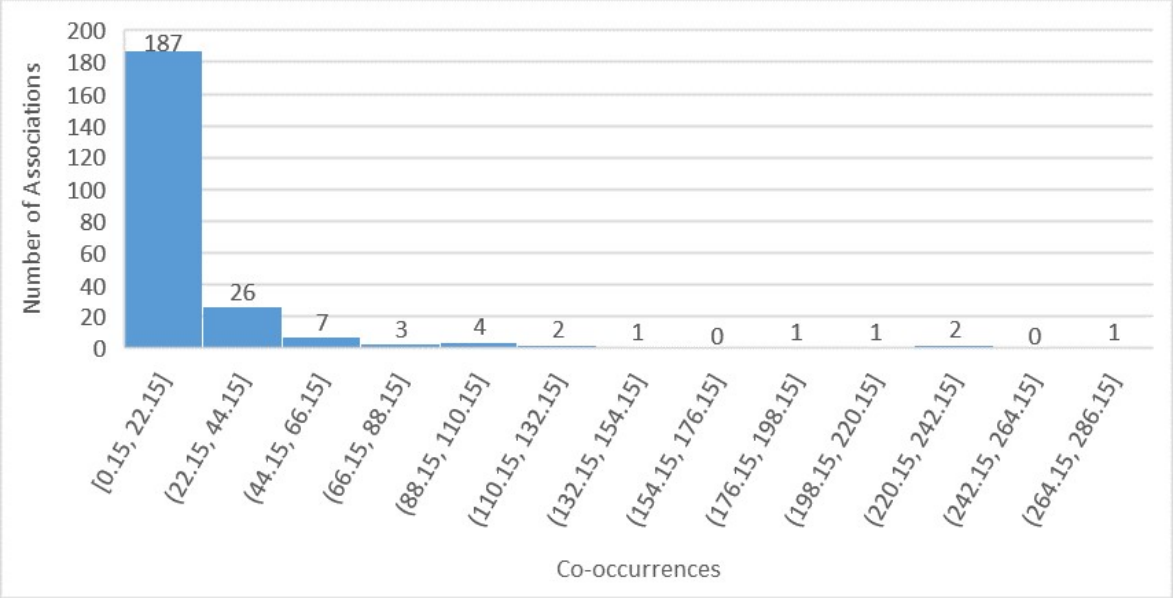


Figure 3.8: Distribution of Co-occurrences for Curated Disease-Diet Associations

processing of the database. It supports Cypher Query Language (CQL) for accessing graph database. The curated database generated as CSV file has been uploaded as Neo4j database using the load query as shown below:

LOAD CSV WITH HEADERS FROM "*location of file*" AS line
MERGE(n:disease{Name:line.DISEASE})
MERGE(m:diet{Name:line.DIET}) with m,n,line
MATCH (n:disease),(m:diet) where line.cooccurrence_new <>0
CREATE (n) - [r:linked_to {cooccurrence: toFloat (line.cooccurrence_new), relation: line.rel_new}] → (m) return m,n

- *Disease-Diet Network Visualization:* After creating the Neo4j database, network can be visualized using query like:

MATCH p=()-[r: linked_to] → () RETURN p
This creates a network consisting all the relationships connecting nodes in the database. The graph database can also be queried in other forms, thus creating different graphs for different scenarios.

3.4 Conclusion

This Chapter undertakes the presentation of the proposed network model DID-NEM through two important segments. Firstly, it describes a customized technique DIDACE designed for curation of disease-diet database, followed by the results retrieved in this segment. Secondly, it unfolds the steps involved in network construction of the extracted database through the second segment. Thus, a comprehensive disease-diet associations database is extracted, refined, validated and represented in this Chapter. The next Chapter utilizes the curated database to manipulate networks and perform analysis using advanced techniques. For this purpose, two different case studies are undertaken and prediction frameworks are customized for each.

Chapter 4

PredNEM: Proposed Prediction Approach using Network Model

The previous Chapter presented the proposed network model DID-NEM along with its two segments. The first segment is an approach towards curating and quantifying disease-diet associations, followed by its network contruction in the second segment. The main aim is to extract, refine and represent a comprehensive database and then identify associations amongst disease and diet.

The current chapter is a thorough description of the essentials required to design a prediction approach for manipulation and analysis of the developed network. The proposed approach PredNEM aims to predict unknown disease-diet associations using network algorithms, network parameters and machine learning. The implementation of the proposed approach is demonstrated and validated through two different case studies corresponding to Covid-19 and Inflammatory Bowel Disease (IBD).

The organization of this Chapter is as follows: The chapter begins with a description of PredNEM in Section 4.1, followed by a description of case studies Covid-19 and IBD in Section 4.2. The experimental validation of PredNEM in first and second case studies are described in Section 4.3 and Section 4.4 respectively. The chapter is concluded in Section 4.5.

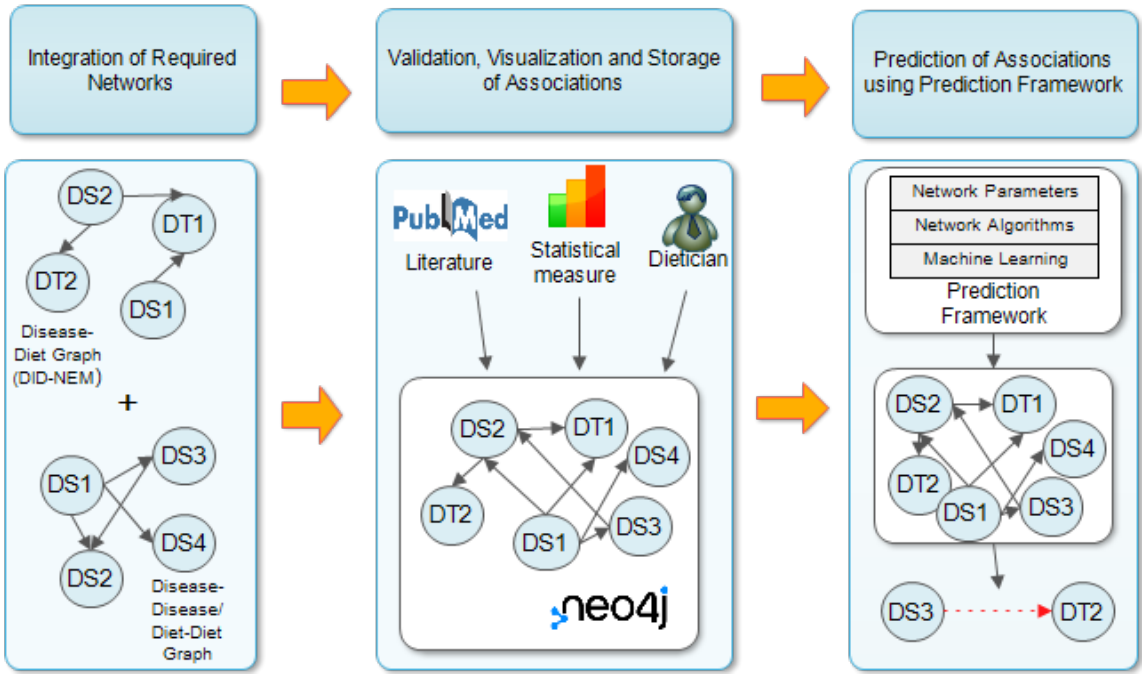


Figure 4.1: Steps followed for PredNEM (DS refers to a Disease Term whereas DT refers to a Diet Term)

4.1 PredNEM Description

A network poses many opportunities for understanding the dynamics of its nodes. A careful manipulation and analysis of complex networks would be able to infer interesting associations. Thus, a Prediction Approach using Network Model (PredNEM) has been proposed in this work to achieve the objective of predicting unknown associations between diseases and diets. The approach aims to integrate different networks and exploit different learning methods based on Network Analysis and/or its alliance with machine learning. The steps involved in designing of the Proposed Prediction Approach are depicted in Figure 4.1 and discussed in the subsequent sections.

4.1.1 Integration of Required Networks

DID-NEM proposed in this work consists of associations between diseases and diets nodes curated from medical literature. However, dependencies among these nodes should also be considered for a robust analysis. For instance, integration with disease-disease or diet-diet associations network would ensure diversity in the network, thereby

facilitating a deeper understanding of associations. The required associations can be retrieved from already available tools like some toolkits offer semantic similarity between diseases. Further, diet-diet associations can be curated from literature using a curation technique like DIDACE or by customizing some other machine learning algorithm. The retrieved associations are integrated with DID-NEM to generate a network consisting of diverse but related associations.

4.1.2 Validation, Visualization and Storage of Associations

The associations curated or extracted from various resources should be validated wherever possible to ensure the authenticity of data in question. It can be performed by confirming through PubMed literature, a certified dietician or statistical observation. Further, the network made from integrated associations in the previous step need to be stored and visualized. A graph database like Neo4j which is specifically designed for representing and storing associations of a graph/network is the best solution for this scenario.

4.1.3 Prediction of Associations using Prediction Framework

Prediction in the integrated network is performed with the aid of a carefully designed framework. Combination of machine learning and network parameters/algorithms in the form of learning methods fabricate the foundation of Prediction Framework. The framework incorporates pre-processing of the data and utilizing these learning methods for performing prediction. There are two different learning methods realized for construction of the framework. The methods mainly differ in the way amalgamation of network parameters/algorithms and machine learning has been utilized to cater to the prediction of associations. The different learning methods utilized in this study are enlisted as:

- *Top to Bottom Method (TBM):* In the proposed method, the complete network is first divided into communities using a community detection algorithm, followed by ranking of the nodes. The ranking is used as a similarity metric for find-

ing most related nodes. The method is named Top to Bottom because it starts from finding communities in the whole network and further divides it for finding the most similar nodes. Combination of these network algorithms and machine learning algorithms is found to be quite efficient as demonstrated in the first case study for predicting diet associations for novel disease Covid-19. The network and machine learning algorithms configured for the prediction framework in this method are Louvain Algorithm for community detection, Page Rank algorithm (PR) for ranking and K nearest neighbors (KNN) for finding similarity. The algorithms are described as:

i) Louvain algorithm: Louvain algorithm is an unsupervised graph algorithm for inferring communities [256]. Nodes reported in the same community are known to be more related to one another than those in other communities. The algorithm works on the principle of maximizing modularization gain, developing communities and reiterating the previous steps until no more changes are evident. The modularity gain of a node in a community is defined as:

$$M = \left(\frac{\sum_{in} + 2k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right) - \left(\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right) \quad (4.1)$$

where m is the sum of weights of all the relationships in the graph, sum_{in} is the sum of relationships in the community, $k_{i,in}$ is the sum of weights of relationships starting from node i to other nodes in the community, k_i is the sum of weights of relationships incident to i and sum_{tot} is the sum of weights of relationships incident to nodes in the community.

ii) Page Rank: Page Rank (PR) calculates the rank of a node by using the number of nodes with which it is connected [42]. It was introduced in Google for ranking the webpages so as to optimize the search.

The machine learning algorithm configured for the prediction framework in this case study is described as:

i) K Nearest Neighbors (KNN): This algorithm utilizes the training data and distance metric to find K closest instances of a new input and predicts output

variable based on those instances [257].

- *Topological Features Method (TFM):* Another alternative realized for learning is to use topological properties of the network as features for machine learning algorithms which substantially enhance the learning capability. Thus, network properties and network algorithms are crafted as features for performing machine learning in this method. Several different network properties and algorithms configured in this method are described as:

  i) Triangle counting: It is a community detection graph algorithm that is used to determine the number of triangles passing through each node in the graph [231]. A triangle consists of three nodes, where each node has a relationship to all other nodes. For every pair of nodes, maximum and minimum count are taken as features.

  ii) Local Clustering Coefficient algorithm: The algorithm calculates the probability of neighbours of a node to be connected [231]. It is computed for each node in the graph using the formula:

$$C_x = \frac{2T_x}{n_x(n_x - 1)} \tag{4.2}$$

  where C_x is the local clustering coefficient, n_x is the degree of node x and T_x is the triangle count of node x. For every pair of nodes, maximum and minimum coefficient is taken as feature.

  iii) Common Neighbours: It is one of the many local similarity indexes used in link prediction based on the principle that two nodes having more common neighbours, have more chances of connection [258]. Common neighbours (F1) for any two nodes a and b is calculated by:

$$F_1(a, b) = |n(a) \cap n(b)| \tag{4.3}$$

  where $n(x)$ corresponds to nodes adjacent to x.

  iv) Preferential Attachment:

  Preferential attachment follows the concept of nodes with higher degree having

more ability to connect with new nodes. It is used for developing scale free property which is evident in complex networks [258, 259]. It is computed as:

$$F_2(a, b) = |n(a)| * |n(b)| \tag{4.4}$$

where $n(x)$ corresponds to degree of x. A higher value suggests more closeness.

v) Total Neighbours: Similar to common neighbours, this index is used to find the closeness based on adjacent unique nodes [231]. The formula for calculating total neighbours is as follows:

$$F_1(a, b) = |n(a) \cup n(b)| \tag{4.5}$$

where $n(x)$ corresponds to degree of x.

vi) All Pairs Shortest Path algorithm: This algorithm computes shortest distances between all the pairs in the graph using the weights of edges [231]. A shorter length of path implies more connected nodes.

vii) Weakly connected components: This algorithm is used to find sets of nodes which are connected [231]. This means that nodes in the same set are reachable to one another.

Several different machine learning algorithms have been utilized in this study which are described as:

i) Support Vector Machine (SVM): It is a supervised algorithm which outputs an optimal hyperplane which best divides the given data in feature space into respective classes. The hyperplane is learnt using a suitable kernel function [257].

ii) Classification And Regression Trees (CART): CART, also known as decision trees is an algorithm which generates a tree like structure consisting of attributes as nodes, their values as decisions/branches and leaf node as the class label [257].

iii) Gradient Boosting (GB): It is an algorithm in which weak learners (decision trees) are gradually added in a gradient descent manner. In each step, a loss function is minimized so as to retrieve a better final model [257].

iv) Naïve Bayes (NB): It is an algorithm based on Bayes' Theorem for predicting

probability of each class and selecting the one which has a higher value [257].

v) Stacking Ensemble (SE): Stacking ensemble is a machine learning method to combine the predictions of multiple models so as to improve the performance [260]. This is done by means of two layers in which the first layer consists of two or more models known as base models. The model in the second layer is known as meta model and it learns to combine the predictions of base models in the best possible manner. The base models perform predictions for data not used to train them so that the predictions and expected outputs can be treated as inputs to the meta model, thus making it more efficient.

DID-NEM and PredNEM were orignally designed for depicting and manipulating global view of associations, but validation of such a large dataset is not possible without a team of experts and dieticians. Thus, only a subset of diseases have been undertaken for study in this work. Experimental validation of PredNEM has been performed by undertaking two case studies corresponding to diseases, Covid-19 and IBD. Moreover, the two case studies also demonstrate two different learning methods of the prediction framework. Description and other details of case studies undertaken in this work are discussed in the subsequent sections.

4.2 Case Studies Description

Two different diseases, Covid-19 and IBD have been undertaken as case studies in this work. Although, the two studies utilize a common prediction approach PredNEM, but yet they are entirely different, since the first case study revolves around a novel disease and the other one is a well known disease related to diet. A detailed description of the involved diseases in the case studies is presented in this section.

4.2.1 Case Study I: Covid-19

Covid-19 is an infectious disease developed due to transmission of novel coronavirus (2019-nCoV), causing respiratory problems and deaths across the globe [261]. Covid-19

was first detected in China in 2019 and in a matter of few months, Covid-19 was declared a global pandemic affecting millions of lives. World scientists and researchers are putting in all the efforts to explore different methods related to containment of coronavirus. In the initial phase of Covid-19, it was identified to be specifically deleterious to children and older adults, but with the reappearance of second wave and increased cases, it has been identified to be associated with many other factors like Type 2 diabetes, obesity, respiratory illness etc. A recent research [262] suggests to closely observe Covid-19 patients having comorbidities like diabetes or pneumonia. Similarly, many other studies and meta- analysis revealed hypertension, chronic obstructive pulmonary disease, diabetes, cerebrovascular disease and cardiovascular disease as risk factors of Covid-19 [263], [264], [265].

Apart from comorbidities, another deciding factor of the progression of Covid-19 is the immune system [266], [267]. Due to variance in the immunity of individuals, their response to the virus is variable. Some patients are asymptomatic and recover through isolation and medicines, while others have mild to severe symptoms requiring hospitalization. In this time of crisis, there has been an urgent need to improve immune system to minimize the risk of Covid-19 and other related diseases. Diet is a vital component for immune system and thus considered important for reducing risk of Covid-19 [268]. The nutrients found in fruits and vegetables have anti-inflammatory effects and are suggested for Covid-19 risk management. Vitamin A, vitamin C, Selenium and Zinc are known to be potential options for prevention from Covid-19 as they are effective for immune functions [269]. Deficiency of vitamin D is known to be related to several diseases like diabetes, hypertension and obesity, which are associated with Covid-19 risk. Thus, many researchers [262, 263, 264, 265] are suggesting vitamin D intake to protect against infection. Few recommendations have been suggested for optimal nutrition at different levels of health model such as leafy vegetables, dairy products, nuts and citrus fruits, consisting of nutrients like iron, zinc, vitamin A, C, B6 and B12 [270]. Other works [267, 268] suggest foods like kiwifruit, broccoli, red pepper, strawberries and citrus foods which are rich in vitamin C, and foods rich in vitamin A like carrot, spinach, sweet potato, and vegetable oils, seeds, spinach or supplements for vitamin D

and E. Similar recommendations have been given like adding foods containing vitamin B6 and omega-3 polyunsaturated fatty acids to the list [271]. Consumption of diets rich in saturated fats, refined carbohydrates and sugars named as Western Diet is known to severely affect immune system leading to impairment against viruses and is thus not being recommended [272].

With more and more research about Covid-19, it is now a well-established fact that it is associated with immune system and thus its prognosis is closely related to diet. Table 4.1 describes the recent research works related to association of food with Covid-19. Another factor, which is of importance is that Covid-19 has been found to be associated with a few other diseases [273, 274, 275, 276, 277] including Non Alcoholic Fatty Liver Disease (NAFLD) and Diabetes Mellitus Type 2 (T2DM). Due to the complex interdependencies present in this scenario, the problem of understanding the dynamics of Covid-19, diet and other diseases is good fit for demonstrating the use of networks and algorithms. The available information can be integrated and manipulated for predicting diet associations through PredNEM. Being a case study for a novel disease, such an approach would provide fast and efficient predictions to benefit domain experts for further studying them. The approach might also act as a baseline to be used for predicting associations for a novel disease in future.

Table 4.1: Recent Studies Related to Understanding Diet Associations with Covid-19

Authors	**Year**	**Type of Study**	**Approach**	**Methodology**	**Results**
Razaghi-Moghadam *et al.* [278]	2021	Observational study	To test association between Zinc, Selenium and Covid-19 severity	Zinc and serum Selenium data of Covid-19 patients were analyzed using statistical tests	Significant association found between disease severity and Zn and Se levels
Losso *et al.* [279]	2021	Review	Plant based food, young age and microbiota are factors for less infection	Survey of literature to identify most consumed food items and their anti-inflammatory properties	Foods consumed in Sub-Saharan Africa including banana, sweet potato and yam possess anti-inflammatory properties
Rocha *et al.* [280]	2021	Review	Role of several macro and micro nutrients	Survey of literature to identify role of macro and micro nutrients	Certain dietary recommendations provided for management of Covid-19
Abdulah *et al.* [281]	2020	Statistical Study	To test association of diet with mortality and infection rate of Covid-19	National dietary data for countries statistically analyzed for calculating mortality and infection rate	Fruits and sugar-sweetened beverage has a positive impact on mortality and infection rate unlike beans and legumes
Gasmi *et al.* [282]	2020	Review	Micronutrients for management	Survey of literature to identify micronutrients for Covid-19 management	Vitamin D, C, A and E, Zinc, Selenium, Magnesium, N-acetylcysteine and polyphenolic compounds are helpful
Jayawardena *et al.* [283]	2020	Review	Improve immunity for viral infections	Survey of literature to identify nutrition for management	Vitamin D, Vitamin A, Zinc, Selenium, and certain probiotics are found helpful
Meltzer *et al.* [284]	2020	Retrospective cohort study	To test association between Vitamin D levels and Covid-19 test results	Multivariate analysis performed for Covid-19 patients whose vitamin D levels were known	Testing positive for Covid-19 is found to be associated with lack of Vitamin D

Continued on next page

Table 4.1 – *(Continued)*

Authors	**Year**	**Type of Study**	**Approach**	**Methodology**	**Results**
Abobaker [285]	2020	Review	Role of Vitamin C	Survey of literature to identify role of Vitamin C	Due to its anti-inflammatory and anti-viral properties, it is useful in management
Budhwar *et al.* [286]	2020	Review	Role of nutrition in prevention	Survey of literature to identify helpful nutritional interventions	Several nutrients and recipes identified for boosting immunity

4.2.2 Case Study II: IBD

IBD is a condition caused due to chronic inflammation of the gastrointestinal tract [287, 288]. It has two subtypes namely Ulcerative Colitis (UC) and Crohn's Disease(CD). The two subtypes differ in terms of the affected location and type of inflammation. UC occurs in large intestine and rectum whereas CD can be seen in any part of the tract. In UC, damaged areas are continuous and inflammation is present in the innermost lining of colon. This is in contrast to CD in which the damaged areas are found in patches and the inflammation might reach multiple layers of the tract. While environmental and genetic components are found to be factors behind its onset, dietary factors are also being considered important towards this subject [289, 290]. Being well known through extensive research, IBD has already been established as a disease related to dietary factors over the years. A vast literature corresponding to exploration of diet associations for IBD and its subtypes is available in the form of meta-analysis, prospective studies or case-control studies [291, 292, 293]. This leads to an opportunity for demonstrating the use of computational techniques for an in-depth analysis of disease-diet associations. PredNEM can be utilized in such scenario for finding potential disease-diet associations. Such an approach would benefit caregivers for planning a healthy lifestyle by eliminating toxic eating patterns.

4.3 Experimental Validation of Case Study I: Covid-19

4.3.1 PredNEM for Case Study I

Foods are known to be related to many diseases in literature, which are in turn related to Covid-19. This transitive relation has been used in this case study to infer unknown diet associations for Covid-19 and other diseases. Detailed description of steps performed for exploring associations in this case study are shown in Figure 4.2. Experimental validation of PredNEM designed for this case study is discussed in subsequent sections.

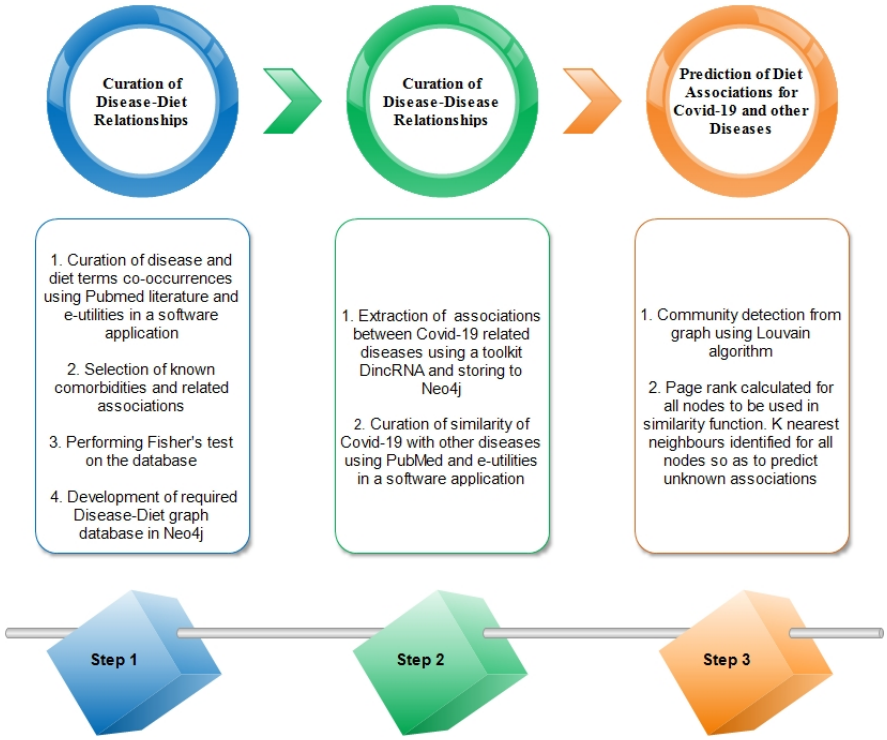


Figure 4.2: Description of Steps for Exploring Associations among Covid-19, Diet and Other Diseases

4.3.1.1 Integration of Required Networks

Covid-19 is a novel disease and research regarding its associations with other diets is still in its infancy. However, diseases having evidence in literature regarding its relation to Covid-19 can be undertaken for representing disease-diet associations in DID-NEM. Thus, the first and foremost step of PredNEM is utilizing DID-NEM as a baseline for further extracting experimentally supported associations amongst diet and Covid-19 related diseases. It is further integrated with disease-disease associations network. Integration of these networks yield a graph containing Disease-Diet and Disease-Disease associations. The various steps of integration are discussed as:

- Selection of Diets Associations with Covid-19 related Diseases

  A study [273] explored the gene expression patterns of Covid-19 and found high similarities with the characteristic patterns of a very few other diseases. The diseases include T2DM, leukemia, psoriasis, Pulmonary arterial hypertension and NAFLD. The similarities suggest that persons suffering from these diseases might

need to take extra care for prevention of Covid-19 risk. Other studies also suggest that patients suffering from T2DM and NAFLD are known to be at a greater risk of developing infections and thus Covid-19 [274, 275, 276, 277]. Thus, out of these diseases only T2DM and NAFLD have been considered for study due to presence of strong evidence in literature regarding association with Covid-19. DID-NEM with 274131 experimentally supported associations became a baseline for selecting records pertaining only to diseases known to be related to Covid-19. A total of 235 relations between NAFLD-Diets and T2DM-Diets have been selected from the curated database.

- Curation of Disease-Disease Associations

  After the Covid-19 related Diseases and Diets associations have been established; the next step was to understand the associations between the underlying diseases. The associations between NAFLD and T2DM could be extracted using similarity measures. One approach could be to use the curated Disease-Diet association graph and find similarity of diseases based on graph using traditional measures [294] like Cosine similarity or Euclidean distance, but this might make the entire database redundant. The association between NAFLD and T2DM has been extracted using semantic similarity which is based on finding relatedness from semantic meanings of terms. The following tasks have been performed in this step (also shown in Figure 4.3):

  i) Finding similarity between NAFLD and T2DM: A toolkit named DincRNA [295] provides DisSim tool offering five state-of-art methods to choose for calculating similarity of diseases. In this work, Wang's method [296] is used which is based on finding semantic similarity between terms using hierarchical structure in the ontology. This method is used because it adds a different perspective (using semantic meanings) to the dataset unlike other traditional similarity metrics. The retrieved data is further normalized using min max scaler with a range of 1-10. This range is same as the previous step so that complete dataset follows same distribution.

  ii) Finding similarity of Covid-19 with other diseases: If Covid-19 node is in-

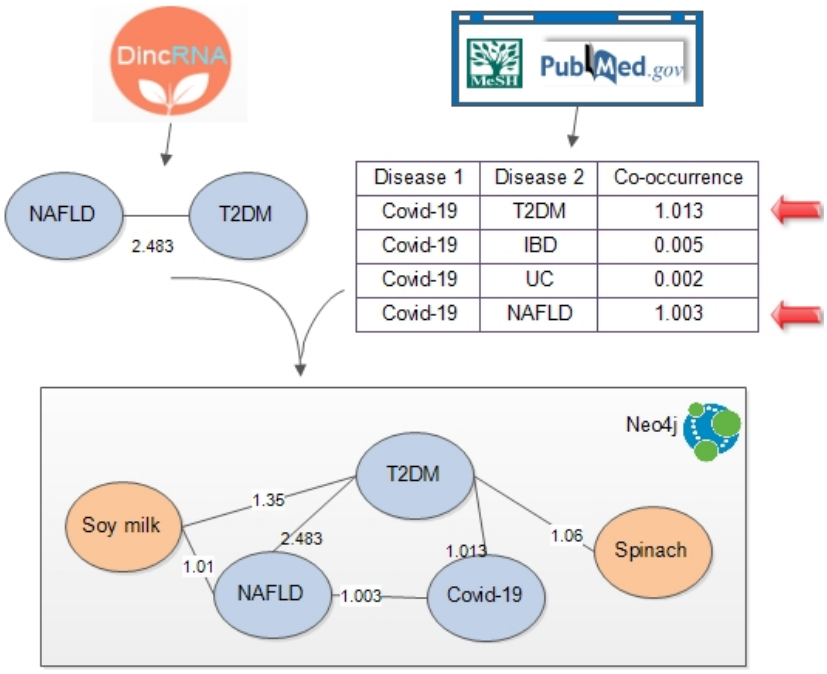


Figure 4.3: Steps for Curation of Disease-Disease Associations Graph

troduced in this graph along with its similarity with the other diseases as relationships, then its association with diets can be calculated. Due to its novelty, associations database between Covid-19 and other diseases is not available as of now. Thus, similarity of Covid-19 with other diseases is computed by curating its co-occurrences using PubMed and e-utilities. The retrieved dataset is normalized using min-max scaler (range 1-10) and records pertaining to relationships with NAFLD and T2DM are used to develop a graph containing Disease-Disease associations.

4.3.1.2 Validation, Visualization and Storage of Associations

The retrieved Disease-Diets database contained associations between two variables, namely disease and diet. A statistical significance test needed to be performed to examine the significance of their associations. Since more than 20% of the relations have expected frequencies (co-occurrences) less than 5, Fisher's exact test was performed on this data to comprehend statistical significance [297]. The null hypothesis in this case indicated that the disease terms were independent to diet terms, whereas the

alternative hypothesis suggested a relationship between the two variables. The obtained p-value=0.0004997 was less than 0.05, thus the null hypothesis was rejected at 5% significance level. This affirmed a significant relationship between diseases and diets in the retrieved database and thus ensured authenticity of the database. The associations database retrieved as in the previous steps have been further stored as a graph in Neo4j. The graph initially contained 137 diet nodes, 2 disease nodes (T2DM and NAFLD) and 235 relationships between them. A subgraph of this graph database is shown in Figure 4.4 in which the weight of edges depicts normalized co-occurrences using min-max scaler. Min-max normalization has been used so that the complete dataset is viewed on a standard scale even if data from different sources is integrated. The most commonly used range for scaling is 0 to 1 [298], but in this case study, a value of 0 would mean absence of association. For example in this case study, the minimum value of normalized co-occurrence is 0.1538 (association between NAFLD and infant food). If this value is scaled on the range 0-1, it would convert to 0, which will lead to an ambiguous record. Moreover, the co-occurrences retrieved vary from very small values to very large values. For instance, the maximum value of normalized co-occurrence is 274.65 (association between T2DM and dietary fiber). A range from 0 to 1 would not be a good fit for such a vast difference between minimum and maximum values. Thus, a broader range not starting from 0 i.e. a range from 1 to 10 has been selected for this purpose. This implies association between NAFLD and infant food is taken as 1, T2DM and dietary fiber as 10 and other values lie between 1 to 10. As depicted in the figure, T2DM and NAFLD are mutually related to beer, coffee, energy drinks, carbonated beverage and many other diet terms. Further, Covid-19 node along with relationships between the three diseases are integrated with the graph.

4.3.1.3 Prediction of Associations using Prediction Framework

In this step, inferences are drawn using Prediction Framework based on TBM learning method. It has been designed using algorithms namely Louvain algorithm (LA) and KNN using PR. These two algorithms (LA and KNN) are selected because they utilize different learning properties i.e. unsupervised and supervised respectively to infer

Figure 4.4: A Subgraph of Retrieved Covid-19 related Diseases and Diets Associations Graph

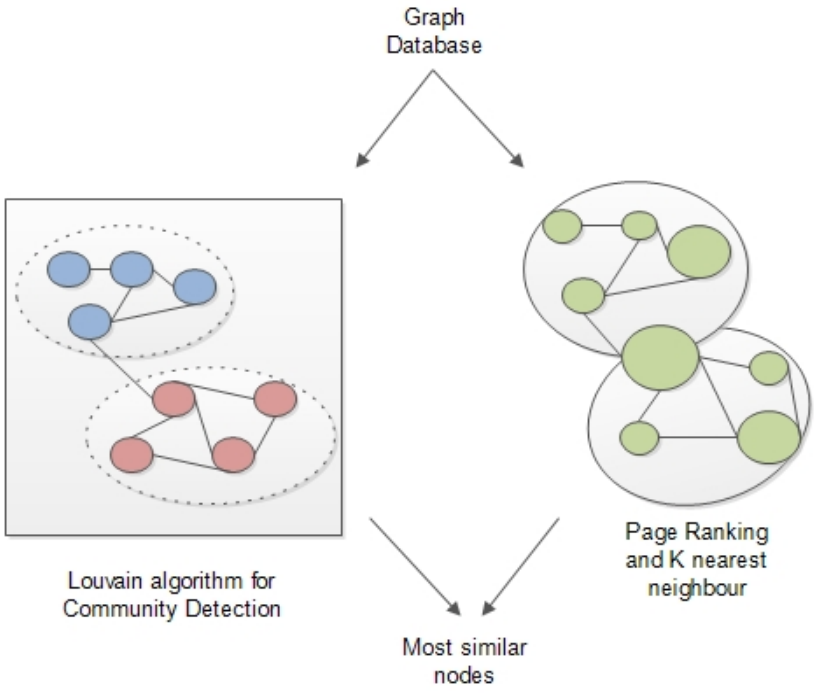


Figure 4.5: Steps for Prediction of Diet Associations for Covid-19 and other Diseases (Bigger size of node depicts a better rank in Page Ranking Algorithm)

relations. Initially, Louvain algorithm is used to divide the graph database into communities. Since, only two communities are identified in this step, another algorithm KNN is introduced to further refine the analysis. In KNN, Page Rank algorithm is applied to rank all the nodes in graph. This is done so that the ranks can be further used for finding similarity. Thus, LA finds communities in the graph whereas KNN fetches the top most related nodes in the graph. The steps performed are discussed as follows (also shown in Figure 4.5):

i) Louvain algorithm is selected for this step because it has no prior assumptions regarding community of nodes and is specifically designed for graphs. Graph retrieved from the above steps is stored in Neo4j and Louvain algorithm is run using cypher queries. Two communities are identified containing a total of 140 nodes.

ii) KNN algorithm is selected because it is a simple technique with good accuracy and less runtime, which is suitable for small dataset in this study. It finds top K similar nodes for each node using a similarity function. Similarity function is calculated using a given property of nodes in the graph. In this study, page rank of nodes is taken as the given property based on the principle that more connected nodes are more similar. Since page rank is a float value in this study, cosine similarity is used as similarity function by KNN. A higher sample rate increases the accuracy of KNN, thus it is chosen as 0.8 along with delta threshold of 0.001. Different variations of algorithm are performed with value of K ranging from 2 to 20.

4.3.2 Evaluation of Results

The two communities identified using Louvain algorithm contain 46 and 94 nodes respectively. One of the communities contains NAFLD and Covid-19 terms along with 44 diet terms. The second community contains 93 diet terms related to T2DM. The disease and diet nodes retrieved in different communities are described in Table 4.2. Initially two topmost similar nodes are retrieved, but as the value of K increases, other most similar nodes are added as described in Table 4.3. The obtained associations are evaluated in the following manner:

- *Validation through PubMed literature:* The retrieved associations have been val-

idated by searching PubMed research papers with respective diet terms. Table 4.4 represents important associations validated from Pubmed literature. All the associations predicted for T2DM have been identified from literature as shown in Table 4.4 (records for terms "sweetening agents", "flavoring agents" and "food additives" are displayed collectively because of same literature involved). Interestingly, most of the diets associated with NAFLD have also been identified except bread, dietary fats unsaturated and edible grains. Being a novel disease, research related to Covid-19 is in its infancy, thus lesser number of diets have been validated for it. The diets predicted and validated for Covid-19 risk management include egg yolk, celery, sesame oil, strawberry, raspberries, honey, carrot, kefir and onion. The predicted associations which could not be validated include food items like beets, watermelon, shellfish, cucumber, egg white, pumpkin, brazil nuts, infant formula, raw food, peach and carbonated water.

- *Evaluation through Precision:* Evaluation for this study is done using precision as a measure due to the nature of database. Being a prediction from dietary database and relating directly to human health, precision has utmost importance in this case. A high precision might mean returning less results, but most of it is correct. This kind of database cannot afford low precision with high number of incorrect results. Precision is calculated as ratio of true positives and all the positives where true positives refer to actual associations (confirmed using PubMed literature validation) and all positives refer to the predicted associations. Since relation between Covid-19 and diets is a very fresh domain, validation of its corresponding results in literature is difficult. If only NAFLD and T2DM are considered, then precision is around 92.5% as shown in Table 4.5. The table depicts different combinations of data samples of results and their corresponding precision. The precision with all the data samples (76.7%) is also quite interesting because of the presence of results related to a novel disease.

Table 4.2: Disease and Diet Nodes Detected in Different Communities

Community Detection	**Diseases**	**Diets**
1	Covid-19 and NAFLD	Beer, Carbonated Water, Kefir, Infant Formula, Soy Milk, Chocolate, High Fructose Corn Syrup, Corn Oil, Sesame Oil, Soybean Oil, Egg Proteins Dietary, Egg White, Egg Yolk, Infant Food, Honey, Poultry, Shellfish, Raw Foods, Apple, Avocado, Beets, Blueberries, Brazil nuts, Cabbages, Carrot, Celery, Coconut, Corn, Cranberries, Cucumber, Grapes, Citrus, Orange, Mushroom, Onion, Peach, Raspberries, Strawberry, Tomato, Walnut, Watermelon, Plum, Pumpkin, Apricot
2	T2DM	Wine, Carbonated Beverages, Coffee, Energy Drinks, Buttermilk, Milk (Human), Whey, Tea, Teas Herbal, Bread, Candy, Chewing Gum, Condiments, Spices, Edible Grain, Whole Grains, Dairy Products, Butter, Ghee, Cultered Milk Products, Cheese, Yogurt, Ice Cream, Margarine, Milk Proteins, Whey Proteins, Dietary Carbohydrates, Dietary Sucrose, Dietary Fats, Dietary Fats Unsaturated, Cottonseed Oil, Olive Oil, Safflower Oil, Dietary Fiber, Vegetable Proteins, Dietary Supplements, Eggs, Fast Foods, Flour, Food Additives, Fat Substitutes, Flavoring Agents, Sweetening Agents, Non-Nutritive Sweeteners, Nutritive Sweeteners, Food Preservatives, Fruit, Meat, Meat Products, Poultry Products, Red Meat, Seafood, Fish Products, Vitamins, Provitamins, Nuts, Seeds,Vegetables, Vegetable Products, Soy Foods, Soybean Proteins, Almonds, Asparagus, Banana, Buckwheat, Cantaloupe, Cashew, Cherry, Coriander, Pepper, Lettuce, Mango, Millets, Mustard, Oats, Pear, Peas, Pecans, Pineapple, Pistachio nuts, Potato, Radish, Spinach, Sweet potato, Wheat, Barley, Cloves, Hazelnut, Jackfruit, Kidney beans, Papaya, Rice, Turnip

Table 4.3: Associations Identified using K Nearest Neighbours and Page Rank Algorithms

S. No.	**Value of K**	**Source Node**	**Predicted K nearest nodes**
1		T2DM	Dietary Fat, Flavoring Agents
2	2	NAFLD	Sweetening Agents, Flavoring Agents
3		COVID-19	Egg yolk, Celery
4		T2DM	Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives
5	5	NAFLD	Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives
6		COVID-19	Egg yolk, Celery, Beets, Sesame oil, Watermelon
7		T2DM	Dietary Fats, Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives, Dietary Sucrose, Nutritive Sweeteners, Dietary Supplements, Vegetables
8	10	NAFLD	Dietary Fats, Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives, Dietary Sucrose, Nutritive Sweeteners, Dietary Supplements, Vegetables
9		COVID-19	Egg yolk, Celery, Beets, Sesame oil, Watermelon, Peach, Carrot, Strawberry, Raspberries, Shellfish
10		T2DM	Dietary Fats, Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives, Dietary Sucrose, Nutritive Sweeteners, Dietary Supplements, Vegetables, Whole grains, Bread, Dietary Fat Unsaturated, Fruits, Red Meat, Coffee, Nuts, Edible grains, Meat, High Fructose Corn Syrup
11	20	NAFLD	Dietary Fats, Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives, Dietary Sucrose, Nutritive Sweeteners, Dietary Supplements, Vegetables, Whole grains, Bread, Dietary Fat Unsaturated, Fruits, Red Meat, Coffee, Nuts, Edible grains, Meat, High Fructose Corn Syrup
12		COVID-19	Egg yolk, Celery, Beets, Sesame oil, Watermelon, Peach, Carrot, Strawberry, Raspberries, Shellfish, Brazil nuts, Pumpkin, Infant Formula, Raw foods, Cucumber, Egg white, Carbonated water, Kefir, Onion, Honey

Table 4.4: Validation of Predicted Associations for Diseases using Pubmed Literature

References	Disease	Diet	Association
[299, 300, 301, 302, 303, 304, 305]	T2DM	Bread	Modified bread/ whole grain bread helps in management. Bread intake is associated with blood glucose.
[222, 223, 306]	T2DM	Edible grains	Decrease of risk found with increase of consumption of whole grains. Refined grains are found to be associated with increased risk.
[223, 307, 308, 309]	T2DM	Vegetables	A juice consisting spinach, celery, chickpea, broccoli, and green beans is found to helpful for diabetic patients.A careful selection of vegetables have been found to be helpful for preventing the disease.
[306, 310, 311]	T2DM	Meat	Processed meats have been found to be related to increased risk.
[312, 313]	T2DM	Dietary fat unsaturated	Polyunsaturated fatty acid might be protective towards risk or management.
[222, 223, 301]	T2DM	Whole grains	Protective effect of high consumption of whole grains.
[224, 314, 315, 316]	T2DM	Coffee	Brewed coffee is found to be helpful for risk prevention.
[223, 309, 315]	T2DM	Fruits	Polyphenols found in fruit might be helpful for risk prevention.
[309, 315]	T2DM	Nuts	Polyphenols found in nuts might be helpful for risk prevention.
[223, 317, 318]	T2DM	Red meat	More consumption (specially processed) might increase risk.
[319, 320]	T2DM	Dietary Fats	Intake of Saturated and Trans Fatty Acids should be decreased whereas a balance between omega 3 and omega 6 fatty acid helps in management
[321, 322, 323]	T2DM	Dietary Fibers	Dietary fiber including cereal fiber is protective against T2DM
[324, 325]	T2DM	Dietary Carbohydrates	Low carbohydrates ketogenic diets can be helpful in prevention and reversal, yet carbohydrate intake should be individualized
[326, 327, 328]	T2DM	Sweetening/ Flavoring Agents and Food additives	Sugar sweetened beverages and flavouring agents increase risk of disease
[329]	T2DM	Dietary Sucrose	It has been found to be associated with decreased risk in a meta analysis

Continued on next page

Table 4.4 – *(Continued)*

References	**Disease**	**Diet**	**Association**
[327]	T2DM	Nutritive Sweeteners	Natural sugars have not been accounted for disease risk
[330, 331]	T2DM	Dietary Supplements	In a review, enough evidence was not found for benefits of supplements but Vitamin D supplements have been found to be helpful in some studies
[328, 332]	T2DM	High Fructose Corn Syrup	High fructose corn syrup present in sugar sweetened beverages is found to associated with increased risk
[333, 334, 335, 336]	NAFLD	Whole grains	Beneficial for management and risk prevention.
[333, 334, 335, 336]	NAFLD	Vegetables	Beneficial for management and risk prevention.
[337, 338, 339]	NAFLD	Coffee	It might be helpful in risk prevention and management.
[333, 334, 335, 336]	NAFLD	Fruits	Beneficial for management and risk prevention.
[333, 336]	NAFLD	Nuts	Beneficial for management and risk prevention.
[340, 341]	NAFLD	Red Meat	High consumption might increase risk.
[342, 343]	NAFLD	High fructose corn syrup	Might increase risk of disease.
[340]	NAFLD	Meat	Processed meat might increase risk.
[344, 345]	NAFLD	Dietary Fats	Saturated fat may lead to development of disease.
[346, 347, 348]	NAFLD	Dietary Fibers	Beneficial for management.
[349]	NAFLD	Dietary Carbohydrates	Dietary carbohydrates specially fructose might be involved in development
[327, 342, 350]	NAFLD	Sweetening/ Flavouring Agents and Food additives	Reducing added sugars and sugar sweetened beverages might be helpful in management.
[342]	NAFLD	Dietary Sucrose	Role in increased risk
[351]	NAFLD	Nutritive sweeteners	Natural sweeteners like stevia might have protective effects

Continued on next page

Table 4.4 – *(Continued)*

References	Disease	Diet	Association
[352, 353]	NAFLD	Dietary Supplements	Certain supplements and herbs might be helpful
[354]	Covid-19	Kefir	It is protective agent against viruses, thus might be considered for Covid.
[355]	Covid-19	Carrot	Fresh carrot juice is suggested and further being explored for preventing damage of multiple organs in case of Covid.
[356, 357]	Covid-19	Onion	Due to its therapeutic property, it might be considered for Covid.
[358, 359, 360, 361]	Covid-19	Egg yolk	Egg yolk antibodies are found to be preventive against Covid-19
[355]	Covid-19	Celery	Fresh juice from celery is suggested and further being explored for preventing damage of multiple organs
[362]	Covid-19	Sesame oil	Due to its high linoleic acid concentration, sesame oil might be helpful for protection
[363]	Covid-19	Strawberry	Strawberry might present inhibitory potential
[364]	Covid-19	Raspberries	Leucoefdin found in raspberry present inhibitory potential
[365, 366]	Covid-19	Honey	Due to its anti-viral properties, it might be helpful

Table 4.5: Precision for Different Data Samples of Results

Data Sample	**True positives**	**All positives**	**Precision**
T2DM	20	20	100%
NAFLD	17	20	85%
T2DM and NAFLD	37	40	92.5%
T2DM, NAFLD and Covid-19	46	60	76.7%

4.3.3 Insights

Important inferences drawn from the retrieved results are described as follows:

- Similar diets have been predicted for T2DM and NAFLD which confirms the overlap seen in Figure 4.6. Diet terms including almonds, cashew, mustard etc. are clearly visible in one cluster with T2DM whereas pumpkin, apricot etc. are visible in other cluster with Covid-19 and NAFLD. Certain diets are also visible in the overlapping of both clusters including whole grains, nuts etc. The diets include Dietary Fats, Dietary Fibre, Dietary Carbohydrates, Sweetening Agents, Flavoring Agents, Food Additives, Dietary Sucrose, Nutritive Sweeteners, Dietary Supplements, Vegetables, Whole grains, Bread, Dietary Fat Unsaturated, Fruits, Red Meat, Coffee, Nuts, Edible grains, Meat and High Fructose Corn Syrup. All the predicted diets have been validated from literature for T2DM. This is due to the existence of a substantial research regarding association of T2DM with diets. In case of NAFLD, all the diets except bread, dietary fats unsaturated and edible grains have been validated.

- It is evident from the results and validations that diets like whole grains, vegetables, fruits and nuts are found to be helpful for both NAFLD and T2DM diseases. The reason behind this relation lies in the fact that many researchers suggest Mediterranean Diet (MD) for prevention and management of chronic diseases like T2DM and NAFLD [333, 367]. Mediterranean Diet corresponds to more consumption of plant based foods and fish but less consumption of dairy products and meat, which very well agrees with the results obtained.

- The top 20 predicted diets for Covid-19 include Egg yolk, Celery, Beets, Sesame

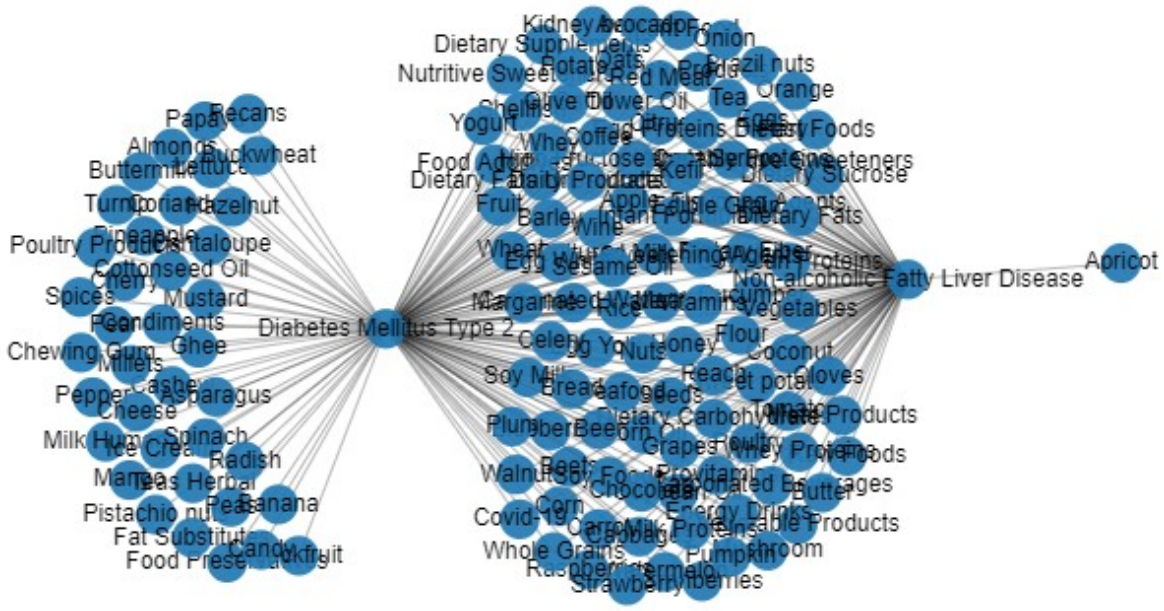


Figure 4.6: Important Communities Identified using Louvain Algorithm

oil, Watermelon, Peach, Carrot, Strawberry, Raspberries, Shellfish, Brazil nuts, Pumpkin, Infant Formula, Raw foods, Cucumber, Egg white, Carbonated water, Kefir, Onion, Honey. Out of these, 9 associations have been validated including egg yolk, celery, sesame oil, strawberry, raspberries, honey, carrot, kefir and onion. Validated diets for Covid-19 imply that they are currently being inspected for future prospectives of disease prevention and management. Food items containing Vitamin C and A like strawberry and carrot respectively along with those exhibiting antiviral properties like kefir are found in the list of validated diets.

- The diet-based associations predicted for Covid-19 are depicted in Figure 4.7 along with their similarity scores. The most similar diets as seen in the figure include egg yolk, celery, beets, sesame oil and watermelon. Out of these 5 diets, 3 diets including egg yolk, celery and sesame oil have been validated in literature. These kinds of validations are interesting, considering the novelty of disease.

- Considering the high similarity scores retrieved for diets with Covid-19 (refer Figure 4.7), associations which could not be validated in literature should be studied for further research. These include beets, watermelon, shellfish, cucumber, egg

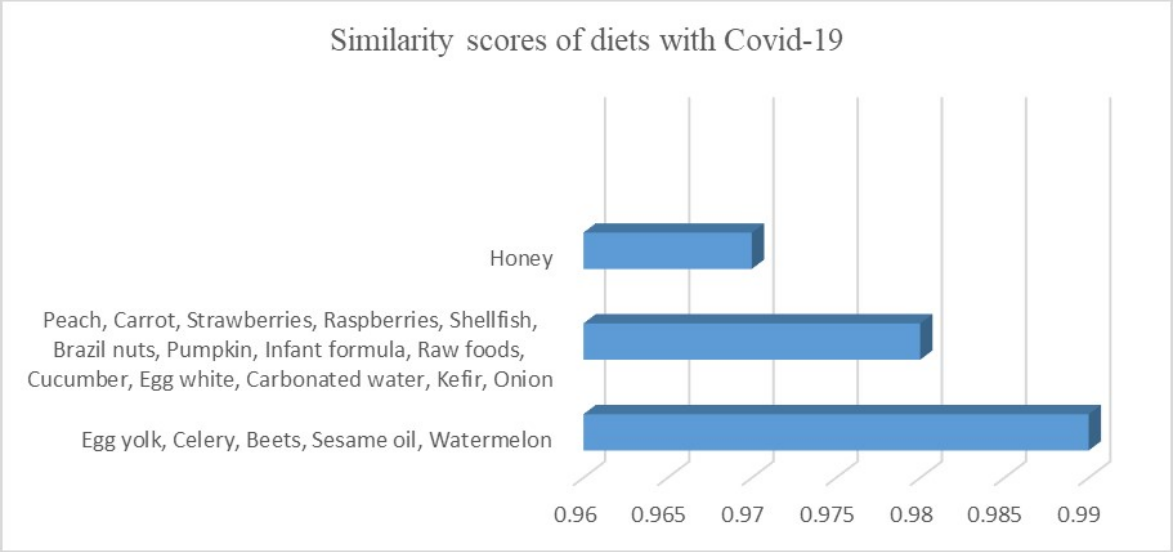


Figure 4.7: Top 20 Associations Identified for Covid-19 along with Similarity Scores

white, pumpkin, brazil nuts, infant formula, raw food, peach and carbonated water which must be considered for research in this time of pandemic. The significance of associations can be confirmed by performing randomized trial or cohort study. Similarly, associations that could not be validated for NAFLD including bread, dietary fats unsaturated and edible grains should also be inspected for future associations.

- Covid-19 and NAFLD are seen in the same community, thus the diets present in the community must be explored for utilizing them in case of patients suffering from both the diseases. The confirmed associations can then be used for planning a better diet for patients suffering from Covid-19, T2DM or NAFLD or combinations of these. This approach can help to expedite the process of testing different diets for Covid-19 and act as a baseline to discover dynamics of Covid-19 with different comorbidities as per patients' health status.

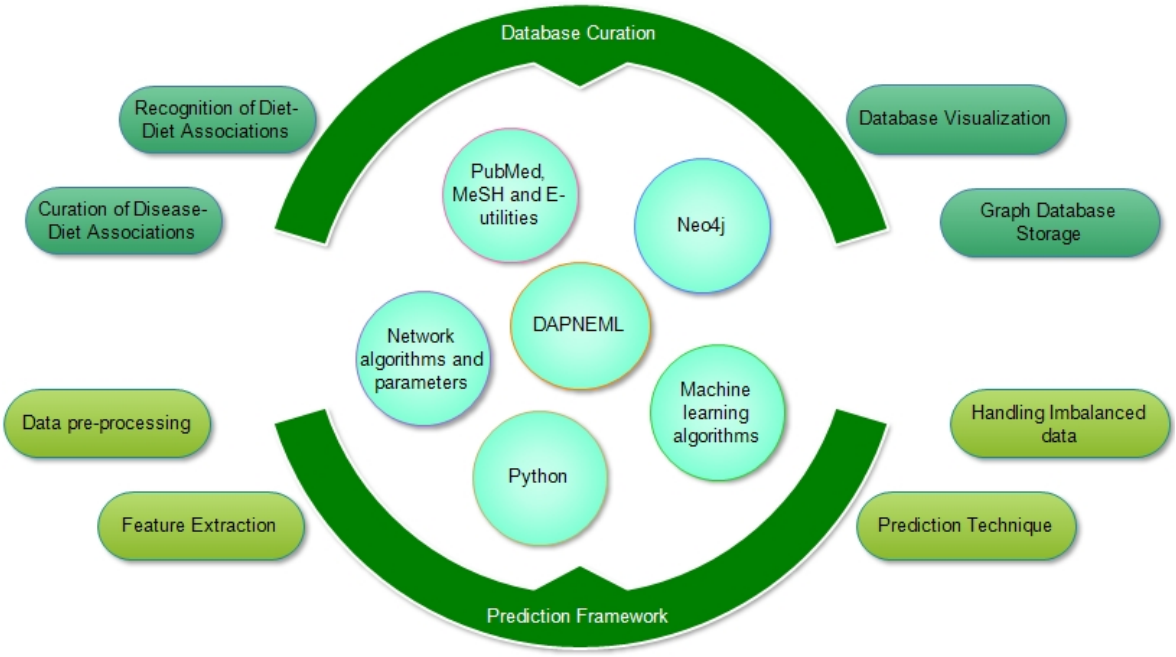


Figure 4.8: Elements of Proposed Approach PredNEM

4.4 Experimental Validation of Case Study II: IBD

4.4.1 PredNEM for Case Study II

Understanding the dynamics of the disease and its subtypes for diets are undertaken in this case study. Moreover, validating a well known diet-related disease like IBD is feasible via medical literature and a certified dietician. Thus, IBD and its subtypes CD and UC have been selected for this case study. PredNEM aims to utilize disease-diet and diet-diet relationships for prediction. The purpose is to predict unknown associations for CD by utilizing already known diet associations of IBD and UC. Elements of the proposed approach PredNEM are described in Figure 4.8. The steps involved in PredNEM are discussed in the subsequent sections.

4.4.1.1 Integration of Required Networks

Similar to the previous case study, DID-NEM has been used to extract required disease-diet network in this study as well. Database of diet-diet associations is also not as such

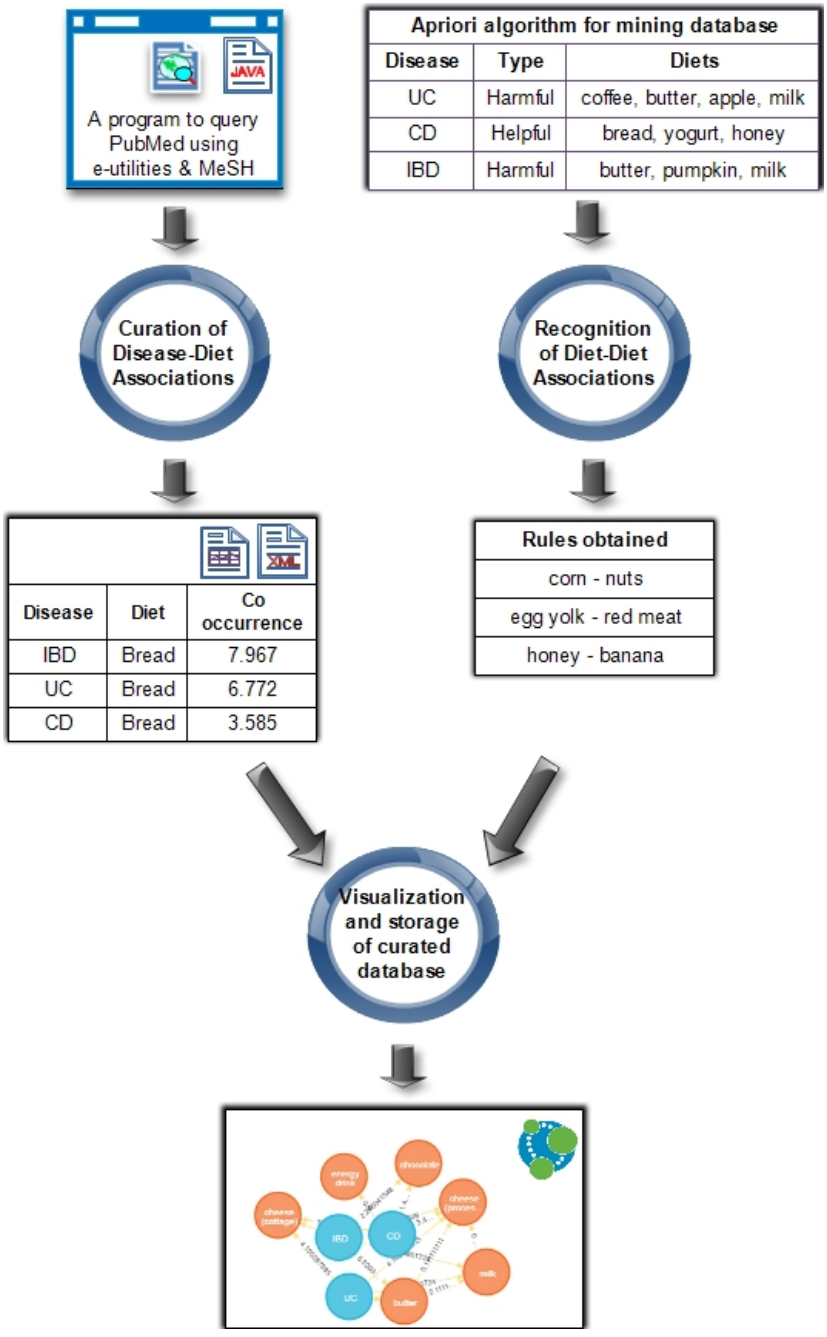
available in literature, thus associations were recognized in this study by performing pattern mining. Integration of networks containg associations corresponding to disease-diet and diet-diet has been performed as shown in Figure 4.9 and discussed as follows:

- Selection of Diets Associations with IBD and UC

  DID-NEM has been utilized for selection of diet associations with IBD and UC. The selected associations were validated in real life by a certified dietician. Along with that, PubMed was also searched with the respective terms so as to confirm the relationships. Thus, a filter of PubMed validation and dietician validation was applied to the database. Relationships satisfying any one of the conditions were taken into consideration. The improved curated database consists of 2 diseases (IBD and UC) and their 109 curated and validated relationships with different diet terms along with their normalized co-occurrences.

- Recognition of Diet-Diet associations

  Records curated in the previous step were labelled by reading PubMed papers. The labels include little harmful, harmful, very harmful, neutral, little helpful, helpful and very helpful. The labelled dataset was validated from a certified dietician and then transformed such that all the diets which are related to a disease in the same way (i.e. same label) form one record, for example diets helpful for IBD form one record of the database, as shown through a snapshot in Figure 4.9. This was done so as to apply Apriori algorithm to these records. It is an algorithm for creating association rules from a dataset using frequent items [368]. It obeys an iterative procedure to uncover associations between items. 5304 rules have been generated using minimum confidence=0.5. Redundant associations were removed by using a function (is.redundant) in R. A rule is said to be redundant if a more general rule with the same or higher confidence exists. This significantly reduced the number of rules with a total of 86. Some of the rules were further eliminated if they consisted of two or more terms on either side of the relation. This was done to reduce complexity and avoid redundancy. Thus, only diet pairs were considered for this purpose. Hence, a total of 20 significant



Apriori algorithm for mining database

Disease	Type	Diets
UC	Harmful	coffee, butter, apple, milk
CD	Helpful	bread, yogurt, honey
IBD	Harmful	butter, pumpkin, milk

Disease	Diet	Co occurrence
IBD	Bread	7.967
UC	Bread	6.772
CD	Bread	3.585

Rules obtained
corn - nuts
egg yolk - red meat
honey - banana

Figure 4.9: Overview of Steps for Integration of Required Networks

rules were retrieved. The weight of retrieved associations is also defined using tf-idf to create a standard measure of edges in the graph. In the case of diet-diet associations, there is no need to weigh the term frequency with inverse document frequency because every diet term is considered equally significant to any other diet term. Thus, term frequency which is the same as support value is taken as the weight of edges.

4.4.1.2 Validation, Visualization and Storage of Associations

Only validated diet associations of IBD and UC were considered in the study as discussed earlier. The retrieved diet-diet associations were also validated in PubMed literature which include:

1. Corn and nuts are both harmful for Eosinophilic esophagitis as suggested in [369, 370].
2. Corn and nuts, both are known to be helpful for cholesterol [371].
3. Corns and nuts are also harmful for IBD [372].
4. Egg yolk and red meat are found to be harmful for renal impairment [373, 374]
5. Egg yolk and red meat are found to be harmful for stroke as well [374].
6. Honey and banana, both are helpful for hypertension [375].

Both kinds of associations (disease-diet and diet-diet) have been stored in Neo4j in the last step as shown in Figure 4.9. A closer look at the graph database is provided in Figure 4.10. The figure represents curated associations of IBD and UC with diet terms along with the associations realized between different diets. As depicted in the figure, IBD and UC are mutually related to apple, coffee and milk and similarly other curated relations are evident. The weight of an edge in this graph depicts the tf-idf value of the association.

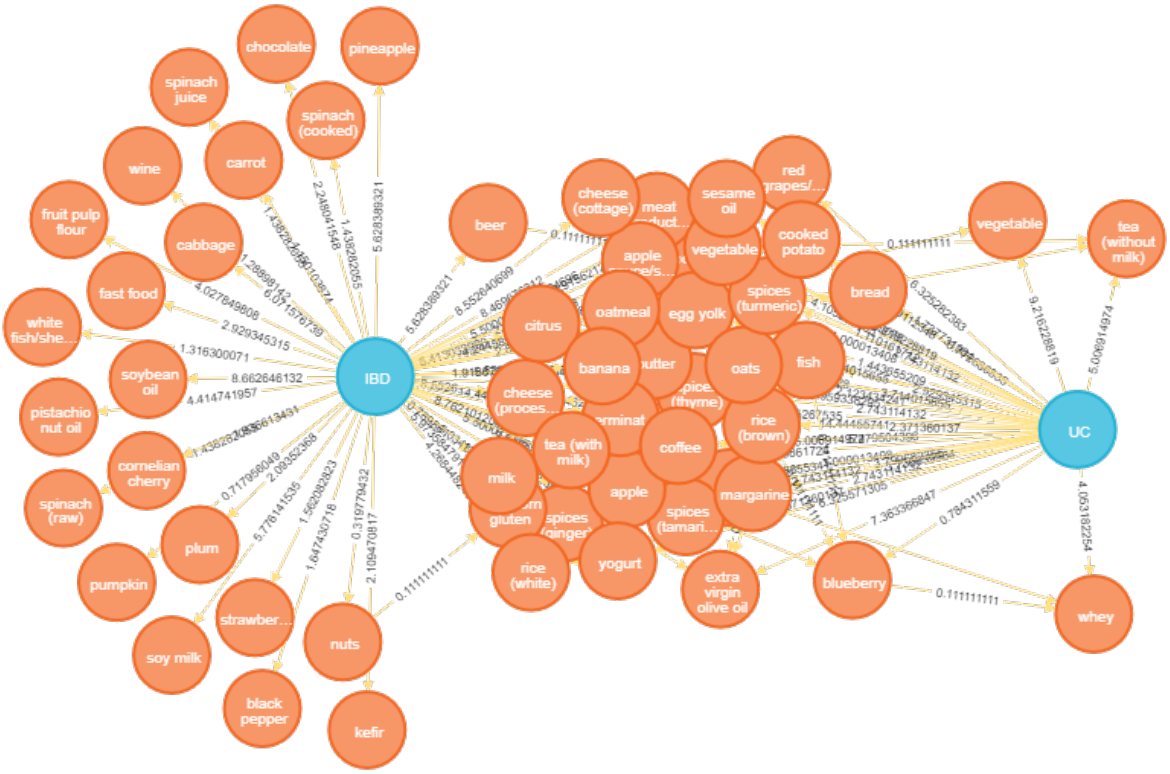


Figure 4.10: Visualization of Curated Disease-Diet and Diet-Diet Associations Corresponding to IBD, UC and Diets)

4.4.1.3 Prediction of Associations using Prediction Framework

The prediction framework utilizes TFM for analysis. It has been divided into two phases, with the first phase predicting the possibility of relation between diseases and diets and second phase predicting the nature of relation (harmful or helpful). The steps followed in both the phases are similar and are also depicted in Figure 4.11. Same steps are followed in both the phases but on different databases as discussed below:

i) Creation of labels

The curated database contains diseases, diets and their co-occurrences. Distribution of co-occurrences between diseases (IBD and UC) and diets are shown in Figure 4.12. IBD is connected to vegetables (both raw and soft) with the highest co-occurrence, followed by red grapes/grape juice and yogurt. The highest co-occurrence of UC is with germinated barley followed by vegetables (both raw and soft). The first step of prediction framework is preparation of data from the graph. In order to prepare the data for prediction, there is a need to have positive and negative labels. The labels

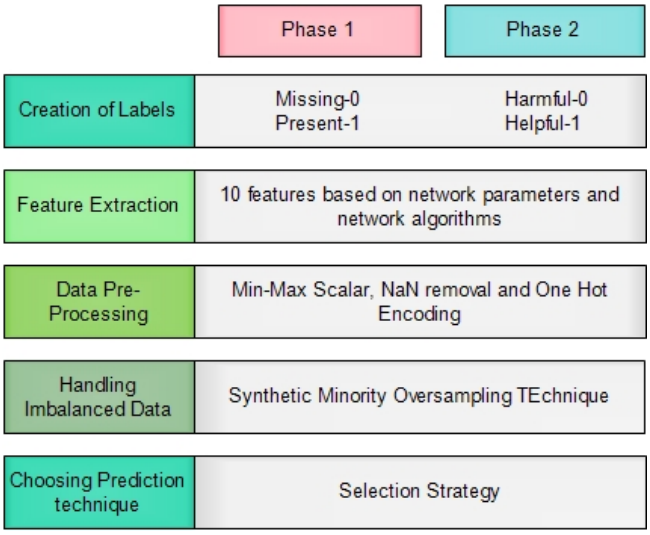
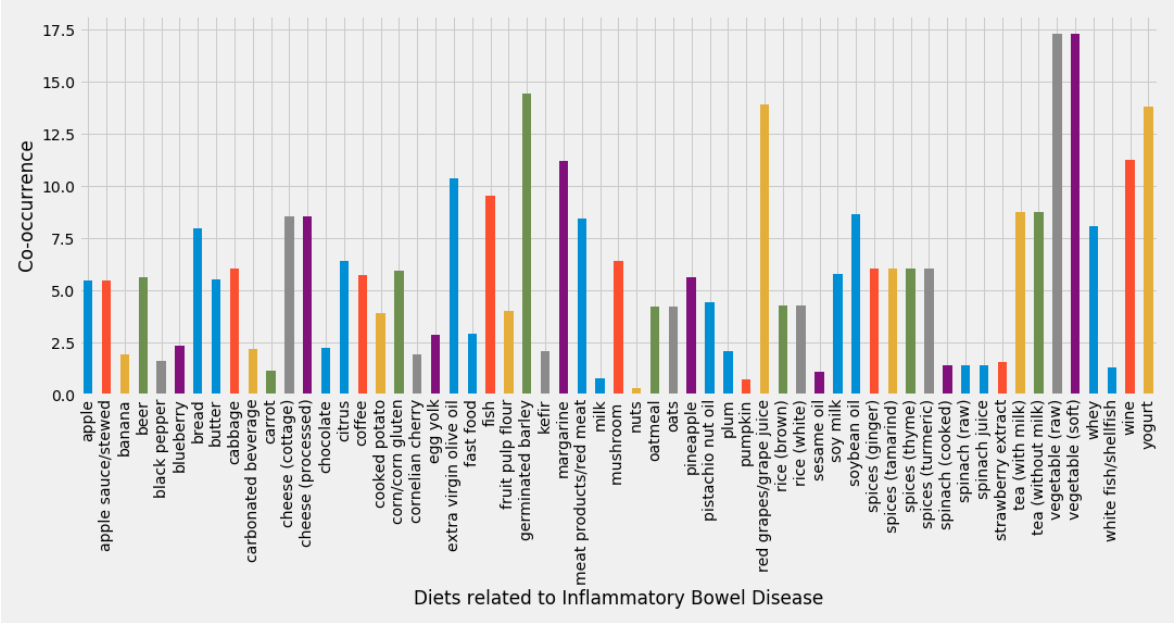


Figure 4.11: Steps of Designed Prediction Framework followed in both Phases

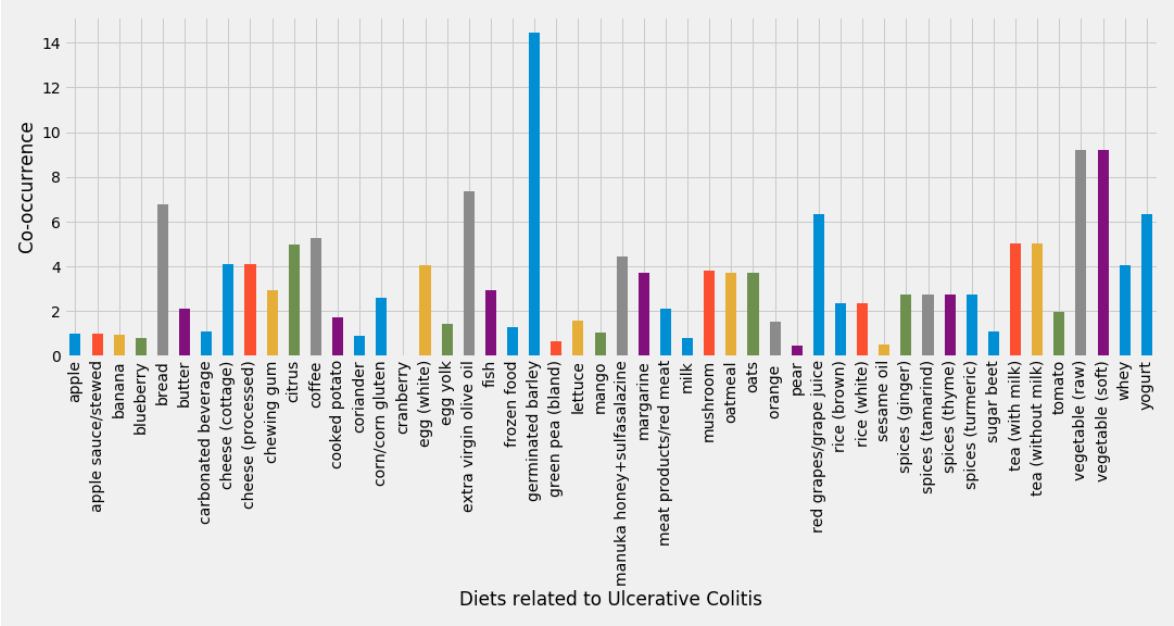
differ in each phase due to the type of prediction to be performed as discussed:

1. *Phase 1:* This phase has been designed to predict the probability of disease and diet being associated. Thus, disease-diet connections already present in the graph were labelled as positive whereas, diseases and diets not known to be connected (or missing) were labelled as negative. The distribution of relationships can be seen for each type of disease (IBD and UC) in the form of its count of present or missing links as shown in Figure 4.13. 50 diet relationships are present with UC whereas 59 with IBD. In case of missing links, there are 19 relationships that do not have any reference in literature for IBD whereas 28 in UC for the same. Thus, a total of 47 relationships are missing and 109 relationships are present, amounting to a total of 156 relationships.

2. *Phase 2:* This phase has been designed to predict the nature of diet association for disease as being harmful or helpful. The associations validated and labelled by a certified dietician were taken into consideration for this phase. Out of 109, 106 relationships pertaining to IBD and UC were taken. The remaining 3 associations were not considered because they were labelled as neutral, which does not lie in one of the two considered classes as negative (harmful) or positive (helpful). Out of the selected relationships, 75 were labelled as helpful, whereas 31 as harmful.



(a)



(b)

Figure 4.12: a) Distribution of Diet Associations for IBD b) Distribution of Diet Associations for UC

ii) Feature extraction

The purpose of using TFM is to be able to extract important features using network properties and algorithms. The features were extracted using different algorithms so

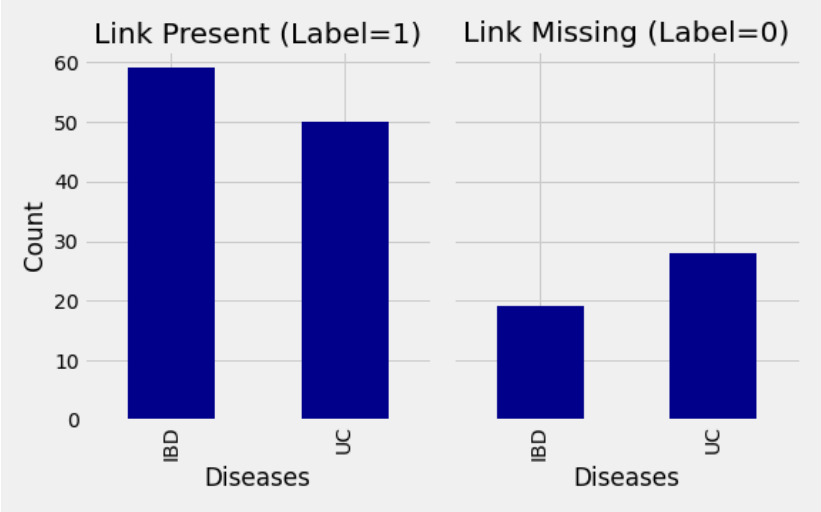


Figure 4.13: Distribution of Links (Present and Missing) for IBD and UC

as to capture multiple perspectives. Thus, there were features extracted using link prediction algorithms (common neighbours, preferential attachment, total neighbours), triangle counting, clustering coefficient, community detection algorithm (louvain algorithm) and path analytics (all pairs shortest paths). A total of 10 features have been extracted to be used in both the phases including Sum of co-occurrences (named as "Co-occurrence" in the curated database), Local Clustering Coefficient algorithm (Min Coefficient and Max Coefficient), Common Neighbours, Preferential Attachment, Total Neighbours, All Pairs Shortest Path algorithm (distance), Louvain algorithm (louvain) and Weakly connected components (cc). The feature named "Co-occurrence" is designed specifically for this database to represent a property. This property evident from the graph visualization as seen in Figure 4.10 is the presence of transitive relations. For example, in the figure, UC is connected to apple which is further also connected to IBD. IBD is also connected to pineapple but pineapple is not directly connected to UC. Presence of this transitivity implies that a relation between UC and pineapple might be likely. This feature has been extracted using a Neo4j cypher query by taking its value as the sum of cooccurrences along all the paths observed as transitive relations between two nodes, for example in the discussed scenario, value of this feature would be the sum of co-occurrences along all the transitive paths from UC to pineapple. Apart from this, two communities have been identified in this study using louvain algorithm;

thus, each pair is assigned either 1 (same community) or 0 (different communities).

iii) Pre-processing of data

In order to eliminate bias in the prediction framework due to different scales, min-max scalar has been applied to all the features (except Louvain algorithm and weakly connected components). Since, Louvain algorithm categorizes each node into either same or different community, it is considered as a categorical variable and one hot encoder has been used to encode this feature. Similarly, Weakly connected component is also categorical, thus one hot encoder has been used. Missing or NaN values of features have been converted to 0. A heatmap of extracted features as shown in Figure 4.14 displays a high correlation between total neighbours and max triangles. Common neighbours and min triangles also show high correlation. It can be seen that features based on different algorithms/ properties have low correlation. For example, all pairs shortest distance (Distance) has a low correlation with min triangles, max coefficient, min coefficient and common neighbours. Similarly, louvain algorithm (sp) has the least correlation with co-occurrence.

iv) Handling imbalanced data

The databases constructed in both the phases contain an imbalance between the number of positive and negative labels. Moreover, the extracted datasets are also small in size. Thus, a technique named Synthetic Minority Oversampling Technique (SMOTE) has been used to balance the distribution of classes [376]. In this technique, minority samples which are closer in feature space are selected and a line is fit to these points. This line is used to draw further points and generate synthetic minority samples. A python library named Imbalanced-Learn has been used for this purpose.

v) Choosing Prediction Technique

The most important step of the framework is to select the best machine learning model for which a selection strategy has been designed. The steps followed in the selection strategy are as follows:

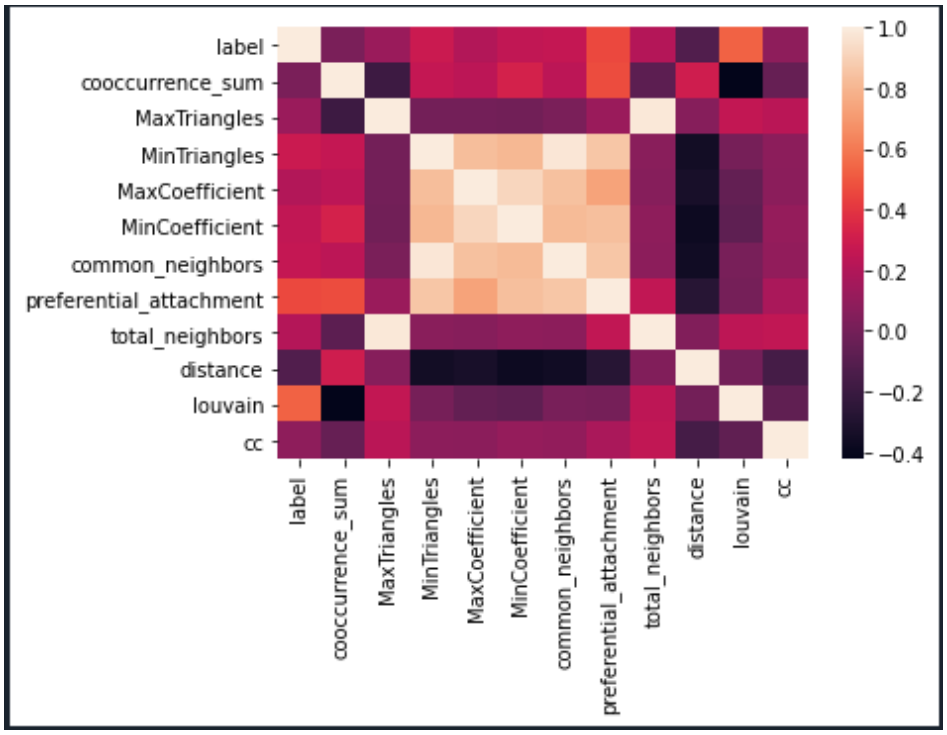


Figure 4.14: Heatmap of Extracted Features

1. *Perform a dry run of basic algorithms:* Algorithms which have different procedure of learning were selected and an initial run was performed. An 80:20 train test split ratio was used along with a stratified k fold cross validation. This was used so as to ensure same class distribution in each split of data. In this cross validation, complete dataset is shuffled and split into k folds, from which 1 fold is taken as test set and remaining as train set. It is observed that if number of folds are increased, then standard deviation of algorithms increase substantially in the first phase, whereas outliers are removed significantly in the second phase. Thus 5 folds are used in the first phase and 10 folds are used in the second phase.

2. *Select 3 best algorithms:* After the initial run, algorithms were compared and three best performing among them were selected. This was done to reduce computational cost and time. In case of imbalanced classes, ROC AUC is a good metric, thus algorithms were compared based on ROC AUC.

3. *Hyperparameter Optimization:* Parameter optimization of the selected three algorithms was performed either manually or using GridSearchCV. GridSearchCV

is a library in which different range of parameter values are input to the function and the best combination of values is selected. This aims to improve accuracy of algorithms.

4. *Run 3 optimized algorithms:* The parameters of algorithms were set as per the previous step and 3 optimized algorithms were run again.

5. *Select the best algorithm:* The best performing algorithm was selected among the three by comparing their boxplots.

6. *Perform optimization again for the selected algorithm:* The selected algorithm performs best in hyperparameter optimization but it might be overfitting to the prediction data. Thus, in this step, the selected best algorithm is again optimized so as to avoid selection of parameters leading to an overfit algorithm. For this, the pattern of results obtained with different parameters are compared as discussed in the next section.

   The selection strategy performed in both the phases is as described as follows and as in Figure 4.15.

1. *Phase 1:* In the first phase, machine learning algorithms selected for dry run included KNN, GB, SVM, NB and SE. These algorithms have been selected because they have different learning procedures. KNN has been selected due to its simplicity and its suitability for small sized data. Similarly, among CART and random forests, CART has been selected for this study because of the small number of prediction classes (associated or not) in the dataset. Random forests would have added layers of complexity to a simple and small dataset. SVM is also a relatively simple machine learning algorithm, thus selected for small dataset. A Naïve Bayes classifier works well for small dataset and Gradient boosting is known to improve performance. Thus, GB and NB were added in the selection strategy. A dry run of these algorithms retrieved values depicted in Table 4.6 along with a boxplot shown in Figure 4.16.

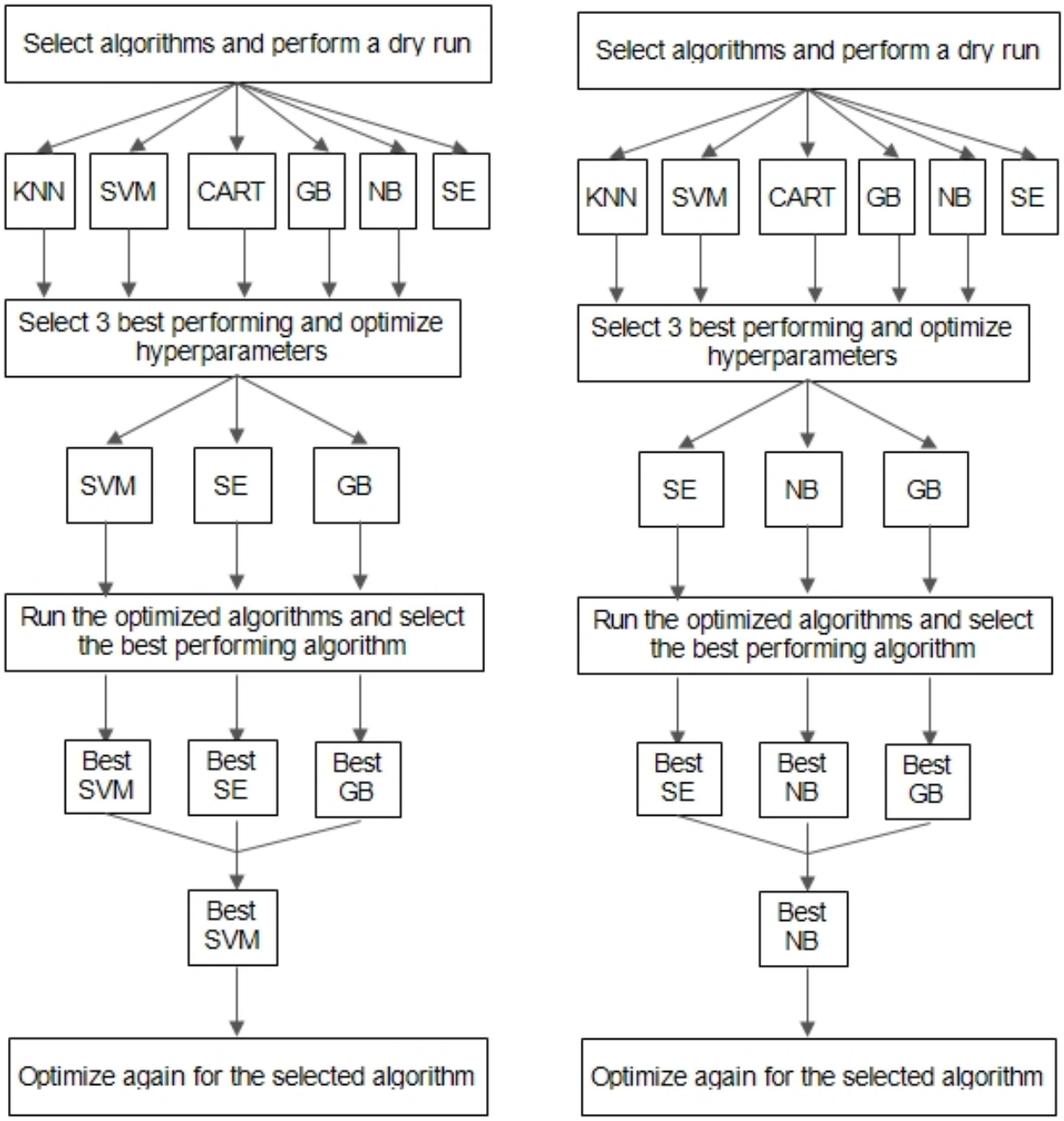


Figure 4.15: Algorithm Selection Strategy for a) Phase 1 b) Phase 2

Table 4.6: Retrieved Values and Standard Deviation from Dry Run in Phase 1

Algorithm	ROC AUC	Std
KNN	95.8	0.028
GB	98.0	0.021
SVM	97.0	0.019
NB	92.8	0.043
CART	93.4	0.055
SE	98.4	0.016

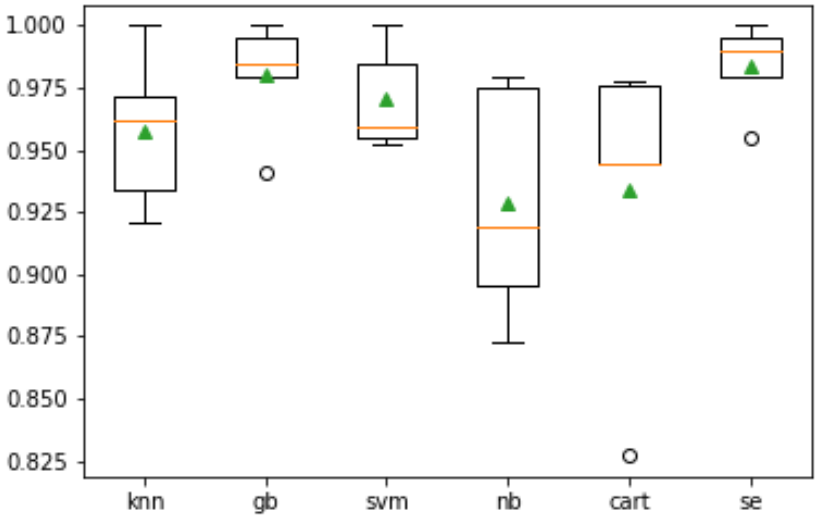


Figure 4.16: Boxplot for ROC AUC Retrieved from Dry Run in Phase 1

Table 4.7: Hyperparameter Optimization of GB using GridSearchCV in Phase 1

Parameter	Values Input	Value Selected
N estimators	10, 50, 100, 500	100
Learning rate	1, 0.1, 0.01, 0.001, 0.0001	1
Subsample	0.5, 0.7, 1	0.5
Max depth	3, 7, 9	3

Table 4.8: Hyperparameter Optimization of SVM using GridSearchCV in Phase 1

Parameter	**Values Input**	**Value Selected**
C	0.1, 1, 10, 100, 1000	10
Gamma	1, 0.1, 0.01, 0.001, 0.0001	1
Kernel	RBF	RBF

Table 4.9: Retrieved Values and Standard Deviation after Optimization in Phase 1

Algorithm	**ROC AUC**	**Std**
GB	97.1	0.039
SVM	97.3	0.038
SE	97.6	0.039

SVM, GB and SE were further selected due to better performances. Hyperparameters were tuned for SVM and GB using GridSearchCV, whereas for SE, optimization was done using tuned SVM and GB as base models. For GB, parameters were optimized with the range of parameters as depicted in Table 4.7. The best value is achieved with the selected parameters shown in the table. In SVM, optimization was performed and the values of parameters are as shown in Table 4.8. ROC AUC achieved after optimization are shown in Table 4.9 along with boxplot representation as Figure 4.17.

Both SVM and SE achieve similar values and can be used as final classifier but SVM is selected in this study because it is a relatively simple algorithm with less computational costs. In order to avoid overfitting of algorithm, a manual hyperparameter optimization is also performed. Figure 4.18 depicts the ROC AUC achieved with different values of C and gamma for SVM. As seen in the graph, most good values lie in the region with small values of gamma and C with values around 100. Thus, value of C is selected as 100 and gamma as 0.001 for the final model.

2. *Phase 2:* In the second phase, same algorithms are run in the first step. A dry run of the algorithms retrieved values depicted in Table 4.10 along with boxplot in Figure 4.19. The best performing algorithm among these are GB, CART, SE

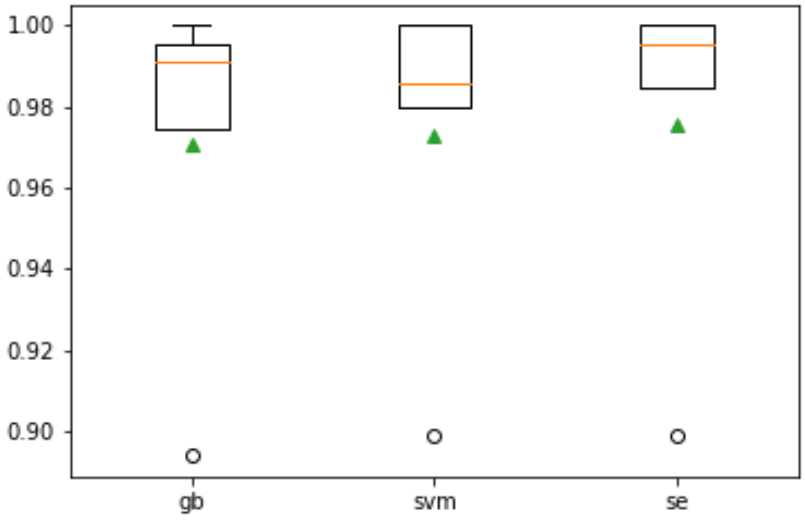


Figure 4.17: Boxplot for ROC AUC Retrieved after Optimization in Phase 1

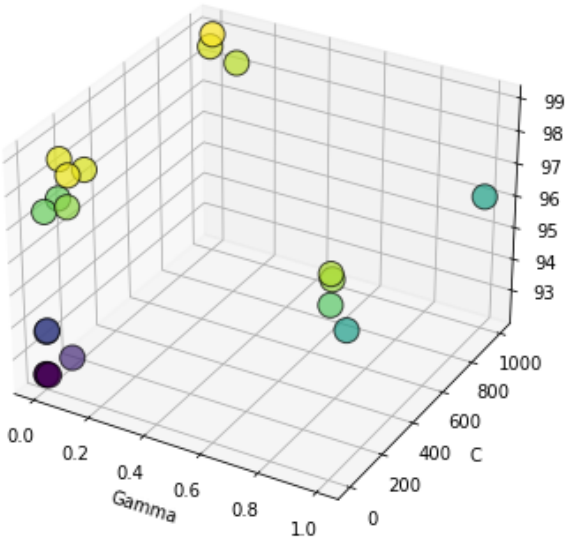


Figure 4.18: ROC AUC Retrieved with Different Values of C and Gamma for SVM algorithm

Table 4.10: Retrieved Values and Standard Deviation from Dry Run in Phase 2

Algorithm	**ROC AUC**	**Std**
KNN	66.0	0.120
GB	76.3	0.080
SVM	71.6	0.088
NB	72.5	0.069
CART	72.5	0.123
SE	78.3	0.074

Table 4.11: Hyperparameter Optimization of GB in Phase 2

Parameter	**Values Input**	**Value Selected**
N estimators	10, 50, 100, 500	100
Learning rate	1, 0.1, 0.01, 0.001, 0.0001	0.1
Subsample	0.5, 0.7, 1	0.5
Max depth	3, 7, 9	3

and NB. CART and NB represent similar values, but standard deviation of NB is lesser, thus it is selected. The selected algorithms are further optimized by tuning the hyperparameters. The optimizations retrieved for GB is depicted in Table 4.11. In case of NB, values were passed from 0 to -9 on a log scale to retrieve 100 samples for variable smoothing parameter. These samples were then fit with stratified cross validation technique. The best value achieved for variable smoothing is 0.23109. Values achieved after running optimized algorithms are as shown in Table 4.12 along with boxplot in Figure 4.20. Although, SE and GB perform better, but they contain outliers. Thus, NB is selected among the three as the predictive model for this phase. Manual optimization of the algorithm is performed next so that it does not overfit.

Table 4.12: Retrieved Values and Standard Deviation after Optimization in Phase 2

Algorithm	**ROC AUC**	**Std**
SE	78.3	0.108
NB	71.6	0.097
GB	76.5	0.108

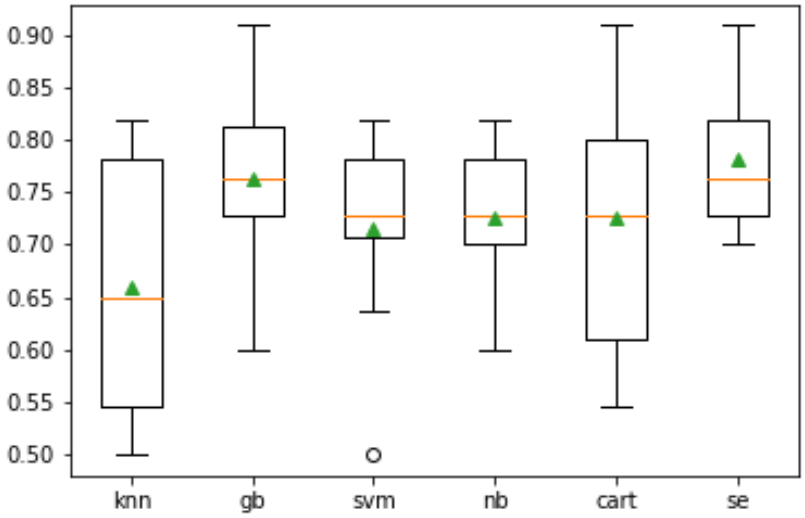


Figure 4.19: Boxplot for ROC AUC Retrieved from Dry Run in Phase 2

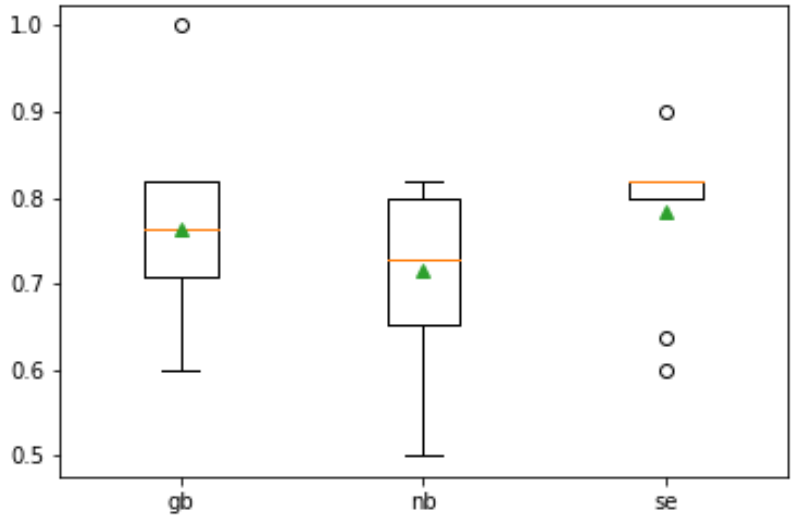
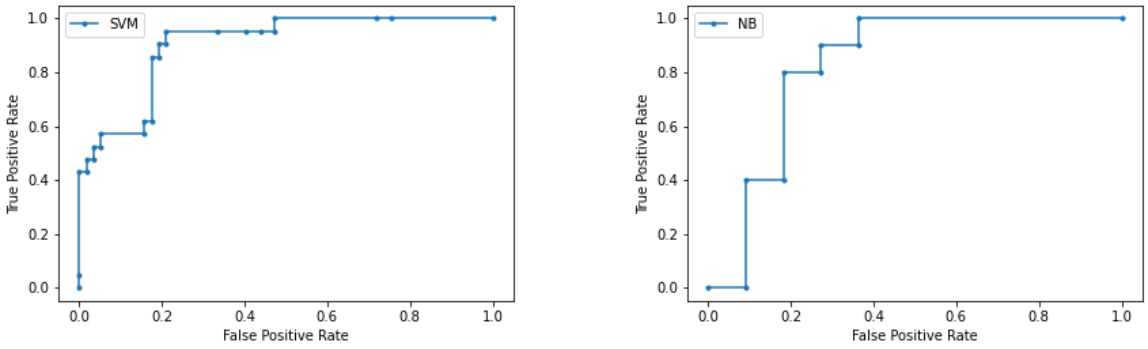


Figure 4.20: Boxplot for ROC AUC Retrieved after Optimization in Phase 2



(a) (b)

Figure 4.21: ROC Curve for a) Phase 1 b) Phase 2

4.4.2 Evaluation of Results

SVM model has been selected in the first phase to predict existence of associations between CD and diet terms. From a total set of 78 diet terms, 30 terms have been predicted to be related to CD considering a probability threshold value of 85%. Out of these 30 diets, 23 diet terms are found to have prediction probability greater than 95% and 7 with probability lying between 85-95% as shown in Table 4.13. The results were validated by using metrics including ROC curve, F1 score and accuracy. Due to the nature of this work, false positives are dangerous than false negatives because they will mislead the domain experts in decision making. Thus, the results were also shared with the certified dietician for validating the precision in real life. Each of the metric has been discussed as follows:

- *ROC curve and ROC AUC:* ROC curve is a plot between true positive rate and false positive rate for different thresholds. True positive rate is computed by dividing count of true positives by the sum of count of true positives and false negatives. False positive rate is computed by dividing count of false positives by the sum of count of false positives and true negatives. The ROC curve for both the phases is depicted in Figure 4.21. AUC of this curve is 90.2% and 82.7% in first and second phase respectively. This implies the model works good in separating the two classes in both the phases.

Table 4.13: Predicted Associations of Diets with CD

No. of Diets	**Prediction Probability**	**Diet terms**	**Validation as per Dietician**
23	Above 95%	Carbonated Beverage	✓
		Margarine	✓
		Energy Drink	✓
		Yogurt	✓
		Whitefish/shellfish	✓
		Mushroom	✓
		Whey	✓
		Corn/corn gluten	✓
		Nuts	✓
		Chocolate	✓
		Cheese (processed)	✓
		Cheese (cottage)	✓
		Butter	✓
		Blueberry	✓
		Vegetable (raw)	✓
		Egg yolk	X
		Meat products/red meat	✓
		Rice (brown)	✓
		Extra virgin olive oil	✓
		Fish	✓
		Apple	✓
		Milk	✓
		Honey	✓
7	85-95%	Beer	✓
		Safflower oil	✓
		Grape	✓
		Fruits	✓
		Boiled potato	✓
		Tea (without milk)	✓
		Banana	✓

- *F1 score:* A precision metric measures the performance of model by depicting the proportion of predicted results which are correct, whereas a recall metric refers to the proportion of total correct results which are correctly identified by the model. An F1-score metric is a harmonic mean of precision and recall as described:

$$F = \frac{2 * (P * R)}{(P + R)} \tag{4.6}$$

  F1-score is 74.5% and 73.7% in the first and second phase respectively.

- *Accuracy:* Accuracies of 83% and 76% were achieved in the phases respectively. Considering the lack of a large dataset, the accuracies achieved are quite good. Moreover, this is novel research with lots of upcoming opportunities in which accuracies and other metrics can be improved by enhancing the dataset.

- *Validation by a Certified Dietician and Literature:* In the first phase, 29 associations have been predicted correctly whereas only 1 association has been identified incorrectly as per the validation also shown in the Table 4.13. In the second phase, prediction of type of association between CD and diets was performed using Naïve Bayes. Instead of using all the 78 diet terms, 21 CD and diet records were taken for this task. This is because these records were confirmed to be associated by the certified dietician as well as were already found to be associated in literature. Table 4.14 depicts the nature of associations predicted using the Proposed Prediction Approach along with the comments shared by the certified dietician for validation. Research observations like meta- analysis or other case studies reported for the associations are also depicted in the table. As per the certified dietician or research observation, nature of 16 associations have been predicted correctly.

4.4.3 Insights

In this case study, PredNEM has been used to predict associations corresponding to CD, which is also a subtype of IBD. Important inferences drawn from the retrieved

results are as follows:

- Considering the small dataset which was used in the case study, achieved accuracies are quite good. The results can be improved if validation of a larger dataset can be facilitated by employing a team of certified dieticians.
- The associations correctly predicted helpful for CD include boiled potato, whitefish, mushroom, whey, bread, safflower oil and fruits. It is quite significantly observable that most of these foods are soft and can be easily digested by colon.
- The associations correctly predicted harmful for CD include carbonated beverage, margarine, beer, corn, nuts, cheese (processed and cottage), energy drink and chocolate. It is again significantly observable that most of these foods are not easy to digest or irritates the colon.
- Such kind of in-depth analysis would be helpful in a scenario when a person is suffering from multiple diseases. The effect of comorbidities will be reflected in the complex interdependencies and thus, diets will be predicted taking such relations in consideration.

4.5 Conclusion

This Chapter presents a detailed description of the proposed prediction approach PredNEM which uses Network algorithms/parameters and machine learning algorithms for exploring unknown disease-diet associations. Two different learning methods have been demonstrated by undertaking case studies revolving around two different diseases. Some possible diet associations have been predicted for diseases.
Objective of the next Chapter is to describe the deployment of prediction approach developed for the second case study over a cloud platform. The developed service facilitates an adaptive system which can be transformed into a large scale application in future.

Table 4.14: Predicted Nature of Associations of Diet with CD

S. No.	Diet	Prediction	Validation by Dietician	Research Observations and/or Comments by Dietician
1	Boiled potato	Helpful	✓	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. A beneficial response reported for boiled potato [377] It is helpful because after boiling, starch present can be easily digested by colon.
2	Carbonated beverage	Harmful	✓	A meta analysis concluded that high intake of soft drinks might increase the risk [291] These are very harmful because they can irritate the digestive tract and cause acidity and/or trigger the disease.
3	Margarine	Harmful	✓	A prospective study of patients using dietary questionnaire suggests increased consumption of margarine might increase the risk [292] It is harmful due to presence of Poly Unsaturated Fatty Acids
4	Banana	Harmful	✗	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. A beneficial response reported for banana [377] Ripen and soft bananas are well tolerated by CD patients
5	Yogurt	Harmful	✗	A prospective study of patients using dietary questionnaire suggests increased consumption of yogurt leads to decreased risk [292] Yogurt is helpful since it is antacid in nature
6	Whitefish	Helpful	✓	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. A beneficial response reported for whitefish [377] It is helpful due to its anti-inflammatory mechanisms
7	Mushroom	Helpful	✓	A clinical trial with CD patient suggests that a mushroom extract AndoSanTM might be helpful for CD patients [378] Mushrooms have no peel and can be easily digested in cooked or boiled form

Continued on next page

Table 4.14 – *(Continued)*

S. No.	Diet	Prediction	Validation by Dietician	Research Observations and/or Comments by Dietician
8	Whey	Helpful	✓	Glutamine and whey protein display significant improvements in intestinal permeability and morphology of CD patients as studied in a randomized controlled trial [379] It contains many bioactive peptides which provide potential health benefits. Thus, it is a high protein supplement which is very helpful
9	Bread	Helpful	✓	A comparative study of patients suggests that white bread was least likely to invoke any symptoms [293] It is useful because it is soft and bland.
10	Safflower oil	Helpful	✓	A clinical trial reports that n-6 Poly Unsaturated Fatty Acids present in safflower oil decreased neutrophil based responses, which is associated with inflammation in CD [380]
11	Grape	Helpful	✗	It is not helpful for CD patients since it is hard to digest for them
12	Fruits	Helpful	✓	A prospective study of patients using dietary questionnaire suggests fruits are found to be associated with a decreased risk [292]
13	Beer	Harmful	✓	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. An unfavorable response reported for beer [377]
14	Corn/corn gluten	Harmful	✓	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. An unfavorable response reported for corns and its different forms like corn flakes, corn crackers and popcorn [377] A comparative study of patients suggests that corn is likely to invoke any symptoms [293] Corn flakes display a strong association with CD in a controlled trial [381] Corn is highly fibrous carbohydrate which is not easy to digest, thus it is harmful.
15	Nuts	Harmful	✓	Patients reported an unfavorable response for nuts consumption in a study [372]

Continued on next page

Table 4.14 – *(Continued)*

S. No.	Diet	Prediction	Validation by Dietician	Research Observations and/or Comments by Dietician
				Nuts and seeds are hard to digest. These can be eaten only after the colon is healed properly
16	Cheese (processed)	Harmful	✓	More antibodies were found against processed cheese in CD patients than in healthy controls in a case-control study, which might lead to them being harmful [382] It irritates colon and trigger the disease
17	Cheese (cottage)	Harmful	✓	A case-control study suggests that high intake of cheese increases the risk of CD [382] It irritates colon and trigger the disease
18	Energy drink	Harmful	✓	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. An unfavorable response reported for energy drinks [377] It is not helpful since it might irritate the bowel
19	Wine	Helpful	✗	CD patients from New Zealand responded to dietary questionnaire and data was statistically analyzed. An unfavorable response reported for red wine [377]
20	Honey	Harmful	✗	A prospective study of patients using dietary questionnaire suggests increased intake leads to decreased risk [292]
21	Chocolate	Harmful	✓	Chocolate has been reported as a trigger [383]

Chapter 5

Deployment of Network based Analysis over Cloud

The previous Chapter details the proposed prediction approach PredNEM designed for exploring disease-diet associations using Network analysis and machine learning algorithms. Different methods have been employed for this purpose in two case studies revolving around diseases Covid-19 and IBD. Such an approach is found to be beneficial for inferences of some important diet associations.

The current chapter describes the deployment of network based disease-diet prediction as a service. This is demonstrated by deploying the PredNEM designed for IBD and its subtypes in the cloud environment. The objective of this work is to develop an efficient, accessible and adaptable service for assisting disease-diet associations prediction.

The organization of this Chapter is as follows: Section 5.1 discusses the need of a cloud platform for this work, followed by the details of deployment discussed in Section 5.2. The retrieved results are presented in 5.3. Evaluation of the service is done in Section 5.4 along with conclusion of the Chapter in Section 5.5.

5.1 Deployment as a Service: Cloud Menu

PredNEM derived from DID-NEM is a major step forward for utilizing computational techniques for understanding disease-diet associations. However, an expeditious growth of the literature can turn out be a hindrance in this domain. This is due to the fact that the database for analysis changes with the novel advancements evident in research, which in turn affects the retrieved results as well as the performance. Thus, there is a need of a platform which assists effective analysis, keeps the database up-to-date and does not compromise with the performance in case of higher load. These conditions can only be accomplished by use of adaptive technologies like Cloud computing. Cloud technology believes in providing everything as a service including infrastructure, storage, development interface and computations. The resources are available on demand and a pay-as-you-use model is provisioned. In this manner, it supports enhancements in utilization of resources and costs cutting. Using such a platform will automate the system and the ever-increasing literature data can be stored, incorporated and manipulated smoothly for generating up-to-date results. Development of a user interface in the Cloud environment makes it easy to use, flexible, economical, efficient and easy to access. Thus, PredNEM designed for case study of IBD is deployed on the cloud for demonstrating an adaptive solution. This leads to development of a cloud-based service named Cloud Menu which assists seamless integration of Network Analysis, machine learning and Cloud technology for predicting disease-diet associations. The rationale behind this work is to design a proficient service for effectively analysing disease-diet associations and recommending diets to assist health professionals or dieticians in decision making.

5.2 Experimental Details

The architecture of the deployed service is depicted in Figure 5.1. As evident in the figure and previous Chapter, disease-disease and diet-diet associations have been extracted with the aid of an application program and pattern mining. The developed network is stored in Neo4j on a Cloud platform and TFM is employed. The extracted

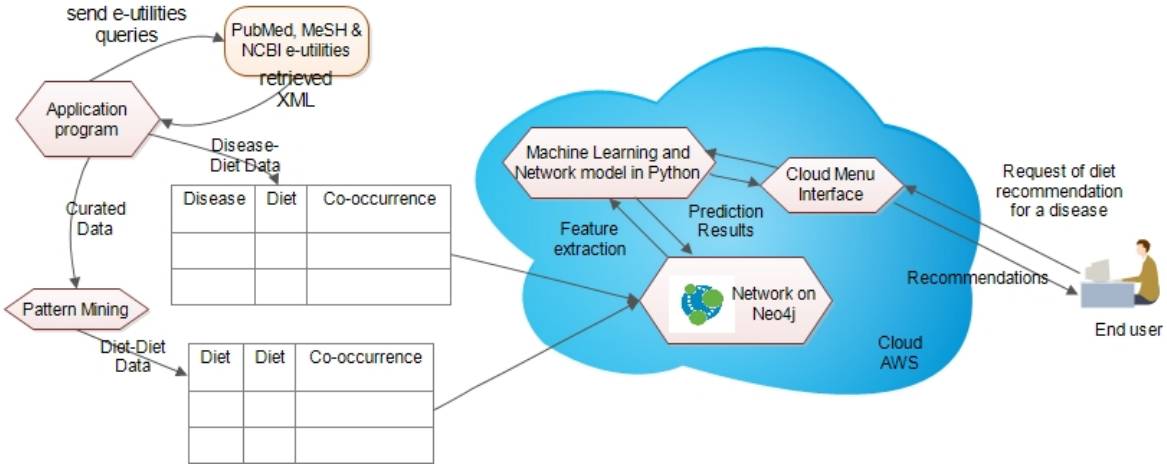


Figure 5.1: Architecture of CloudMenu

Table 5.1: Configuration Details of EC2 and Neo4j Instances

	EC2 instance	**Neo4j instance**
Type	t2.micro	m4.large
vCPUs	1	2
Memory	8 GiB	8 GiB
Network Performance	Low to Moderate	Moderate
IPv6 support	Yes	Yes
Processor	Intel Xeon 2.5 GHz	Intel Xeon 2.4 GHz

features database is evaluated over the Cloud using machine learning and the outcomes are shared on a cloud-based interface. This cloud service can be accessed by anyone through the internet. Furthermore, due to its scalability, more resources can be acquired if needed. The following strategy has been exercised to deploy the approach as a cloud service:

- *Create an Elastic Cloud Compute(EC2) instance*- An instance was required for developing and hosting the application. As a result, Ubuntu Server 20.04 LTS (HVM) (64 bit) free tier was chosen from AWS marketplace. It was secured by configuring entities including Identity and Access Management (IAM) role, elastic IP address and firewall. Configuration details of the instance are as depicted in Table 5.1.

- *Create a Neo4j instance*- The graph extracted in the previous step was also de-

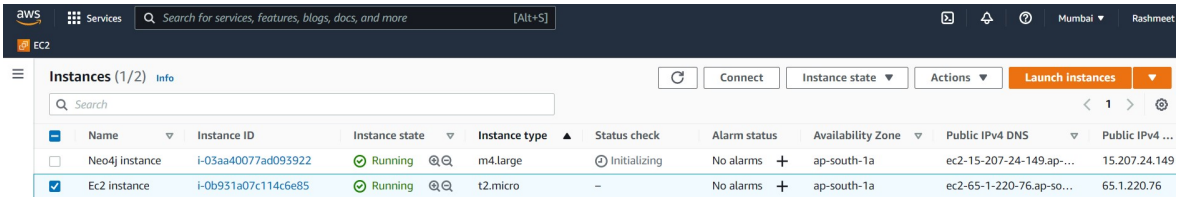


Figure 5.2: Snapshot of Neo4j and EC2 Instances created on AWS

```
ubuntu@ip-172-31-37-31:~$ cd disdiet
ubuntu@ip-172-31-37-31:~/disdiet$ python3 project4.py
* Loading Keras model and Flask starting server...please wait until server has fully started
 * Serving Flask app "project4" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Running on all addresses.
   WARNING: This is a development server. Do not use it in a production deployment.
 * Running on http://172.31.37.31:8002/ (Press CTRL+C to quit)
 * Restarting with stat
* Loading Keras model and Flask starting server...please wait until server has fully started
 * Debugger is active!
 * Debugger PIN: 424-453-239
```

Figure 5.3: Snapshot of Running Code on EC2 instance

ployed in the cloud environment so that it can be accessed remotely. The development was done on community edition of Neo4j available in AWS Cloud. Neo4j 4.3 which is compatible with Ubuntu 16 bit has been selected for this purpose. The configuration details of Neo4j instance are also depicted in Table 5.1. The ports were configured and elastic IP address was associated. Further, the configuration file was updated so that the graph can be accessed remotely. A graph corresponding to diet relations containing IBD and UC was developed using Cypher queries. The network is available at https://15.207.24.149:7473.

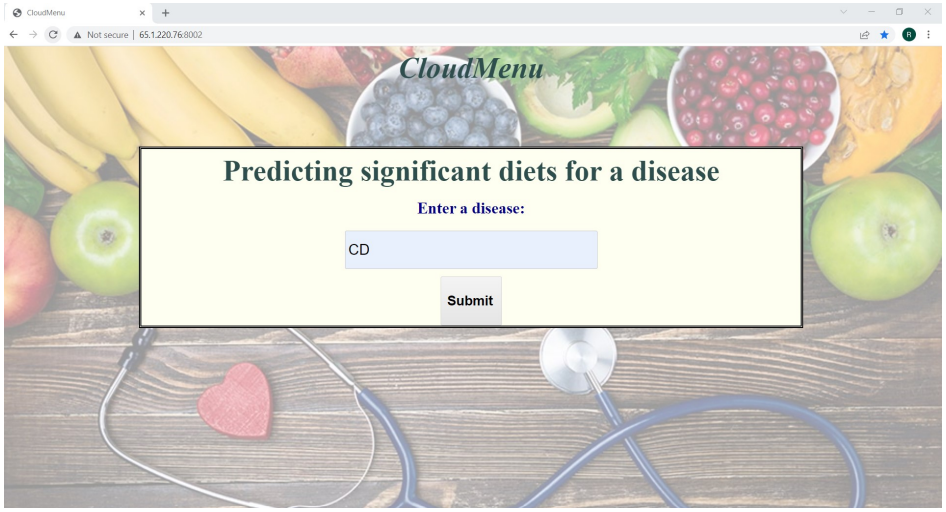
- *Deploy the Web application*- The EC2 instance was connected using EC2 instance connect service in AWS. Similarly, the Neo4j instance was connected to leverage graph database. The two instances from AWS have been shown in Figure 5.2. The code was run on EC2 instance connected remotely to the Neo4j server as shown in Figure 5.3. The results retrieved after processing were displayed on the web application. The web application is available at https://65.1.220.76:8002.

Table 5.2: Confusion Matrix of Predicted Outcomes

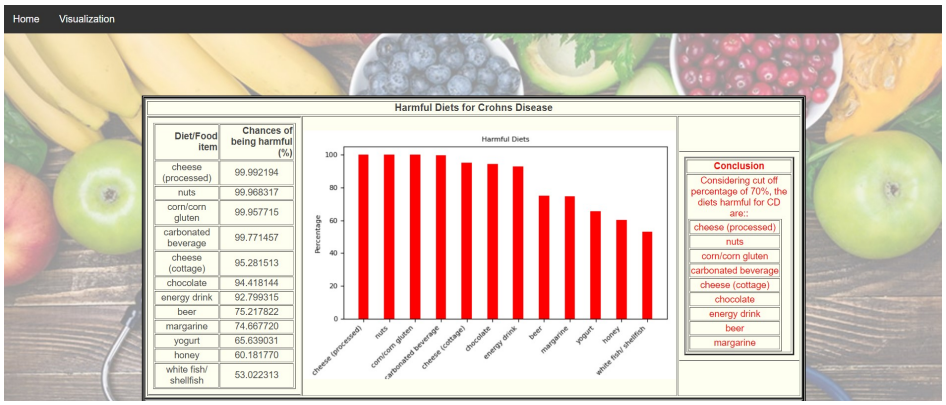
N=17	**Negative**	**Positive**
Negative	9	2
Positive	0	6

5.3 Results

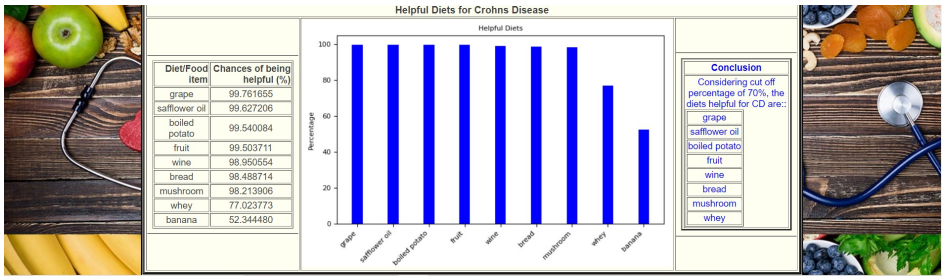
The prediction approach PredNEM fabricated for the second phase of the case study regarding IBD has been used. Thus, NB classifier was selected for learning from known associations of IBD and UC to predict unknown associations of CD. Snapshots of the Cloud-based interface developed for this service are shown in Figure 5.4 and Figure 5.5. A graphical comparison of predicted diets and their retrieved probabilities which is also available on the interface is as shown in Figure 5.6. The outcomes of the service were evaluated through dietician and by searching in medical literature. The diets predicted to be harmful for CD include cheese (processed), nuts, corn/corn gluten, carbonated beverage, cheese (cottage), chocolate, margarine, energy drink and beer, considering a probability threshold of 70%. As per a certified dietician, these diets are indeed not suitable for patients suffering from CD. The diets predicted to be helpful with the same threshold include grape, safflower oil, boiled potato, fruits, wine, bread, mushroom and whey. As per the dietician, except for grapes and wine, other diets predicted helpful are indeed beneficial for CD patients. The evaluation of proposed framework has already been done in the previous Chapter based on metrics including F1-score, accuracy and confusion matrix. An F1-score of 73.7% and accuracy of 76% were achieved using NB classifier. A confusion matrix for the obtained outcomes is shown in Table 5.2. Harmful associations are taken as negative whereas helpful associations are taken as positive. No false negatives are evident whereas only 2 false positives are seen, leading to an efficient analysis framework.



(a)



(b)



(c)

Figure 5.4: Snapshot of a) Main Page of Cloud Menu in which Disease in question is to be entered b) Page displaying the Results with Harmful Diets for entered disease which is CD here c) Page displaying Helpful Diets for CD

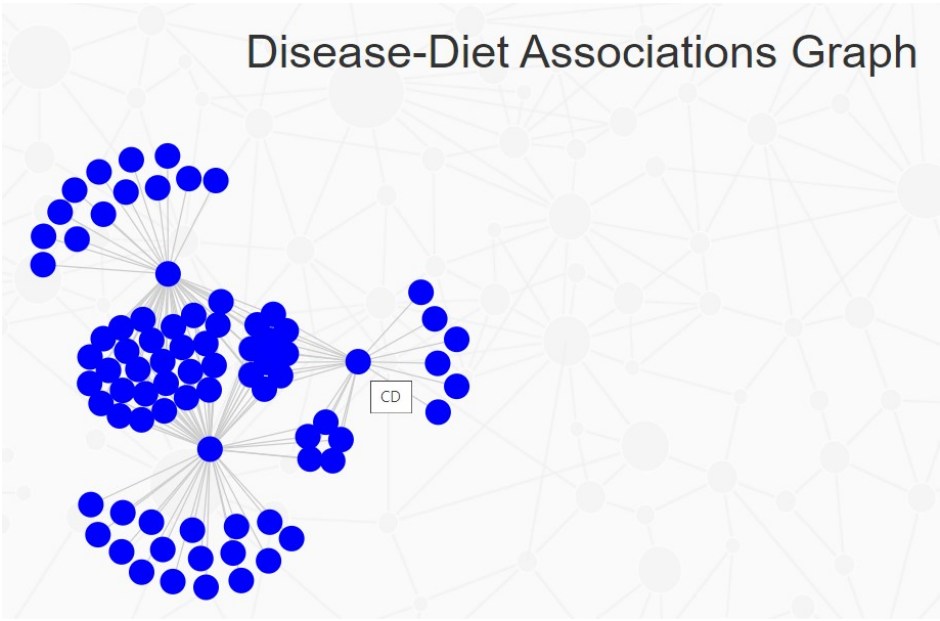
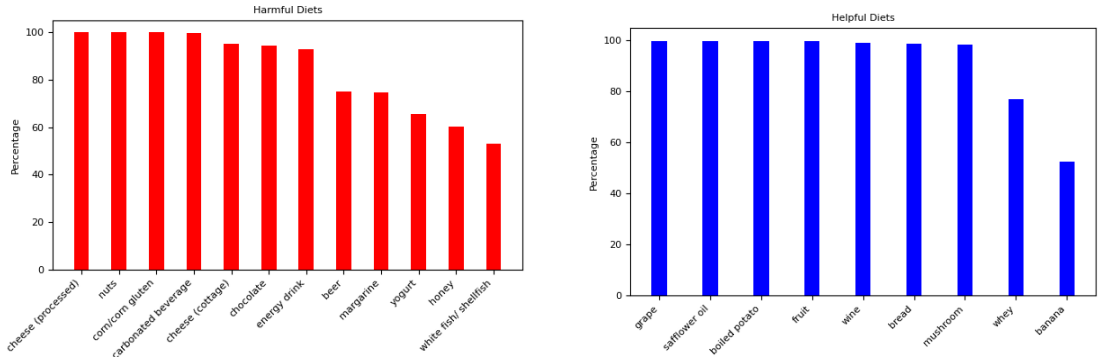


Figure 5.5: Visualization Tab displaying Resultant Graph Constructed



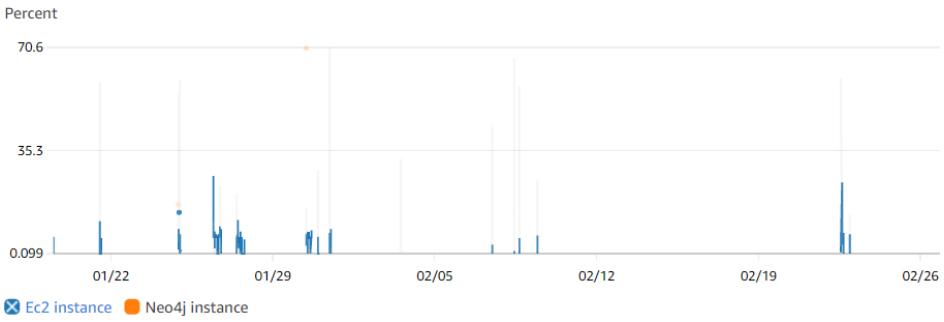
(a) (b)

Figure 5.6: Graphical Representation of Predicted Diets (Harmful and Helpful) and their Retrieved Probabilities Percentage

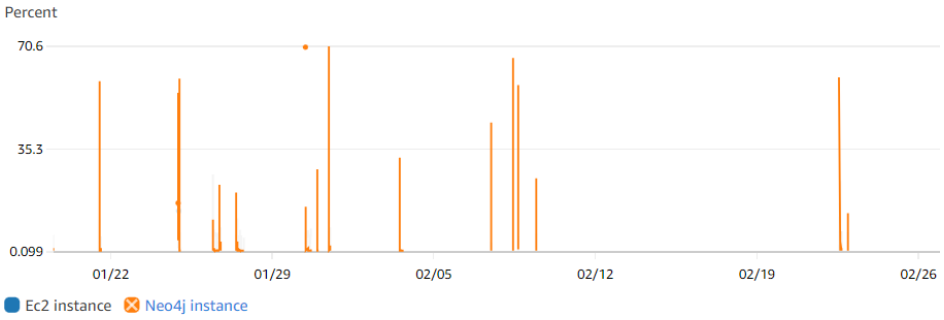
5.4 Evaluation

Evaluation of the service has been performed by monitoring two key metrics of performance namely CPU Utilization and Throughput. These are discussed as:

- *CPU utilization*: It measures the proportion of allotted compute components that are active in terms of percentage. CPU utilization for both instances (EC2 and Neo4j) has been monitored for a month as depicted in Figure 5.7. As evident in the figure, only a few compute components have been exercised for EC2 instance whereas nominal usage has been seen for Neo4j instance. Minimum, maximum and average utilizations for a period of 3 months are as described in Table 5.3. Not much percentage was utilized on an average by both the instances, being 3.09% and 1.75% respectively for EC2 instance and Neo4j instance. Even the maximum utilizations have reached a peak of 58% and 70.6% respectively. This implies an ideal CPU utilization and consequently better resource configuration.



(a)



(b)

Figure 5.7: CPU Utilization of Resources for a Month a) EC2 Instance b) Neo4j Instance

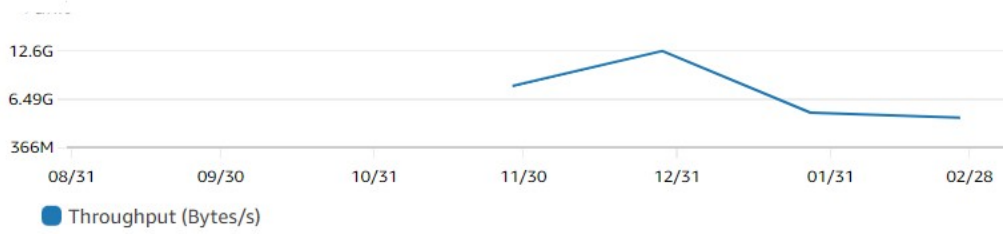
Table 5.3: CPU Utilization Metrics for 3 months

	Min CPU Utilization (100%)	**Max CPU Utilization (100%)**	**Average CPU Utilization (100%)**
EC2 Instance	0.098	58.099	3.097
Neo4j Instance	0.348	70.678	1.753

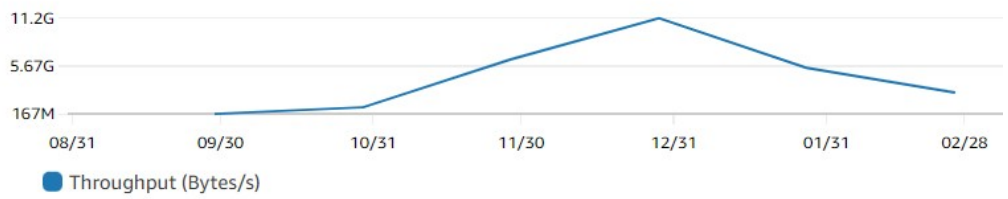
- *Throughput*: It is a measure of how many units of information a system can process in a given amount of time. Average throughput in bytes/second for an instance can be calculated using the following formula derived from metrics:

$$Throughput = \frac{(Sum(VolumeReadBytes) + Sum(VolumeWriteBytes))}{(Period - Sum(VolumeIdleTime))} \tag{5.1}$$

where Sum of $VolumeReadBytes$ and $VolumeWriteBytes$ depicts total number of bytes transferred during the specified period. It is divided by the time when the volume is active. Active time of volume is calculated by subtracting its idle time from the period. The retrieved throughput in both EC2 and Neo4j instances are depicted in Figure 5.8.



(a)



(b)

Figure 5.8: Throughput for a Month a) EC2 Instance b) Neo4j Instance

The maximum throughput for EC2 and Neo4j instances is 12.6 GB/s and 11.2 GB/s respectively, whereas the minimum throughputs are around 3.9 GB/s and 167MB/s respectively. The values retrieved demonstrate a good performance and fast service.

5.5 Conclusion

This Chapter describes the development of a Cloud-based service to assist prediction of disease-diet associations. For this purpose, PredNEM described in the previous Chapter is deployed on the Cloud. Use of Cloud technology aids in automation of the proposed approach and the ever-increasing literature data can be incorporated and manipulated smoothly for generating up-to-date results.

The next Chapter concludes this work and enlightens the readers with future prospects.

Chapter 6

Conclusions and Future Scope

This chapter culminates the thesis by discussing the overall conclusions of the research work along with future scope. It also provides a detailed discussion of thesis contributions.

This work realizes the importance of predictive healthcare and its associated techniques including Network Analysis and Cloud computing. It provides an extensive review of Network Analysis in healthcare domain, focussing on different network scenarios for exploring disease based associations. Disease-diet associations are not available as such in literature, thus a Network Model DID-NEM has been proposed to cater the needs of curation and visualization of these associations. Analysis of DID-NEM has been further performed through a proposed prediction approach PredNEM which utilizes network properties/ algorithms and machine learning. The approach is demonstrated via two case studies corresponding to Covid-19 and IBD. In order to implement this Network based analysis seamlessly, it has been deployed over the Cloud. This makes it an accessible and cost effective service. The service is available for healthcare professionals to make informed decisions.

The overall conclusions rendered from this work have been discussed in Section 6.1. Important contributions of this work have been discussed thoroughly in Section 6.2. The chapter also suggests some future research directions and possible extensions of this work as discussed in Section 6.3.

6.1 Conclusions

This work is an attempt to understand the importance of predictive analytics in healthcare specifically focusing on the disease and diet domain. Network Analysis and Cloud computing are upcoming technologies in this domain, which are used for exploring disease-diet associations. Some important conclusions derived from this work are discussed as:

- This work reviews the current status of research in understanding relation between diet and health. It is realized that there is a dire need of advanced techniques for exploring disease-diet associations. Being extensively used for understanding disease based associations, Network Analysis has been thoroughly investigated in this work. The current picture of Network Analysis in healthcare domain has been discussed, along with provision of its proposed usage and posed challenges. Existing Cloud based healthcare applications have also been summarized. These technologies seem promising for exploring disease-diet associations as indicative from the survey.

- Realizing the importance of Network Analysis, a network model DID-NEM has been proposed in this work for depicting disease-diet associations as a network and perform quantification of the associations. Data pertaining to these associations is required to be curated using literature mining through a customized technique DIDACE. After curation, follows the unfolding of steps involved in network construction of the extracted database. A comprehensive disease-diet associations database is extracted, refined, validated and represented in this work. The proposed network model is novel and can be utilized for depicting any other similar medical associations as well.

- This work proposes a prediction approach PredNEM which uses combination of Network algorithms/parameters and machine learning algorithms as learning method for exploring unknown disease-diet associations. Two different learning methods, TBM and TFM have been demonstrated by undertaking case studies

revolving around two different diseases, Covid-19 and IBD. Some possible diet associations have been predicted for each disease. The retrieved results suggest that such an approach is useful for understanding diet relations both with a novel disease and a well known disease. The approach can be used for understanding effects of comorbidities on diet and help professionals in making informed decision.

- The prediction approach PredNEM is further deployed on the Cloud. Use of Cloud technology aids in its automation. In this way, ever-increasing literature data can be incorporated and manipulated smoothly for generating up-to-date results. Such a service is accessible to researchers and medical professionals for decision making and future studies.

6.2 Thesis Contributions

This work is an attempt to improve medical or nutritional care by introducing predictive analytics to the health and diet domain. The deployed service will assist healthcare professionals in gaining research knowledge and thus making informed decisions. In this way, traditional solutions can be replaced by knowledge consisting of updated novel associations. It is also beneficial for healthcare researchers in further exploring the predicted associations. The main contributions of this thesis are as follows:

- An extensive survey of the work carried out in the domain of Network Analysis in healthcare has been performed. This is done in order to understand the dynamics of networks in disease based associations for designing an effective model.

- Developed a curation technique DIDACE for extracting disease-diet associations present in literature which automates the process and will aid in updating the database in future.

- A novel disease-diet association database is another contribution which can be further utilized by other researchers for applying their own computational methods or techniques.

- Proposed a Network model DID-NEM which includes curation, visualization and modelling of complex disease-diet associations.
- Designed a prediction approach PredNEM for two different case studies incorporating network algorithms/parameters and machine learning algorithms for advanced analysis of disease and diet relations.
- Developed a web interface CloudMenu by deploying the network based analysis over Cloud. This makes the service easy to use, flexible and economical.

6.3 Future Scope

This section presents some suggestions which can be implemented for extension of this work in future:

- The Network based analysis has been designed for disease-diet associations as a pair, but it can be customized for realizing association between disease and a set of different food items.
- In order to make curated database up-to-date with the progress made in healthcare research, curation should happen in real time. An important step to achieve this is by performing curation entirely over Cloud platform. This will not only improve the accuracy due to analysis of an inclusive dataset, but also will make the system truly global.
- Curation of database can be further sped up by incorporating a team of diet experts. In this manner, a large database can be curated and validated providing opportunities for deep learning. Thus, the work can be implemented on large scale thereby improving its scope and accuracy.
- The service has been designed to predict disease-diet associations, however the methodology used can be easily replicated for other entities which are present in the medical literature including PubMed and MeSH vocabulary for example, disease and drugs can be curated and analysed using the same procedure. Thus,

this service can be transformed for predicting any entities other than diet-disease. This might prove to be extremely beneficial for medical and computational researchers.

- In case of conditions like food allergies, diabetes, or pregnancy, individuals might need alternate diet recommendations. This work does not take this scenario in consideration. A similar approach can be devised in which associations between diets can be known based on their associated diseases. If two diet/food items are associated with many mutual diseases, they will be more similar. Based on this principle, their similarity can be predicted using similarity algorithms and a graph can be generated Alternate diets can be recommended using inferred communities from the relationships.

Bibliography

[1] M. Bhattacharyya, "Disease dietomics," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 21, no. 4, pp. 38–44, 2015.

[2] E. Horvitz, "From data to predictions and decisions: Enabling evidence-based healthcare," *Computing Community Consortium*, vol. 6, 2010.

[3] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," *International journal of medical informatics*, vol. 114, pp. 57–65, 2018.

[4] A. Hoerbst and E. Ammenwerth, "Electronic health records," *Methods of information in medicine*, vol. 49, no. 04, pp. 320–336, 2010.

[5] E. S. Berner, *Clinical decision support systems*, vol. 233. Springer, 2007.

[6] R. O. Duda and E. H. Shortliffe, "Expert systems research," *Science*, vol. 220, no. 4594, pp. 261–268, 1983.

[7] S. Soni, A. Khunteta, and M. Gupta, "A review on intelligent methods used in medicine and life science," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, pp. 703–706, 2011.

[8] G. I. Mihalas, "Evolution of trends in european medical informatics," *Acta Informatica Medica*, vol. 22, no. 1, p. 37, 2014.

[9] G. Cosma, D. Brown, M. Archer, M. Khan, and A. G. Pockley, "A survey on computational intelligence approaches for predictive modeling in prostate cancer," *Expert systems with applications*, vol. 70, pp. 1–19, 2017.

[10] E. E. Tripoliti, T. G. Papadopoulos, G. S. Karanasiou, K. K. Naka, and D. I. Fotiadis, "Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques," *Computational and structural biotechnology journal*, vol. 15, pp. 26–47, 2017.

[11] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[12] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar, "Computational health informatics in the big data age: a survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, pp. 1–36, 2016.

[13] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *Journal of Big data*, vol. 1, no. 1, pp. 1–35, 2014.

[14] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, pp. 1–10, 2014.

[15] F. F. Costa, "Big data in biomedicine," *Drug discovery today*, vol. 19, no. 4, pp. 433–440, 2014.

[16] M. I. Pramanik, R. Y. Lau, M. A. K. Azad, M. S. Hossain, M. K. H. Chowdhury, and B. Karmaker, "Healthcare informatics and analytics in big data," *Expert Systems with Applications*, vol. 152, p. 113388, 2020.

[17] H. Alharthi, "Healthcare predictive analytics: An overview with a focus on saudi arabia," *Journal of infection and public health*, vol. 11, no. 6, pp. 749–756, 2018.

[18] L. Wang and C. A. Alexander, "Big data analytics in medical engineering and healthcare: methods, advances and challenges," *Journal of medical engineering & technology*, vol. 44, no. 6, pp. 267–283, 2020.

[19] M. Bamiah, S. Brohi, S. Chuprat, *et al.*, "A study on significance of adopting cloud computing paradigm in healthcare sector," in *2012 International Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM)*, pp. 65–68, IEEE, 2012.

[20] A. Motwani, P. K. Shukla, and M. Pawar, "Novel framework based on deep learning and cloud analytics for smart patient monitoring and recommendation (spmr)," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2021.

[21] M. Al-Khafajiy, T. Baker, C. Chalmers, M. Asim, H. Kolivand, M. Fahim, and A. Waraich, "Remote health monitoring of elderly through wearable sensors," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24681–24706, 2019.

[22] G. K. Gupta, *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd., 2014.

[23] M. R. Dallagassa, C. dos Santos Garcia, E. E. Scalabrin, S. O. Ioshii, and D. R. Carvalho, "Opportunities and challenges for applying process mining in healthcare: a systematic mapping study," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2021.

[24] S. Hassanpour and C. P. Langlotz, "Unsupervised topic modeling in a large free text radiology report repository," *Journal of digital imaging*, vol. 29, no. 1, pp. 59–62, 2016.

[25] S. Hassanpour and C. P. Langlotz, "Information extraction from multi-institutional radiology reports," *Artificial intelligence in medicine*, vol. 66, pp. 29–39, 2016.

[26] H. Hao, K. Zhang, W. Wang, and G. Gao, "A tale of two countries: International comparison of online doctor reviews between china and the united states," *International journal of medical informatics*, vol. 99, pp. 37–44, 2017.

[27] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in india," *International Journal of Diabetes in Developing Countries*, vol. 36, no. 4, pp. 519–528, 2016.

[28] A.-L. Barabási, "Network medicine—from obesity to the "diseasome"," 2007.

[29] R. Toor and I. Chana, "Network analysis as a computational technique and its benefaction for predictive analysis of healthcare data: A systematic review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 1689–1711, 2021.

[30] "Personalized Nutrition Project." http://newsite.personalnutrition.org/WebSite/Home.aspx. Accessed: 2022-03-07.

[31] "Hundred Person Wellness Project, Institute for Systems Biology." https://isbscience.org/research/100k-wellness-project/. Accessed: 2022-03-07.

[32] "Food4me, Research-backed Nutrition." https://www.food4me.org/. Accessed: 2022-03-07.

[33] R. San-Cristobal, S. Navas-Carretero, C. Celis-Morales, K. M. Livingstone, B. Stewart-Knox, A. Rankin, A. L. Macready, R. Fallaize, C. B. O'Donovan, H. Forster, *et al.*, "Capturing health and eating status through a nutritional perception screening questionnaire (npsq9) in a randomised internet-based personalised nutrition intervention: the food4me study," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 14, no. 1, pp. 1–12, 2017.

[34] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2016.

[35] S. Wasserman, K. Faust, *et al.*, "Social network analysis: Methods and applications," 1994.

[36] D. Chambers, P. Wilson, C. Thompson, and M. Harden, "Social network analysis in healthcare settings: a systematic scoping review," 2012.

[37] "Network Analysis Lecture Series by Zhukov Leonid." https://www.youtube.com/watch?v=UHnmPu8Zevg&index=1&list=FLlJVsj8bum0N30-hsAR7ddg. Accessed: 2022-04-04.

[38] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[39] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[40] J. Powell, A. Clarke, *et al.*, "The www of the world wide web: who, what, and why?," *Journal of Medical Internet Research*, vol. 4, no. 1, p. e854, 2002.

[41] S. Golder and J. McCambridge, "Alcohol, cardiovascular disease and industry funding: A co-authorship network analysis of systematic reviews," *Social Science & Medicine*, vol. 289, p. 114450, 2021.

[42] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[43] A.-L. Barabási, "Network science," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1987, p. 20120375, 2013.

[44] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022.

[45] "Network Analysis Lecture Series." https://www.youtube.com/watch?v=1T5-BG6yngM&list=PLriUvS7IljvkGesFRuYjqRz4lKgodJgh2. Accessed: 2022-03-07.

[46] C. Pastrello, E. Pasini, M. Kotlyar, D. Otasek, S. Wong, W. Sangrar, S. Rahmati, and I. Jurisica, "Integration, visualization and analysis of human interactome,"

Biochemical and biophysical research communications, vol. 445, no. 4, pp. 757–773, 2014.

[47] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[48] M. Flores, G. Glusman, K. Brogaard, N. D. Price, and L. Hood, "P4 medicine: how systems medicine will transform the healthcare sector and society," *Personalized medicine*, vol. 10, no. 6, pp. 565–576, 2013.

[49] R. Toor and I. Chana, "Application of it in healthcare: a systematic review," *ACM SIGBioinformatics Record*, vol. 6, no. 2, pp. 1–8, 2016.

[50] "Cloud Computing by AWS." https://aws.amazon.com/what-is-cloud-computing/. Accessed: 2022-04-06.

[51] "Cloud Computing by IBM." https://www.ibm.com/in-en/cloud/learn/cloud-computing. Accessed: 2022-04-06.

[52] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[53] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[54] R. L. Grossman, "The case for cloud computing," *IT professional*, vol. 11, no. 2, pp. 23–27, 2009.

[55] L. Qian, Z. Luo, Y. Du, and L. Guo, "Cloud computing: An overview," in *IEEE international conference on cloud computing*, pp. 626–631, Springer, 2009.

[56] B. Furht, A. Escalante, *et al.*, *Handbook of cloud computing*, vol. 3. Springer, 2010.

[57] V. Navale and P. E. Bourne, "Cloud computing applications for biomedical science: A perspective," *PLoS computational biology*, vol. 14, no. 6, p. e1006144, 2018.

[58] Q. Yao, X. Han, X.-K. Ma, Y.-F. Xue, Y.-J. Chen, and J.-S. Li, "Cloud-based hospital information system as a service for grassroots healthcare institutions," *Journal of medical systems*, vol. 38, no. 9, pp. 1–7, 2014.

[59] F. Abidi, H. J. Abidi, and S. Armani, "Cloud libraries: A novel application of cloud computing," in *International Conference on Education and e-Learning Innovations*, pp. 1–4, IEEE, 2012.

[60] "Hathi Trust Digital Library." https://www.hathitrust.org/. Accessed: 2022-04-08.

[61] "NutriChem Server 2.0." http://sbb.hku.hk/services/NutriChem-2.0/. Accessed: 2022-03-07.

[62] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Nutrichem: a systems chemical biology resource to explore the medicinal value of plant-based foods," *Nucleic acids research*, vol. 43, no. D1, pp. D940–D945, 2015.

[63] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level," *PLoS computational biology*, vol. 10, no. 1, p. e1003432, 2014.

[64] H.-I. Lin, H.-M. Chen, C.-C. Hsu, H.-J. Lin, J.-J. Wang, S.-F. Weng, Y. Kao, and C.-C. Huang, "Associations between dietary patterns and stages of chronic kidney disease," *BMC nephrology*, vol. 23, no. 1, pp. 1–13, 2022.

[65] N. Z. Siddiqui, A. N. Nguyen, S. Santos, and T. Voortman, "Diet quality and cardiometabolic health in childhood: the generation r study," *European journal of nutrition*, vol. 61, no. 2, pp. 729–736, 2022.

[66] A. Sezaki, T. Imai, K. Miyamoto, F. Kawase, Y. Shirai, C. Abe, M. Sanada, A. Inden, T. Rato, N. Sugihara, *et al.*, "Association between the mediterranean diet score and healthy life expectancy: A global comparative study," *The journal of nutrition, health & aging*, pp. 1–7, 2022.

[67] S.-J. Liu, P.-D. Huang, J.-M. Xu, Q. Li, J.-H. Xie, W.-Z. Wu, C.-T. Wang, and X.-B. Yang, "Diet and gastric cancer risk: an umbrella review of systematic reviews and meta-analyses of prospective cohort studies," *Journal of Cancer Research and Clinical Oncology*, pp. 1–14, 2022.

[68] A. Roustazadeh, H. Mir, S. Jafarirad, F. Mogharab, S. A. Hosseini, A. Abdoli, and S. Erfanian, "A dietary pattern rich in fruits and dairy products is inversely associated to gestational diabetes: a case-control study in iran," *BMC Endocrine Disorders*, vol. 21, no. 1, pp. 1–9, 2021.

[69] M. Salomé, L. Arrazat, J. Wang, A. Dufour, C. Dubuisson, J.-L. Volatier, J.-F. Huneau, and F. Mariotti, "Contrary to ultra-processed foods, the consumption of unprocessed or minimally processed foods is associated with favorable patterns of protein intake, diet quality and lower cardiometabolic risk in french adults (inca3)," *European Journal of Nutrition*, vol. 60, no. 7, pp. 4055–4067, 2021.

[70] Z. Heidari, E. Mohammadi, V. Aghamohammadi, S. Jalali, A. Rezazadeh, F. Sedaghat, M. Assadi, and B. Rashidkhani, "Dietary approaches to stop hypertension (dash) diets and breast cancer among women: a case control study," *BMC cancer*, vol. 20, no. 1, pp. 1–10, 2020.

[71] N. P. Bondonno, L. C. Blekkenhorst, A. L. Bird, J. R. Lewis, J. M. Hodgson, N. Shivappa, J. R. Hébert, R. J. Woodman, G. Wong, D. A. Kerr, *et al.*, "Dietary inflammatory index and the aging kidney in older women: a 10-year prospective cohort study," *European Journal of Nutrition*, vol. 59, no. 7, pp. 3201–3211, 2020.

[72] A. Arab, S. Mehrabani, S. Moradi, and R. Amani, "The association between diet and mood: A systematic review of current literature," *Psychiatry research*, vol. 271, pp. 428–437, 2019.

[73] R. Mendonça, N. Carvalho, J. Martin-Moreno, A. Pimenta, A. Lopes, A. Gea, M. Martinez-Gonzalez, and M. Bes-Rastrollo, "Total polyphenol intake, polyphenol subtypes and incidence of cardiovascular disease: The sun cohort study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 29, no. 1, pp. 69–78, 2019.

[74] A. Salari-Moghaddam, A. H. Keshteli, A. Esmaillzadeh, and P. Adibi, "Adherence to the pro-inflammatory diet in relation to prevalence of irritable bowel syndrome," *Nutrition journal*, vol. 18, no. 1, pp. 1–10, 2019.

[75] N. Shivappa, C. Niclis, J. B. Coquet, M. D. Román, J. R. Hébert, and M. d. P. Diaz, "Increased inflammatory potential of diet is associated with increased odds of prostate cancer in argentinian men," *Cancer Causes & Control*, vol. 29, no. 9, pp. 803–813, 2018.

[76] A. H. Liu, C. P. Bondonno, J. Russell, V. M. Flood, J. R. Lewis, K. D. Croft, R. J. Woodman, W. H. Lim, A. Kifley, G. Wong, *et al.*, "Relationship of dietary nitrate intake from vegetables with cardiovascular disease mortality: a prospective study in a cohort of older australians," *European journal of nutrition*, vol. 58, no. 7, pp. 2741–2753, 2019.

[77] N. Namazi, B. Larijani, and L. Azadbakht, "Association between the dietary inflammatory index and the incidence of cancer: A systematic review and meta-analysis of prospective studies," *Public Health*, vol. 164, pp. 148–156, 2018.

[78] C. Rodríguez-Martin, R. Alonso-Domínguez, M. C. Patino-Alonso, M. A. Gómez-Marcos, J. A. Maderuelo-Fernández, C. Martin-Cantera, L. García-Ortiz, and J. I. Recio-Rodríguez, "The evident diet quality index is associated with cardiovascular risk and arterial stiffness in adults," *BMC Public Health*, vol. 17, no. 1, pp. 1–9, 2017.

[79] E. Mertens, O. Markey, J. M. Geleijnse, J. A. Lovegrove, and D. I. Givens, "Adherence to a healthy diet in relation to cardiovascular incidence and risk

markers: evidence from the caerphilly prospective study," *European journal of nutrition*, vol. 57, no. 3, pp. 1245–1258, 2018.

[80] N. P. Rocha, L. C. Milagres, G. Z. Longo, A. Q. Ribeiro, and J. F. d. Novaes, "Association between dietary pattern and cardiometabolic risk in children and adolescents: a systematic review," *Jornal de pediatria*, vol. 93, pp. 214–222, 2017.

[81] M. A. Martínez-González, E. Fernández-Jarne, M. Serrano-Martínez, A. Marti, J. A. Martinez, and J. M. Martín-Moreno, "Mediterranean diet and reduction in the risk of a first acute myocardial infarction: an operational healthy dietary score," *European journal of nutrition*, vol. 41, no. 4, pp. 153–160, 2002.

[82] F. S. Bernaud, M. V. Beretta, C. do Nascimento, F. Escobar, J. L. Gross, M. J. Azevedo, and T. C. Rodrigues, "Fiber intake and inflammation in type 1 diabetes," *Diabetology & metabolic syndrome*, vol. 6, no. 1, pp. 1–10, 2014.

[83] A. Otaegui-Arrazola, P. Amiano, A. Elbusto, E. Urdaneta, and P. Martínez-Lage, "Diet, cognition, and alzheimer's disease: food for thought," *European journal of nutrition*, vol. 53, no. 1, pp. 1–23, 2014.

[84] S. Rohrmann, K. Overvad, H. B. Bueno-de Mesquita, M. U. Jakobsen, R. Egeberg, A. Tjønneland, L. Nailler, M.-C. Boutron-Ruault, F. Clavel-Chapelon, V. Krogh, *et al.*, "Meat consumption and mortality-results from the european prospective investigation into cancer and nutrition," *BMC medicine*, vol. 11, no. 1, pp. 1–12, 2013.

[85] S. Z. Sun and M. W. Empie, "Lack of findings for the association between obesity risk and usual sugar-sweetened beverage consumption in adults–a primary analysis of databases of csfii-1989–1991, csfii-1994–1998, nhanes iii, and combined nhanes 1999–2002," *Food and chemical toxicology*, vol. 45, no. 8, pp. 1523–1536, 2007.

[86] H. Jiang and Y. Huang, "An effective drug-disease associations prediction model

based on graphic representation learning over multi-biomolecular network," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–17, 2022.

[87] B.-Y. Ji, Z.-H. You, Z.-H. Chen, L. Wong, and H.-C. Yi, "Nempd: a network embedding-based method for predicting mirna-disease associations by preserving behavior and attribute information," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–17, 2020.

[88] D. Yao, X. Zhan, and C.-K. Kwoh, "An improved random forest-based computational model for predicting novel mirna-disease associations," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–14, 2019.

[89] B. Bhasuran and J. Natarajan, "Automatic extraction of gene-disease associations from literature using joint ensemble learning," *PloS one*, vol. 13, no. 7, p. e0200699, 2018.

[90] A. Gutiérrez-Sacristán, À. Bravo, M. Portero-Tresserra, O. Valverde, A. Armario, M. Blanco-Gandía, A. Farré, L. Fernández-Ibarrondo, F. Fonseca, J. Giraldo, *et al.*, "Text mining and expert curation to develop a database on psychiatric diseases and their genes," *Database*, vol. 2017, 2017.

[91] M. Khordad and R. E. Mercer, "Identifying genotype-phenotype relationships in biomedical text," *Journal of biomedical semantics*, vol. 8, no. 1, pp. 1–16, 2017.

[92] B. Haslam and L. Perez-Breva, "Learning disease relationships from clinical drug trials," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 13–23, 2017.

[93] W. Ma, L. Zhang, P. Zeng, C. Huang, J. Li, B. Geng, J. Yang, W. Kong, X. Zhou, and Q. Cui, "An analysis of human microbe–disease associations," *Briefings in bioinformatics*, vol. 18, no. 1, pp. 85–97, 2017.

[94] L. Wang, G. Del Fiol, B. E. Bray, and P. J. Haug, "Generating disease-pertinent treatment vocabularies from medline citations," *Journal of biomedical informatics*, vol. 65, pp. 46–57, 2017.

[95] J. Kim, J.-j. Kim, and H. Lee, "An analysis of disease-gene relationship from medline abstracts by digsee," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[96] D. Jang, S. Lee, J. Lee, K. Kim, and D. Lee, "Inferring new drug indications using the complementarity between clinical disease signatures and drug effects," *Journal of biomedical informatics*, vol. 59, pp. 248–257, 2016.

[97] A. Singhal, M. Simmons, and Z. Lu, "Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature," *Journal of the American Medical Informatics Association*, vol. 23, no. 4, pp. 766–772, 2016.

[98] D. M. Lowe, N. M. O'Boyle, and R. A. Sayle, "Efficient chemical-disease identification and relationship extraction using wikipedia to improve recall," *Database*, vol. 2016, 2016.

[99] C.-C. Huang and Z. Lu, "Discovering biomedical semantic relations in pubmed queries for information retrieval and database curation," *Database*, vol. 2016, 2016.

[100] R. Jiang, "Walking on multiple disease-gene networks to prioritize candidate genes," *Journal of molecular cell biology*, vol. 7, no. 3, pp. 214–230, 2015.

[101] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Analysis of the human diseasome using phenotype similarity between common, genetic and infectious diseases," *Scientific reports*, vol. 5, no. 1, pp. 1–14, 2015.

[102] K. Sun, J. P. Gonçalves, C. Larminie, and N. Pržulj, "Predicting disease associations via biological network analysis," *BMC bioinformatics*, vol. 15, no. 1, pp. 1–13, 2014.

[103] Y. Li and P. Agarwal, "A pathway-based view of human diseases and disease relationships," *PloS one*, vol. 4, no. 2, p. e4346, 2009.

[104] T. N. Trang Tran, M. Atas, A. Felfernig, and M. Stettinger, "An overview of recommender systems in the healthy food domain," *Journal of Intelligent Information Systems*, vol. 50, no. 3, pp. 501–526, 2018.

[105] J. Marshall, P. Jimenez-Pazmino, R. Metoyer, and N. V. Chawla, "A survey on healthy food decision influences through technological innovations," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 2, pp. 1–27, 2022.

[106] R. Y. Toledo, A. A. Alzahrani, and L. Martinez, "A food recommender system considering nutritional information and user preferences," *IEEE Access*, vol. 7, pp. 96695–96711, 2019.

[107] G. Agapito, M. Simeoni, B. Calabrese, I. Caré, T. Lamprinoudi, P. H. Guzzi, A. Pujia, G. Fuiano, and M. Cannataro, "Dietos: A dietary recommender system for chronic diseases monitoring and management," *Computer methods and programs in biomedicine*, vol. 153, pp. 93–104, 2018.

[108] C. Iwendi, S. Khan, J. H. Anajemba, A. K. Bashir, and F. Noor, "Realizing an efficient iomt-assisted patient diet recommendation system through machine learning model," *IEEE Access*, vol. 8, pp. 28462–28474, 2020.

[109] T. Ivaşcu, A. Diniş, and K. Cincar, "A disease-driven nutrition recommender system based on a multi-agent architecture," in *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pp. 1–5, 2018.

[110] T. Ueta, M. Iwakami, and T. Ito, "A recipe recommendation system based on automatic nutrition information extraction," in *International Conference on Knowledge Science, Engineering and Management*, pp. 79–90, Springer, 2011.

[111] M. Bhattacharyya, S. Maity, and S. Bandyopadhyay, "Exploring the missing links between dietary habits and diseases," *IEEE Transactions on NanoBioscience*, vol. 16, no. 3, pp. 226–238, 2017.

[112] C. Aggarwal and K. Subbian, "Evolutionary network analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1–36, 2014.

[113] A. Starke, C. Trattner, H. Bakken, M. Johannessen, and V. Solberg, "The cholesterol factor: Balancing accuracy and health in recipe recommendation through a nutrient-specific metric," in *CEUR Workshop Proceedings*, vol. 2959, 2021.

[114] M. Shrimal, M. Khavnekar, S. Thorat, and J. Deone, "Nutriflow: A diet recommendation system," *Available at SSRN 3866863*, 2021.

[115] D. Mckensy-Sambola, M. Á. Rodríguez-García, F. García-Sánchez, and R. Valencia-García, "Ontology-based nutritional recommender system," *Applied Sciences*, vol. 12, no. 1, p. 143, 2021.

[116] C. Musto, C. Trattner, A. Starke, and G. Semeraro, "Towards a knowledge-aware food recommender system exploiting holistic user models," in *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pp. 333–337, 2020.

[117] N. M. Yusof and S. A. M. Noah, "Semantically enhanced case adaptation for dietary menu recommendation of diabetic patients," in *Joint International Semantic Technology Conference*, pp. 318–333, Springer, 2017.

[118] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2015.

[119] M. A. Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social network data analytics*, pp. 243–275, Springer, 2011.

[120] S. E. Schaeffer, "Graph clustering," *Computer science review*, vol. 1, no. 1, pp. 27–64, 2007.

[121] Y. Sun and J. Han, "Ranking methods for networks.," 2014.

[122] V. Krishna, N. R. Suri, and G. Athithan, "A comparative survey of algorithms for frequent subgraph discovery," *Current Science*, pp. 190–198, 2011.

[123] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature communications*, vol. 5, no. 1, pp. 1–10, 2014.

[124] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization," *Computer networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

[125] B. Guo, D. Zhang, Z. Wang, Z. Yu, and X. Zhou, "Opportunistic iot: Exploring the harmonious interaction between human and the internet of things," *Journal of Network and Computer Applications*, vol. 36, no. 6, pp. 1531–1539, 2013.

[126] M. Soliman, O. Nasraoui, and N. G. Cooper, "Building a glaucoma interaction network using a text mining approach," *BioData mining*, vol. 9, no. 1, pp. 1–25, 2016.

[127] R. Vyas, S. Bapat, E. Jain, M. Karthikeyan, S. Tambe, and B. D. Kulkarni, "Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis," *Computational biology and chemistry*, vol. 65, pp. 37–44, 2016.

[128] H. Carter, M. Hofree, and T. Ideker, "Genotype to phenotype via network analysis," *Current opinion in genetics & development*, vol. 23, no. 6, pp. 611–621, 2013.

[129] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[130] P. Sah, L. O. Singh, A. Clauset, and S. Bansal, "Exploring community structure in biological networks with random graphs," *BMC bioinformatics*, vol. 15, no. 1, pp. 1–14, 2014.

[131] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, 2015.

[132] A. Stevens, C. De Leonibus, D. Hanson, A. Dowsey, A. Whatmore, S. Meyer, R. Donn, P. Chatelain, I. Banerjee, K. Cosgrove, *et al.*, "Network analysis: a new approach to study endocrine disorders," *Journal of molecular endocrinology*, vol. 52, no. 1, pp. R79–R93, 2014.

[133] W. Liu, A. Wu, M. Pellegrini, and X. Wang, "Integrative analysis of human protein, function and disease networks," *Scientific reports*, vol. 5, no. 1, pp. 1–11, 2015.

[134] G. Alanis-Lobato, "Mining protein interactomes to improve their reliability and support the advancement of network medicine," *Frontiers in genetics*, vol. 6, p. 296, 2015.

[135] N. Pržulj, D. A. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.

[136] S. Hasan, B. K. Bonde, N. S. Buchan, and M. D. Hall, "Network analysis has diverse roles in drug discovery," *Drug discovery today*, vol. 17, no. 15-16, pp. 869–874, 2012.

[137] Z. M. Ibrahim and A. Ngom, "The relative vertex clustering value-a new criterion for the fast discovery of functional modules in protein interaction networks," *BMC bioinformatics*, vol. 16, no. 4, pp. 1–14, 2015.

[138] L. Yu, X. Ma, L. Zhang, J. Zhang, and L. Gao, "Prediction of new drug indications based on clinical data and network modularity," *Scientific reports*, vol. 6, no. 1, pp. 1–12, 2016.

[139] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via

network-based inference," *PLoS computational biology*, vol. 8, no. 5, p. e1002503, 2012.

[140] Z. Stanfield, M. Coşkun, and M. Koyutürk, "Drug response prediction as a link prediction problem," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[141] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397–2403, 2009.

[142] Z. Wu, F. Cheng, J. Li, W. Li, G. Liu, and Y. Tang, "Sdtnbi: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning," *Briefings in bioinformatics*, vol. 18, no. 2, pp. 333–347, 2017.

[143] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2016.

[144] A. C. Nascimento, R. B. Prudêncio, and I. G. Costa, "A multiple kernel learning algorithm for drug-target interaction prediction," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–16, 2016.

[145] G. Fu, Y. Ding, A. Seal, B. Chen, Y. Sun, and E. Bolton, "Predicting drug target interactions using meta-path-based semantic network analysis," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–10, 2016.

[146] M. Ernst, Y. Du, G. Warsow, M. Hamed, N. Endlich, K. Endlich, H. Murua Escobar, L.-M. Sklarz, S. Sender, C. Junghanß, *et al.*, "Focusheuristics–expression-data-driven network optimization and disease gene prediction," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[147] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an

integrated human functional linkage network," *Genome biology*, vol. 10, no. 9, pp. 1–17, 2009.

[148] Y. Chen and R. Xu, "Context-sensitive network-based disease genetics prediction and its implications in drug discovery," *Bioinformatics*, vol. 33, no. 7, pp. 1031–1039, 2017.

[149] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. De Moor, and Y. Moreau, "Candidate gene prioritization by network analysis of differential expression using machine learning approaches," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–16, 2010.

[150] Z. Razaghi-Moghadam, R. Abdollahi, S. Goliaei, and M. Ebrahimi, "Hybridranker: integrating network topology and biomedical knowledge to prioritize cancer candidate genes," *Journal of biomedical informatics*, vol. 64, pp. 139–146, 2016.

[151] Y. Guan, C. L. Ackert-Bicknell, B. Kell, O. G. Troyanskaya, and M. A. Hibbs, "Functional genomics complements quantitative genetics in identifying disease-gene associations," *PLoS computational biology*, vol. 6, no. 11, p. e1000991, 2010.

[152] L. Eronen and H. Toivonen, "Biomine: predicting links between biological entities using network models of heterogeneous databases," *BMC bioinformatics*, vol. 13, no. 1, pp. 1–21, 2012.

[153] M. D. Leiserson, J. V. Eldridge, S. Ramachandran, and B. J. Raphael, "Network analysis of gwas data," *Current opinion in genetics & development*, vol. 23, no. 6, pp. 602–610, 2013.

[154] F. Ferrazzi, R. Bellazzi, and F. B. Engel, "Gene network analysis: from heart development to cardiac therapy," *Thrombosis and Haemostasis*, vol. 113, no. 03, pp. 521–531, 2015.

[155] J. H. Kim, K. Y. Son, D. W. Shin, S. H. Kim, J. W. Yun, J. H. Shin, M. S. Kang, E. H. Chung, K. H. Yoo, and J. M. Yun, "Network analysis of human diseases

using korean nationwide claims data," *Journal of Biomedical Informatics*, vol. 61, pp. 276–282, 2016.

[156] A. Khan, S. Uddin, and U. Srinivasan, "Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression," *International journal of medical informatics*, vol. 115, pp. 1–9, 2018.

[157] E. Belyi, P. J. Giabbanelli, I. Patel, N. H. Balabhadrapathruni, A. B. Abdallah, W. Hameed, and V. K. Mago, "Combining association rule mining and network analysis for pharmacosurveillance," *The Journal of supercomputing*, vol. 72, no. 5, pp. 2014–2034, 2016.

[158] C. R. Evans, J.-P. Onnela, D. R. Williams, and S. Subramanian, "Multiple contexts and adolescent body mass index: Schools, neighborhoods, and social networks," *Social science & medicine*, vol. 162, pp. 21–31, 2016.

[159] R. B. Sampaio, M. V. d. A. Fonseca, F. Zicker, *et al.*, "Co-authorship network analysis in health research: method and potential use," *Health research policy and systems*, vol. 14, no. 1, pp. 1–10, 2016.

[160] Y. Liu, Y. Cheng, Z. Yan, and X. Ye, "Multilevel analysis of international scientific collaboration network in the influenza virus vaccine field: 2006–2013," *Sustainability*, vol. 10, no. 4, p. 1232, 2018.

[161] T. McCurdie, P. Sanderson, and L. M. Aitken, "Applying social network analysis to the examination of interruptions in healthcare," *Applied ergonomics*, vol. 67, pp. 50–60, 2018.

[162] N. Moradianzadeh, P. M. Zadeh, Z. Kobti, S. Hansen, and K. Pfaff, "Using social network analysis to model palliative care," *Journal of Network and Computer Applications*, vol. 120, pp. 30–41, 2018.

[163] B. D. Steitz and M. A. Levy, "Temporal and atemporal provider network analysis in a breast cancer cohort from an academic medical center (usa)," in *Informatics*, vol. 5, p. 34, Multidisciplinary Digital Publishing Institute, 2018.

[164] M. De Vries, P. Kenis, M. Kraaij-Dirkzwager, E. J. Ruitenberg, J. Raab, and A. Timen, "Collaborative emergency preparedness and response to cross-institutional outbreaks of multidrug-resistant organisms: a scenario-based approach in two regions of the netherlands," *BMC public health*, vol. 19, no. 1, pp. 1–12, 2019.

[165] M.-h. Kim, S. Banerjee, Y. Zhao, F. Wang, Y. Zhang, Y. Zhu, J. DeFerio, L. Evans, S. M. Park, and J. Pathak, "Association networks in a matched case-control design–co-occurrence patterns of preexisting chronic medical conditions in patients with major depression versus their matched controls," *Journal of biomedical informatics*, vol. 87, pp. 88–95, 2018.

[166] N. Mammone, S. De Salvo, L. Bonanno, C. Ieracitano, S. Marino, A. Marra, A. Bramanti, and F. C. Morabito, "Brain network analysis of compressive sensed high-density eeg signals in ad and mci subjects," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 527–536, 2018.

[167] J. E. Choi and M. S. Kim, "Exploring the knowledge structure of nursing care for older patients with delirium: keyword network analysis," *CIN: Computers, Informatics, Nursing*, vol. 36, no. 5, pp. 216–224, 2018.

[168] A. L. Kjos and G. A. Bryant, "Communication networks of medication management in an ambulatory care setting," *Research in Social and Administrative Pharmacy*, vol. 15, no. 2, pp. 182–192, 2019.

[169] H. Siden and K. Urbanoski, "Using network analysis to map the formal clinical reporting process in pediatric palliative care: a pilot study," *BMC health services research*, vol. 11, no. 1, pp. 1–11, 2011.

[170] N. Yao, X. Zhu, A. Dow, V. K. Mishra, A. Phillips, and S.-P. Tu, "An exploratory study of networks constructed using access data from an electronic health record," *Journal of interprofessional care*, vol. 32, no. 6, pp. 666–673, 2018.

[171] V. Moscato and G. Sperlì, "Community detection over feature-rich information networks: An ehealth case study," *Information Systems*, p. 102092, 2022.

[172] T. Pham, X. Tao, J. Zhang, and J. Yong, "Constructing a knowledge-based heterogeneous information graph for medical health status classification," *Health information science and systems*, vol. 8, no. 1, pp. 1–14, 2020.

[173] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–12, 2017.

[174] Y. Lu, Y. Guo, and A. Korhonen, "Link prediction in drug-target interactions network using similarity indices," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–9, 2017.

[175] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, A. G. Dunn, *et al.*, "Characterizing twitter discussions about hpv vaccines using topic modeling and community detection," *Journal of medical Internet research*, vol. 18, no. 8, p. e6045, 2016.

[176] B. Kaya and M. Poyraz, "Age-series based link prediction in evolving disease networks," *Computers in biology and medicine*, vol. 63, pp. 1–10, 2015.

[177] T. Narayanan and S. Subramaniam, "A newtonian framework for community detection in undirected biological networks," *IEEE transactions on biomedical circuits and systems*, vol. 8, no. 1, pp. 65–73, 2014.

[178] B. Kaya and M. Poyraz, "Supervised link prediction in symptom networks with evolving case," *Measurement*, vol. 56, pp. 231–238, 2014.

[179] S. Fakhraei, B. Huang, L. Raschid, and L. Getoor, "Network-based drug-target interaction prediction with probabilistic soft logic," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 5, pp. 775–787, 2014.

[180] D. Petrochilos, A. Shojaie, J. Gennari, and N. Abernethy, "Using random walks to identify cancer-associated modules in expression data," *BioData mining*, vol. 6, no. 1, pp. 1–25, 2013.

[181] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PloS one*, vol. 8, no. 5, p. e58977, 2013.

[182] C. Lei and J. Ruan, "A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity," *Bioinformatics*, vol. 29, no. 3, pp. 355–364, 2013.

[183] T. N. Jarada, J. G. Rokne, and R. Alhajj, "Snf-nn: computational method to predict drug-disease interactions using similarity network fusion and neural networks," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–20, 2021.

[184] Y. Wang, M. Guo, Y. Ren, L. Jia, and G. Yu, "Drug repositioning based on individual bi-random walks on a heterogeneous network," *BMC bioinformatics*, vol. 20, no. 15, pp. 1–13, 2019.

[185] Z. Tian, Z. Teng, S. Cheng, and M. Guo, "Computational drug repositioning using meta-path-based semantic network analysis," *BMC Systems Biology*, vol. 12, no. 9, pp. 123–134, 2018.

[186] H. Liu, Y. Song, J. Guan, L. Luo, and Z. Zhuang, "Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks," *BMC bioinformatics*, vol. 17, no. 17, pp. 269–277, 2016.

[187] J. Mullen, S. J. Cockell, P. Woollard, and A. Wipat, "An integrated data driven approach to drug repositioning using gene-disease associations," *PloS one*, vol. 11, no. 5, p. e0155811, 2016.

[188] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, and Y. Pan, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, 2016.

[189] H. Moghadam, M. Rahgozar, and S. Gharaghani, "Scoring multiple features to predict drug disease associations using information fusion and aggregation," *SAR and QSAR in Environmental Research*, vol. 27, no. 8, pp. 609–628, 2016.

[190] M. Oh, J. Ahn, and Y. Yoon, "A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions," *PLoS One*, vol. 9, no. 10, p. e111668, 2014.

[191] J. Yang, Z. Li, X. Fan, and Y. Cheng, "Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization," *Journal of chemical information and modeling*, vol. 54, no. 9, pp. 2562–2569, 2014.

[192] Y.-F. Huang, H.-Y. Yeh, and V.-W. Soo, "Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation," *BMC medical genomics*, vol. 6, no. 3, pp. 1–14, 2013.

[193] S. Zhao and S. Li, "A co-module approach for elucidating drug–disease associations and revealing their molecular basis," *Bioinformatics*, vol. 28, no. 7, pp. 955–961, 2012.

[194] D. Borsboom and A. O. Cramer, "Network analysis: an integrative approach to the structure of psychopathology," *Annual review of clinical psychology*, vol. 9, pp. 91–121, 2013.

[195] L.-Y. Dai, J.-X. Liu, R. Zhu, J. Wang, and S.-S. Yuan, "Logistic weighted profile-based bi-random walk for exploring mirna-disease associations," *Journal of Computer Science and Technology*, vol. 36, no. 2, pp. 276–287, 2021.

[196] Y.-W. Niu, G.-H. Wang, G.-Y. Yan, and X. Chen, "Integrating random walk and

binary regression to identify novel mirna-disease association," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–13, 2019.

[197] G. Xie, Z. Fan, Y. Sun, C. Wu, and L. Ma, "Wbnpmd: weighted bipartite network projection for microrna-disease association prediction," *Journal of Translational Medicine*, vol. 17, no. 1, pp. 1–11, 2019.

[198] X. Chen, D.-H. Zhang, and Z.-H. You, "A heterogeneous label propagation approach to explore the potential associations between mirna and disease," *Journal of translational medicine*, vol. 16, no. 1, pp. 1–14, 2018.

[199] J. Luo and Q. Xiao, "A novel approach for predicting microrna-disease associations by unbalanced bi-random walk on heterogeneous network," *Journal of biomedical informatics*, vol. 66, pp. 194–203, 2017.

[200] Z.-H. You, Z.-A. Huang, Z. Zhu, G.-Y. Yan, Z.-W. Li, Z. Wen, and X. Chen, "Pbmda: A novel and effective path-based computational model for mirna-disease association prediction," *PLoS computational biology*, vol. 13, no. 3, p. e1005455, 2017.

[201] X. Chen, C. C. Yan, X. Zhang, Z.-H. You, Y.-A. Huang, and G.-Y. Yan, "Hgimda: Heterogeneous graph inference for mirna-disease association prediction," *Oncotarget*, vol. 7, no. 40, p. 65257, 2016.

[202] C. Gu, B. Liao, X. Li, and K. Li, "Network consistency projection for human mirna-disease associations inference," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[203] P. Xuan, K. Han, Y. Guo, J. Li, X. Li, Y. Zhong, Z. Zhang, and J. Ding, "Prediction of potential disease-associated microrna based on random walk," *Bioinformatics*, vol. 31, no. 11, pp. 1805–1815, 2015.

[204] X. Chen, C. Clarence Yan, X. Zhang, Z. Li, L. Deng, Y. Zhang, and Q. Dai, "Rbmmmda: predicting multiple types of disease-microrna associations," *Scientific reports*, vol. 5, no. 1, pp. 1–13, 2015.

[205] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microrna-disease associations inference," *Scientific reports*, vol. 4, no. 1, pp. 1–10, 2014.

[206] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, *et al.*, "Prediction of micrornas associated with human diseases based on weighted k most similar neighbors," *PloS one*, vol. 8, no. 8, p. e70204, 2013.

[207] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu, and Y. Wang, "Prioritization of disease micrornas through a human phenome-micrornaome network," *BMC systems biology*, vol. 4, no. 1, pp. 1–9, 2010.

[208] Y. Long and J. Luo, "Wmghmda: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–18, 2019.

[209] Z.-A. Huang, X. Chen, Z. Zhu, H. Liu, G.-Y. Yan, Z.-H. You, and Z. Wen, "Pbhmda: path-based human microbe-disease association prediction," *Frontiers in microbiology*, vol. 8, p. 233, 2017.

[210] X. Shen, Y. Chen, X. Jiang, X. Hu, T. He, and J. Yang, "Predicting disease-microbe association by random walking on the heterogeneous network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 771–774, IEEE, 2016.

[211] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, 2017.

[212] S. Roy and V. Filkov, "Strong associations between microbe phenotypes and their network architecture," *Physical Review E*, vol. 80, no. 4, p. 040902, 2009.

[213] R. Xu and Q. Wang, "Phenopredict: a disease phenome-wide drug repositioning approach towards schizophrenia drug discovery," *Journal of biomedical informatics*, vol. 56, pp. 348–355, 2015.

[214] B. Blonder, T. Wey, A. Dornhaus, R. James, and A. Sih, "Temporal dynamics and network analysis. methods ecol evol 3: 958–972," 2012.

[215] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, "A dynamic network approach for the study of human phenotypes," *PLoS computational biology*, vol. 5, no. 4, p. e1000353, 2009.

[216] S. K. Sood and I. Mahajan, "Fog-cloud based cyber-physical system for distinguishing, detecting and preventing mosquito borne diseases," *Future Generation Computer Systems*, vol. 88, pp. 764–775, 2018.

[217] N. Carroll and I. Richardson, "Mapping a careflow network to assess the connectedness of connected health," *Health informatics journal*, vol. 25, no. 1, pp. 106–125, 2019.

[218] J. A. Merrill, B. Sheehan, K. M. Carley, and P. Stetson, "Transition networks in a cohort of patients with congestive heart failure," *Applied clinical informatics*, vol. 6, no. 03, pp. 548–564, 2015.

[219] J. A. Effken, K. M. Carley, S. Gephart, J. A. Verran, D. Bianchi, J. Reminga, and B. B. Brewer, "Using ora to explore the relationship of nursing unit communication to patient safety and quality outcomes," *International journal of medical informatics*, vol. 80, no. 7, pp. 507–517, 2011.

[220] S. Daminelli, J. M. Thomas, C. Durán, and C. V. Cannistraci, "Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks," *New Journal of Physics*, vol. 17, no. 11, p. 113037, 2015.

[221] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[222] D. Aune, T. Norat, P. Romundstad, and L. J. Vatten, "Whole grain and refined grain consumption and the risk of type 2 diabetes: a systematic review and dose–response meta-analysis of cohort studies," *European journal of epidemiology*, vol. 28, no. 11, pp. 845–858, 2013.

[223] L. Schwingshackl, G. Hoffmann, A.-M. Lampousi, S. Knüppel, K. Iqbal, C. Schwedhelm, A. Bechthold, S. Schlesinger, and H. Boeing, "Food groups and risk of type 2 diabetes mellitus: a systematic review and meta-analysis of prospective studies," *European journal of epidemiology*, vol. 32, no. 5, pp. 363–375, 2017.

[224] M. Ding, S. N. Bhupathiraju, M. Chen, R. M. van Dam, and F. B. Hu, "Caffeinated and decaffeinated coffee consumption and risk of type 2 diabetes: a systematic review and a dose-response meta-analysis," *Diabetes care*, vol. 37, no. 2, pp. 569–586, 2014.

[225] M. Jiang, R. Alhajj, and J. Rokne, "Behavior modeling in social networks.," 2018.

[226] J. Wu, G. Zhang, and Y. Ren, "A balanced modularity maximization link prediction model in social networks," *Information Processing & Management*, vol. 53, no. 1, pp. 295–307, 2017.

[227] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, and Y. Ju, "Prediction of microrna-disease associations based on social network analysis methods," *BioMed research international*, vol. 2015, 2015.

[228] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1100–1108, 2011.

[229] M. Pujari and R. Kanawati, "Supervised rank aggregation approach for link prediction in complex networks," in *Proceedings of the 21st international conference on world wide web*, pp. 1189–1196, 2012.

[230] S. Ghasemi and A. Zarei, "Improving link prediction in social networks using local and global features: a clustering-based approach," *Progress in Artificial Intelligence*, vol. 11, no. 1, pp. 79–92, 2022.

[231] M. Needham and A. E. Hodler, *Graph algorithms: practical examples in Apache Spark and Neo4j*. O'Reilly Media, 2019.

[232] "Eating Health Module Dataset — Kaggle." https://www.kaggle.com/bls/eating-health-module-dataset. Accessed: 2022-03-07.

[233] F. Rehman, O. Khalid, K. Bilal, S. A. Madani, *et al.*, "Diet-right: A smart food recommendation system," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 11, no. 6, pp. 2910–2925, 2017.

[234] S. Rani and M. Kumar, "Prediction of the mortality rate and framework for remote monitoring of pregnant women based on iot," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 24555–24571, 2021.

[235] D. S. Rajput, S. M. Basha, Q. Xin, T. R. Gadekallu, R. Kaluri, K. Lakshmanna, and P. K. R. Maddikunta, "Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of india," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2021.

[236] T. A. Ahanger, U. Tariq, M. Nusir, A. Aldaej, I. Ullah, and A. Sulman, "A novel iot–fog–cloud-based healthcare system for monitoring and predicting covid-19 outspread," *The Journal of Supercomputing*, vol. 78, no. 2, pp. 1783–1806, 2022.

[237] V. Balasubramanian, S. Vivekanandhan, and V. Mahadevan, "Pandemic tele-smart: a contactless tele-health system for efficient monitoring of remotely located covid-19 quarantine wards in india using near-field communication and natural language processing system," *Medical & biological engineering & computing*, vol. 60, no. 1, pp. 61–79, 2022.

[238] S. K. Sood, V. Sood, I. Mahajan, *et al.*, "An intelligent healthcare system for predicting and preventing dengue virus infection," *Computing*, pp. 1–39, 2021.

[239] C.-F. Lin, T.-X. Lin, C.-I. Lin, and C.-C. Chang, "A mobile cloud-based health promotion system for cardiovascular diseases," *Wireless Personal Communications*, vol. 108, no. 4, pp. 2179–2193, 2019.

[240] M. Pham, Y. Mengistu, H. M. Do, and W. Sheng, "Cloud-based smart home environment (coshe) for home healthcare," in *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 483–488, IEEE, 2016.

[241] P. Pouladzadeh, S. Shirmohammadi, A. Bakirov, A. Bulut, and A. Yassine, "Cloud-based svm for food categorization," *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5243–5260, 2015.

[242] Y. Zhang, D. Zhang, M. M. Hassan, A. Alamri, and L. Peng, "Cadre: Cloud-assisted drug recommendation service for online pharmacies," *Mobile Networks and Applications*, vol. 20, no. 3, pp. 348–355, 2015.

[243] M. Hussain, A. M. Khattak, W. A. Khan, I. Fatima, M. B. Amin, Z. Pervez, R. Batool, M. A. Saleem, M. Afzal, M. Faheem, *et al.*, "Cloud-based smart cdss for chronic diseases," *Health and Technology*, vol. 3, no. 2, pp. 153–175, 2013.

[244] "Eating Hints: Before, During, and After Cancer Treatment." https://www.cancer.gov/publications/patient-education/eating-hints. Accessed: 2022-03-23.

[245] "Healing foods reference database." https://www.healingfoodreference.com/. Accessed: 2022-03-23.

[246] "U.S. National Library of Medicine." https://clinicaltrials.gov/. Accessed: 2022-03-23.

[247] "Medical Subject Headings." https://www.nlm.nih.gov/mesh/. Accessed: 2022-03-23.

[248] S. Haykin, "Neural networks and learning machines 3rd ed. ny: Nyl pearson prentice hall," 2009.

[249] A. S. Al-Afify, G. El-Akabawy, N. M. El-Sherif, F. E.-N. A. El-Safty, and M. M. El-Habiby, "Avocado soybean unsaponifiables ameliorates cartilage and subchondral bone degeneration in mono-iodoacetate-induced knee osteoarthritis in rats," *Tissue and cell*, vol. 52, pp. 108–115, 2018.

[250] D. Lamichhane, C. Collins, F. Constantinescu, B. Walitt, M. Pettinger, C. Parks, and B. V. Howard, "Coffee and tea consumption in relation to risk of rheumatoid arthritis in the women's health initiative observational cohort," *Journal of clinical rheumatology: practical reports on rheumatic & musculoskeletal diseases*, vol. 25, no. 3, p. 127, 2019.

[251] C. Fang, X. Cai, S. Hayashi, S. Hao, H. Sakiyama, X. Wang, Q. Yang, S. Akira, S. Nishiguchi, N. Fujiwara, *et al.*, "Caffeine-stimulated muscle il-6 mediates alleviation of non-alcoholic fatty liver disease," *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, vol. 1864, no. 3, pp. 271–280, 2019.

[252] "NIH Research Matters." https://www.nih.gov/news-events/nih-research-matters/eating-red-meat-daily-triples-heart-disease-related-chemical. Accessed: 2022-05-05.

[253] "IBD-AID." https://www.umassmed.edu/nutrition/ibd/gastrointestinal/ibd/. Accessed: 2022-05-05.

[254] D. Fernandes and J. Bernardino, "Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb.," in *Data*, pp. 373–380, 2018.

[255] "What is Neo4j." https://neo4j.com/product/neo4j-graph-database/. Accessed: 2022-05-06.

[256] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[257] M. Kuhn, K. Johnson, *et al.*, *Applied predictive modeling*, vol. 26. Springer, 2013.

[258] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[259] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[260] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[261] "Coronavirus disease 2019." https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed: 2022-03-28.

[262] M. C. Chang, Y.-K. Park, B.-O. Kim, and D. Park, "Risk factors for disease progression in covid-19 patients," *BMC infectious diseases*, vol. 20, no. 1, pp. 1–6, 2020.

[263] B. Wang, R. Li, Z. Lu, and Y. Huang, "Does comorbidity increase the risk of patients with covid-19: evidence from meta-analysis," *Aging (albany NY)*, vol. 12, no. 7, p. 6049, 2020.

[264] A. Sanyaolu, C. Okorie, A. Marinkovic, R. Patidar, K. Younis, P. Desai, Z. Hosein, I. Padda, J. Mangat, and M. Altaf, "Comorbidity and its impact on patients with covid-19," *SN comprehensive clinical medicine*, vol. 2, no. 8, pp. 1069–1076, 2020.

[265] J. Yang, Y. Zheng, X. Gou, K. Pu, Z. Chen, Q. Guo, R. Ji, H. Wang, Y. Wang, and Y. Zhou, "Prevalence of comorbidities and its effects in patients infected with sars-cov-2: a systematic review and meta-analysis," *International journal of infectious diseases*, vol. 94, pp. 91–95, 2020.

[266] Q.-X. Long, X.-J. Tang, Q.-L. Shi, Q. Li, H.-J. Deng, J. Yuan, J.-L. Hu, W. Xu, Y. Zhang, F.-J. Lv, *et al.*, "Clinical and immunological assessment of asymptomatic sars-cov-2 infections," *Nature medicine*, vol. 26, no. 8, pp. 1200–1204, 2020.

[267] M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, and L. F. Ng, "The trinity of covid-19: immunity, inflammation and intervention," *Nature Reviews Immunology*, vol. 20, no. 6, pp. 363–374, 2020.

[268] P. C. Calder, "Nutrition, immunity and covid-19," *BMJ Nutrition, Prevention & Health*, vol. 3, no. 1, p. 74, 2020.

[269] A. Gasmi, S. Noor, T. Tippairote, M. Dadar, A. Menzel, and G. Bjørklund, "Individual risk management strategy and potential therapeutic options for the covid-19 pandemic," *Clinical Immunology*, vol. 215, p. 108409, 2020.

[270] F. Naja and R. Hamadeh, "Nutrition amid the covid-19 pandemic: a multi-level framework for action," *European journal of clinical nutrition*, vol. 74, no. 8, pp. 1117–1121, 2020.

[271] G. Muscogiuri, L. Barrea, S. Savastano, and A. Colao, "Nutritional recommendations for covid-19 quarantine," *European journal of clinical nutrition*, vol. 74, no. 6, pp. 850–851, 2020.

[272] M. J. Butler and R. M. Barrientos, "The impact of nutrition on covid-19 susceptibility and long-term consequences," *Brain, behavior, and immunity*, vol. 87, pp. 53–54, 2020.

[273] M. K. Singh, A. Mobeen, A. Chandra, S. Joshi, and S. Ramachandran, "A meta-analysis of comorbidities in covid-19: Which diseases increase the susceptibility of sars-cov-2 infection?," *Computers in biology and medicine*, vol. 130, p. 104219, 2021.

[274] I. Kyrou, T. Robbins, and H. S. Randeva, "Covid-19 and diabetes: No time to drag our feet during an untimely pandemic," *Journal of Diabetes and its Complications*, vol. 34, no. 9, p. 107621, 2020.

[275] P. Portincasa, M. Krawczyk, W. Smyk, F. Lammert, and A. Di Ciaula, "Covid-19 and non-alcoholic fatty liver disease: two intersecting pandemics," *European journal of clinical investigation*, vol. 50, no. 10, p. e13338, 2020.

[276] G. H. Prins, P. Olinga, *et al.*, "Potential implications of covid-19 in non-alcoholic fatty liver disease," *Liver Int*, vol. 40, no. 10, p. 2568, 2020.

[277] R. Huang, L. Zhu, J. Wang, L. Xue, L. Liu, X. Yan, S. Huang, Y. Li, X. Yan, B. Zhang, *et al.*, "Clinical features of patients with covid-19 with nonalcoholic

fatty liver disease," *Hepatology communications*, vol. 4, no. 12, pp. 1758–1768, 2020.

[278] S. Razeghi Jahromi, H. Moradi Tabriz, M. Togha, S. Ariyanfar, Z. Ghorbani, S. Naeeni, S. Haghighi, A. Jazayeri, M. Montazeri, M. Talebpour, *et al.*, "The correlation between serum selenium, zinc, and covid-19 severity: An observational study," *BMC Infectious Diseases*, vol. 21, no. 1, pp. 1–9, 2021.

[279] J. N. Losso, M. N. Losso, J. N. Inungu, J. W. Finley, *et al.*, "The young age and plant-based diet hypothesis for low sars-cov-2 infection and covid-19 pandemic in sub-saharan africa," *Plant Foods for Human Nutrition*, vol. 76, no. 3, pp. 270–280, 2021.

[280] J. Rocha, T. Basra, B. El Kurdi, and C. Venegas-Borsellino, "Effects of potential micro-and macro-nutrients in combatting covid-19," *Current Surgery Reports*, vol. 9, no. 10, pp. 1–6, 2021.

[281] D. M. Abdulah and A. Hassan, "Relation of dietary factors with infection and mortality rates of covid-19 across the world," *The journal of nutrition, health & aging*, vol. 24, no. 9, pp. 1011–1018, 2020.

[282] A. Gasmi, T. Tippairote, P. K. Mujawdiya, M. Peana, A. Menzel, M. Dadar, A. G. Benahmed, and G. Bjørklund, "Micronutrients as immunomodulatory tools for covid-19 management," *Clinical Immunology*, vol. 220, p. 108545, 2020.

[283] R. Jayawardena, P. Sooriyaarachchi, M. Chourdakis, C. Jeewandara, and P. Ranasinghe, "Enhancing immunity in viral infections, with special emphasis on covid-19: A review," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 367–382, 2020.

[284] D. O. Meltzer, T. J. Best, H. Zhang, T. Vokes, V. Arora, and J. Solway, "Association of vitamin d status and other clinical characteristics with covid-19 test results," *JAMA network open*, vol. 3, no. 9, pp. e2019722–e2019722, 2020.

[285] A. Abobaker, A. Alzwi, and A. H. A. Alraied, "Overview of the possible role of vitamin c in management of covid-19," *Pharmacological Reports*, vol. 72, no. 6, pp. 1517–1528, 2020.

[286] S. Budhwar, K. Sethi, and M. Chakraborty, "A rapid advice guideline for the prevention of novel coronavirus through nutritional intervention," *Current Nutrition Reports*, vol. 9, no. 3, pp. 119–128, 2020.

[287] "IBD-types." https://www.cdc.gov/ibd/what-is-IBD.htm. Accessed: 2022-05-30.

[288] Y.-Z. Zhang and Y.-Y. Li, "Inflammatory bowel disease: pathogenesis," *World journal of gastroenterology: WJG*, vol. 20, no. 1, p. 91, 2014.

[289] S. Kakodkar and E. A. Mutlu, "Diet as a therapeutic option for adult inflammatory bowel disease," *Gastroenterology Clinics*, vol. 46, no. 4, pp. 745–767, 2017.

[290] J. K. Hou, B. Abraham, and H. El-Serag, "Dietary intake and risk of developing inflammatory bowel disease: a systematic review of the literature," *Official journal of the American College of Gastroenterology— ACG*, vol. 106, no. 4, pp. 563–573, 2011.

[291] Y. Yang, L. Xiang, and J. He, "Beverage intake and risk of crohn disease: A meta-analysis of 16 epidemiological studies," *Medicine*, vol. 98, no. 21, 2019.

[292] M. Octoratou, E. Merikas, G. Malgarinos, C. Stanciu, and J. Triantafillidis, "A prospective study of pre-illness diet in newly diagnosed patients with crohn's disease.," *Revista medico-chirurgicala a Societatii de Medici si Naturalisti din Iasi*, vol. 116, no. 1, pp. 40–49, 2012.

[293] P. McDonald and V. Fazio, "What can crohn's patients eat?," *European Journal of Clinical Nutrition*, vol. 42, no. 8, pp. 703–708, 1988.

[294] M.-J. Lesot, M. Rifqi, and H. Benhadda, "Similarity measures for binary and numerical data: a survey," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 1, no. 1, pp. 63–84, 2009.

[295] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.

[296] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[297] H.-Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and fisher's exact test," *Restorative dentistry & endodontics*, vol. 42, no. 2, pp. 152–155, 2017.

[298] "6.3. Preprocessing data — scikit-learn 1.0.1 documentation." https://scikit-learn.org/stable/modules/preprocessing.html. Accessed: 2022-03-29.

[299] P. Tessari and A. Lante, "A multifunctional bread rich in beta glucans and low in starch improves metabolic control in type 2 diabetes: a controlled trial," *Nutrients*, vol. 9, no. 3, p. 297, 2017.

[300] A. E. Yanni, N. S. Stamataki, P. Konstantopoulos, M. Stoupaki, A. Abeliatis, I. Nikolakea, D. Perrea, V. T. Karathanos, and N. Tentolouris, "Controlling type-2 diabetes by inclusion of cr-enriched yeast bread in the daily dietary pattern: a randomized clinical trial," *European journal of nutrition*, vol. 57, no. 1, pp. 259–267, 2018.

[301] C. Kyrø, A. Tjønneland, K. Overvad, A. Olsen, and R. Landberg, "Higher whole-grain intake is associated with lower risk of type 2 diabetes among middle-aged men and women: the danish diet, cancer, and health cohort," *The Journal of nutrition*, vol. 148, no. 9, pp. 1434–1444, 2018.

[302] U. Ericson, E. Sonestedt, B. Gullberg, S. Hellstrand, G. Hindy, E. Wirfält, and M. Orho-Melander, "High intakes of protein and processed meat associate with increased incidence of type 2 diabetes," *British journal of nutrition*, vol. 109, no. 6, pp. 1143–1153, 2013.

[303] S. Liatis, P. Tsapogas, E. Chala, C. Dimosthenopoulos, K. Kyriakopoulos, E. Kapantais, and N. Katsilambros, "The consumption of bread enriched with betaglucan reduces ldl-cholesterol and improves insulin resistance in patients with type 2 diabetes," *Diabetes & metabolism*, vol. 35, no. 2, pp. 115–120, 2009.

[304] M. Akhoundan, Z. Shadman, P. Jandaghi, M. Aboeerad, B. Larijani, Z. Jamshidi, H. Ardalani, and M. Khoshniat Nikoo, "The association of bread and rice with metabolic factors in type 2 diabetic patients," *PloS one*, vol. 11, no. 12, p. e0167921, 2016.

[305] H. Haimoto, S. Watanabe, K. Maeda, T. Murase, and K. Wakai, "Reducing carbohydrate from individual sources has differential effects on glycosylated hemoglobin in type 2 diabetes mellitus patients on moderate low-carbohydrate diets," *Diabetes & metabolism journal*, vol. 45, no. 3, pp. 390–403, 2020.

[306] A. Basiak-Rasala, D. Rozanska, and K. Zatonska, "Food groups in dietary prevention of type 2 diabetes," *Roczniki Państwowego Zakładu Higieny*, vol. 70, no. 4, 2019.

[307] S. G. N. Gabrial, M. Shakib, M. S. M. A. Haleem, G. N. Gabrial, and F. A. El-Shobaki, "Hypoglycemic potential of supplementation with a vegetable and legume juice formula in type 2 diabetic patients.," *Pakistan Journal of Biological Sciences: PJBS*, vol. 23, no. 2, pp. 132–138, 2020.

[308] A. K. Tiwari, "Revisiting "vegetables" to combat modern epidemic of imbalanced glucose homeostasis," *Pharmacognosy Magazine*, vol. 10, no. Suppl 2, p. S207, 2014.

[309] S. H. Ley, O. Hamdy, V. Mohan, and F. B. Hu, "Prevention and management of type 2 diabetes: dietary components and nutritional strategies," *The Lancet*, vol. 383, no. 9933, pp. 1999–2007, 2014.

[310] F. Jannasch, J. Kröger, and M. B. Schulze, "Dietary patterns and type 2 diabetes:

a systematic literature review and meta-analysis of prospective studies," *The Journal of nutrition*, vol. 147, no. 6, pp. 1174–1182, 2017.

[311] J. Salas-Salvadó, M. Martinez-Gonzalez, M. Bulló, and E. Ros, "The role of diet in the prevention of type 2 diabetes," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 21, pp. B32–B48, 2011.

[312] G. D. Carrasquilla, H. Jakupović, and T. O. Kilpeläinen, "Dietary fat and the genetic risk of type 2 diabetes," *Current diabetes reports*, vol. 19, no. 11, pp. 1–6, 2019.

[313] J. Tay, N. D. Luscombe-Marsh, C. H. Thompson, M. Noakes, J. D. Buckley, G. A. Wittert, W. S. Yancy Jr, and G. D. Brinkworth, "Comparison of low-and high-carbohydrate diets for type 2 diabetes management: a randomized trial," *The American journal of clinical nutrition*, vol. 102, no. 4, pp. 780–790, 2015.

[314] R. M. M. Santos and D. R. A. Lima, "Coffee consumption, obesity and type 2 diabetes: A mini-review," *European journal of nutrition*, vol. 55, no. 4, pp. 1345–1358, 2016.

[315] M. Guasch-Ferré, J. Merino, Q. Sun, M. Fitó, and J. Salas-Salvadó, "Dietary polyphenols, mediterranean diet, prediabetes, and type 2 diabetes: a narrative review of the evidence," *Oxidative medicine and cellular longevity*, vol. 2017, 2017.

[316] M. Carlström and S. C. Larsson, "Coffee consumption and reduced risk of developing type 2 diabetes: a systematic review with meta-analysis," *Nutrition reviews*, vol. 76, no. 6, pp. 395–417, 2018.

[317] M. Neuenschwander, A. Ballon, K. S. Weber, T. Norat, D. Aune, L. Schwingshackl, and S. Schlesinger, "Role of diet in type 2 diabetes incidence: umbrella review of meta-analyses of prospective observational studies," *bmj*, vol. 366, 2019.

[318] A. Pan, Q. Sun, A. M. Bernstein, M. B. Schulze, J. E. Manson, W. C. Willett, and F. B. Hu, "Red meat consumption and risk of type 2 diabetes: 3 cohorts

of us adults and an updated meta-analysis," *The American journal of clinical nutrition*, vol. 94, no. 4, pp. 1088–1096, 2011.

[319] S. S. Shetty, P. K. Shetty, *et al.*, "ω-6/ω-3 fatty acid ratio as an essential predictive biomarker in the management of type 2 diabetes mellitus," *Nutrition*, vol. 79, p. 110968, 2020.

[320] A. Misra, N. Singhal, and L. Khurana, "Obesity, the metabolic syndrome, and type 2 diabetes in developing countries: role of dietary fats and oils," *Journal of the American College of Nutrition*, vol. 29, no. sup3, pp. 289S–301S, 2010.

[321] K. M. Davison and N. J. Temple, "Cereal fiber, fruit fiber, and type 2 diabetes: Explaining the paradox," *Journal of Diabetes and its Complications*, vol. 32, no. 2, pp. 240–245, 2018.

[322] L. Zhao, F. Zhang, X. Ding, G. Wu, Y. Y. Lam, X. Wang, H. Fu, X. Xue, C. Lu, J. Ma, *et al.*, "Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes," *Science*, vol. 359, no. 6380, pp. 1151–1156, 2018.

[323] N. S. Pcsolyar and B. C. De Jonghe, "Examining the use of dietary fiber in reducing the risk of type 2 diabetes mellitus in latino youth," *Journal of Transcultural Nursing*, vol. 25, no. 3, pp. 249–255, 2014.

[324] C.-H. Jung and K. M. Choi, "Impact of high-carbohydrate diet on metabolic parameters in patients with type 2 diabetes," *Nutrients*, vol. 9, no. 4, p. 322, 2017.

[325] B. J. O'Neill, "Effect of low-carbohydrate diets on cardiometabolic risk, insulin resistance, and metabolic syndrome," *Current Opinion in Endocrinology, Diabetes and Obesity*, vol. 27, no. 5, pp. 301–307, 2020.

[326] P. Qin, Q. Li, Y. Zhao, Q. Chen, X. Sun, Y. Liu, H. Li, T. Wang, X. Chen, Q. Zhou, *et al.*, "Sugar and artificially sweetened beverages and risk of obesity, type 2 diabetes mellitus, hypertension, and all-cause mortality: a dose–response

meta-analysis of prospective cohort studies," *European journal of epidemiology*, vol. 35, no. 7, pp. 655–671, 2020.

[327] L. Paglia, "The sweet danger of added sugars," *Eur. J. Paediatr. Dent*, vol. 20, p. 89, 2019.

[328] F. B. Hu and V. S. Malik, "Sugar-sweetened beverages and risk of obesity and type 2 diabetes: epidemiologic evidence," *Physiology & behavior*, vol. 100, no. 1, pp. 47–54, 2010.

[329] C. S. Tsilas, R. J. de Souza, S. B. Mejia, A. Mirrahimi, A. I. Cozma, V. H. Jayalath, V. Ha, R. Tawfik, M. Di Buono, A. L. Jenkins, *et al.*, "Relation of total sugars, fructose and sucrose with incident type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies," *Cmaj*, vol. 189, no. 20, pp. E711–E720, 2017.

[330] V. Behrouz, A. Dastkhosh, and G. Sohrab, "Overview of dietary supplements on patients with type 2 diabetes," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 325–334, 2020.

[331] A. G. Pittas, R. Jorde, T. Kawahara, and B. Dawson-Hughes, "Vitamin d supplementation for prevention of type 2 diabetes mellitus: to d or not to d?," *The Journal of Clinical Endocrinology & Metabolism*, vol. 105, no. 12, pp. 3721–3733, 2020.

[332] M. E. Patterson, J. K. Yee, P. Wahjudi, C. S. Mao, and W.-N. P. Lee, "Acute metabolic responses to high fructose corn syrup ingestion in adolescents with overweight/obesity and diabetes," *Journal of nutrition & intermediary metabolism*, vol. 14, pp. 1–7, 2018.

[333] S. Zelber-Sagi, F. Salomone, and L. Mlynarsky, "The mediterranean dietary pattern as the diet of choice for non-alcoholic fatty liver disease: evidence and plausible mechanisms," *Liver International*, vol. 37, no. 7, pp. 936–949, 2017.

[334] M. Razavi Zade, M. H. Telkabadi, F. Bahmani, B. Salehi, S. Farshbaf, and Z. Asemi, "The effects of dash diet on weight loss and metabolic status in adults with non-alcoholic fatty liver disease: a randomized clinical trial," *Liver international*, vol. 36, no. 4, pp. 563–571, 2016.

[335] K. Riazi, M. Raman, L. Taylor, M. G. Swain, and A. A. Shaheen, "Dietary patterns and components in nonalcoholic fatty liver disease (nafld): what key messages can health care providers offer?," *Nutrients*, vol. 11, no. 12, p. 2878, 2019.

[336] E. S. George, A. Forsyth, C. Itsiopoulos, A. J. Nicoll, M. Ryan, S. Sood, S. K. Roberts, and A. C. Tierney, "Practical dietary recommendations for the prevention and management of nonalcoholic fatty liver disease in adults," *Advances in nutrition*, vol. 9, no. 1, pp. 30–40, 2018.

[337] U. Hayat, A. A. Siddiqui, H. Okut, S. Afroz, S. Tasleem, and A. Haris, "The effect of coffee consumption on the non-alcoholic fatty liver disease and liver fibrosis: A meta-analysis of 11 epidemiological studies," *Annals of Hepatology*, vol. 20, p. 100254, 2021.

[338] A. Yesil and Y. Yilmaz, "coffee consumption, the metabolic syndrome and non-alcoholic fatty liver disease," *Alimentary pharmacology & therapeutics*, vol. 38, no. 9, pp. 1038–1044, 2013.

[339] K. Wijarnpreecha, C. Thongprayoon, and P. Ungprasert, "Coffee consumption and risk of nonalcoholic fatty liver disease: a systematic review and meta-analysis," *European journal of gastroenterology & hepatology*, vol. 29, no. 2, pp. e8–e12, 2017.

[340] S. Zelber-Sagi, D. Ivancovsky-Wajcman, N. F. Isakov, M. Webb, D. Orenstein, O. Shibolet, and R. Kariv, "High red and processed meat consumption is associated with non-alcoholic fatty liver disease and insulin resistance," *Journal of hepatology*, vol. 68, no. 6, pp. 1239–1246, 2018.

[341] P. Mirmiran, Z. Amirhamidi, H.-S. Ejtahed, Z. Bahadoran, and F. Azizi, "Relationship between diet and non-alcoholic fatty liver disease: a review article," *Iranian journal of public health*, vol. 46, no. 8, p. 1007, 2017.

[342] T. Jensen, M. F. Abdelmalek, S. Sullivan, K. J. Nadeau, M. Green, C. Roncal, T. Nakagawa, M. Kuwabara, Y. Sato, D.-H. Kang, *et al.*, "Fructose and sugar: A major mediator of non-alcoholic fatty liver disease," *Journal of hepatology*, vol. 68, no. 5, pp. 1063–1075, 2018.

[343] M. S. Mundi, S. Velapati, J. Patel, T. A. Kellogg, B. K. Abu Dayyeh, and R. T. Hurt, "Evolution of nafld and its management," *Nutrition in Clinical Practice*, vol. 35, no. 1, pp. 72–84, 2020.

[344] E. Á. Hernández, S. Kahl, A. Seelig, P. Begovatz, M. Irmler, Y. Kupriyanova, B. Nowotny, P. Nowotny, C. Herder, C. Barosa, *et al.*, "Acute dietary fat intake initiates alterations in energy metabolism and insulin resistance," *The Journal of clinical investigation*, vol. 127, no. 2, pp. 695–708, 2017.

[345] L. Hodson, F. Rosqvist, and S. A. Parry, "The influence of dietary fatty acids on liver fat content and metabolism," *Proceedings of the Nutrition Society*, vol. 79, no. 1, pp. 30–41, 2020.

[346] A. Pérez-Montes de Oca, M. T. Julián, A. Ramos, M. Puig-Domingo, and N. Alonso, "Microbiota, fiber, and nafld: Is there any connection?," *Nutrients*, vol. 12, no. 10, p. 3100, 2020.

[347] M. Krawczyk, D. Maciejewska, K. Ryterska, M. Czerwińka-Rogowska, D. Jamioł-Milc, K. Skonieczna-Żydecka, P. Milkiewicz, J. Raszeja-Wyszomirska, and E. Stachowska, "Gut permeability might be improved by dietary fiber in individuals with nonalcoholic fatty liver disease (nafld) undergoing weight reduction," *Nutrients*, vol. 10, no. 11, p. 1793, 2018.

[348] H. Zhao, A. Yang, L. Mao, Y. Quan, J. Cui, and Y. Sun, "Association between

dietary fiber intake and non-alcoholic fatty liver disease in adults," *Frontiers in nutrition*, p. 269, 2020.

[349] S. Chiu, K. Mulligan, and J.-M. Schwarz, "Dietary carbohydrates and fatty liver disease: de novo lipogenesis," *Current Opinion in Clinical Nutrition & Metabolic Care*, vol. 21, no. 4, pp. 277–282, 2018.

[350] S. Softic, D. E. Cohen, and C. R. Kahn, "Role of dietary fructose and hepatic de novo lipogenesis in fatty liver disease," *Digestive diseases and sciences*, vol. 61, no. 5, pp. 1282–1293, 2016.

[351] K. Kakleas, F. Christodouli, and K. Karavanaki, "Nonalcoholic fatty liver disease, insulin resistance, and sweeteners: a literature review," *Expert review of endocrinology & metabolism*, vol. 15, no. 2, pp. 83–93, 2020.

[352] B. J. Perumpail, A. A. Li, U. Iqbal, S. Sallam, N. D. Shah, W. Kwong, G. Cholankeril, D. Kim, and A. Ahmed, "Potential therapeutic benefits of herbs and supplements in patients with nafld," *Diseases*, vol. 6, no. 3, p. 80, 2018.

[353] B. Kilchoer, A. Vils, B. Minder, T. Muka, M. Glisic, and L. Bally, "Efficacy of dietary supplements to reduce liver fat," *Nutrients*, vol. 12, no. 8, p. 2302, 2020.

[354] R. S. Hamida, A. Shami, M. A. Ali, Z. N. Almohawes, A. E. Mohammed, and M. M. Bin-Meferij, "Kefir: A protective dietary supplementation against viral infection," *Biomedicine & Pharmacotherapy*, vol. 133, p. 110974, 2021.

[355] S. C. Tyagi and M. Singh, "Multi-organ damage by covid-19: congestive (cardiopulmonary) heart failure, and blood-heart barrier leakage," *Molecular and Cellular Biochemistry*, vol. 476, no. 4, pp. 1891–1895, 2021.

[356] S. M. Thota, V. Balan, and V. Sivaramakrishnan, "Natural products as home-based prophylactic and symptom management agents in the setting of covid-19," *Phytotherapy Research*, vol. 34, no. 12, pp. 3148–3167, 2020.

[357] A. Pieroni, I. Vandebroek, J. Prakofjewa, R. W. Bussmann, N. Y. Paniagua-Zambrana, A. Maroyi, L. Torri, D. M. Zocchi, A. T. Dam, S. M. Khan, *et al.*,

"Taming the pandemic? the importance of homemade plant-based foods and beverages as community responses to covid-19," 2020.

[358] R. Somasundaram, A. Choraria, and M. Antonysamy, "An approach towards development of monoclonal igy antibodies against sars cov-2 spike protein (s) using phage display method: A review," *International immunopharmacology*, vol. 85, p. 106654, 2020.

[359] S. Wei, S. Duan, X. Liu, H. Wang, S. Ding, Y. Chen, J. Xie, J. Tian, N. Yu, Y. Li, *et al.*, "Chicken egg yolk antibodies (igys) block the binding of multiple sars-cov-2 spike protein variants to human ace2," *International immunopharmacology*, vol. 90, p. 107172, 2021.

[360] J. M. Pérez de la Lastra, V. Baca-González, P. Asensio-Calavia, S. González-Acosta, and A. Morales-delaNuez, "Can immunization of hens provide oral-based therapeutics against covid-19?," *Vaccines*, vol. 8, no. 3, p. 486, 2020.

[361] Y. Lu, Y. Wang, Z. Zhang, J. Huang, M. Yao, G. Huang, Y. Ge, P. Zhang, H. Huang, Y. Wang, *et al.*, "Generation of chicken igy against sars-cov-2 spike protein and epitope mapping," *Journal of Immunology Research*, vol. 2020, 2020.

[362] M. I. Khalil, M. A. Salih, and A. A. Mustafa, "Broad beans (vicia faba) and the potential to protect from covid-19 coronavirus infection," *Sudanese Journal of Paediatrics*, vol. 20, no. 1, p. 10, 2020.

[363] M. M. Al-Sanea, N. Abelyan, M. A. Abdelgawad, A. Musa, M. M. Ghoneim, T. Al-Warhi, N. Aljaeed, O. J. Alotaibi, T. S. Alnusaire, S. F. Abdelwahab, *et al.*, "Strawberry and ginger silver nanoparticles as potential inhibitors for sars-cov-2 assisted by in silico modeling and metabolic profiling," *Antibiotics*, vol. 10, no. 7, p. 824, 2021.

[364] A. Singh and A. Mishra, "Leucoefdin a potential inhibitor against sars cov-2 mpro," *Journal of Biomolecular Structure and Dynamics*, vol. 39, no. 12, pp. 4427–4432, 2021.

[365] K. S. Hossain, M. G. Hossain, A. Moni, M. M. Rahman, U. H. Rahman, M. Alam, S. Kundu, M. M. Rahman, M. A. Hannan, and M. J. Uddin, "Prospects of honey in fighting against covid-19: pharmacological insights and therapeutic promises," *Heliyon*, vol. 6, no. 12, p. e05798, 2020.

[366] M. A. Al-Hatamleh, M. M. Hatmal, K. Sattar, S. Ahmad, M. Z. Mustafa, M. D. C. Bittencourt, and R. Mohamud, "Antiviral and immunomodulatory effects of phytochemicals from honey against covid-19: Potential mechanisms of action and future directions," *Molecules*, vol. 25, no. 21, p. 5017, 2020.

[367] K. Esposito, M. I. Maiorino, G. Bellastella, P. Chiodini, D. Panagiotakos, and D. Giugliano, "A journey into a mediterranean diet and type 2 diabetes: a systematic review with meta-analyses," *BMJ open*, vol. 5, no. 8, p. e008222, 2015.

[368] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, *et al.*, "Fast discovery of association rules.," *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[369] T. Zhan, A. Ali, J. G. Choi, M. Lee, J. Leung, E. S. Dellon, J. J. Garber, and C. Hur, "Model to determine the optimal dietary elimination strategy for treatment of eosinophilic esophagitis," *Clinical Gastroenterology and Hepatology*, vol. 16, no. 11, pp. 1730–1737, 2018.

[370] C. A. Liacouras, "Eosinophilic esophagitis: treatment in 2005," *Current opinion in gastroenterology*, vol. 22, no. 2, pp. 147–152, 2006.

[371] R. Yang, L. Xue, L. Zhang, X. Wang, X. Qi, J. Jiang, L. Yu, X. Wang, W. Zhang, Q. Zhang, *et al.*, "Phytosterol contents of edible oils and their contributions to estimated phytosterol intake in the chinese diet," *Foods*, vol. 8, no. 8, p. 334, 2019.

[372] A. B. Cohen, D. Lee, M. D. Long, M. D. Kappelman, C. F. Martin, R. S. Sandler, and J. D. Lewis, "Dietary patterns and self-reported associations of diet

with symptoms of inflammatory bowel disease," *Digestive diseases and sciences*, vol. 58, no. 5, pp. 1322–1328, 2013.

[373] J. D. Spence, "Trimethylamine n-oxide: not just red meat—egg yolk and renal function are also important," *European Heart Journal*, vol. 40, no. 42, pp. 3498–3498, 2019.

[374] J. D. Spence, "Nutrition and risk of stroke," *Nutrients*, vol. 11, no. 3, p. 647, 2019.

[375] P. Zou, "Traditional chinese medicine, food therapy, and hypertension control: a narrative review of chinese literature," *The American Journal of Chinese Medicine*, vol. 44, no. 08, pp. 1579–1594, 2016.

[376] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[377] C. M. Triggs, K. Munday, R. Hu, A. G. Fraser, R. B. Gearry, M. L. Barclay, and L. R. Ferguson, "Dietary factors in chronic inflammation: food tolerances and intolerances of a new zealand caucasian crohn's disease population," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 690, no. 1-2, pp. 123–138, 2010.

[378] S. P. Therkelsen, G. Hetland, T. Lyberg, I. Lygren, and E. Johnson, "Effect of a medicinal agaricus blazei murill-based mushroom extract, andosan™, on symptoms, fatigue and quality of life in patients with ulcerative colitis in a randomized single-blinded placebo controlled study," *PLoS One*, vol. 11, no. 3, p. e0150191, 2016.

[379] J. Benjamin, G. Makharia, V. Ahuja, K. Anand Rajan, M. Kalaivani, S. D. Gupta, and Y. K. Joshi, "Glutamine and whey protein improve intestinal permeability and morphology in patients with crohn's disease: a randomized controlled trial," *Digestive diseases and sciences*, vol. 57, no. 4, pp. 1000–1012, 2012.

[380] A. Chawla, P. I. Karl, and S. E. Fisher, "Effect of n-3 polyunsaturated fatty acid supplemented diet on neutrophil-mediated ileal permeability and neutrophil function in the rat.," *Journal of the American College of Nutrition*, vol. 14, no. 3, pp. 258–263, 1995.

[381] A. James, "Breakfast and crohn's disease.," *Br Med J*, vol. 1, no. 6066, pp. 943–945, 1977.

[382] S. Bentz, M. Hausmann, H. Piberger, S. Kellermeier, S. Paul, L. Held, W. Falk, F. Obermeier, M. Fried, J. Schölmerich, *et al.*, "Clinical relevance of igg antibodies against food antigens in crohn's disease: a double-blind cross-over diet intervention study," *Digestion*, vol. 81, no. 4, pp. 252–264, 2010.

[383] L. Kinsey and S. Burden, "A survey of people with inflammatory bowel disease to investigate their views of food and nutritional issues," *European journal of clinical nutrition*, vol. 70, no. 7, pp. 852–854, 2016.

List of Publications

- SCI Indexed Journal [Published]

    1. R Toor, I Chana "Network Analysis as a Computational Technique and Its Benefaction for Predictive Analysis of Healthcare Data: A Systematic Review," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 1689-711, May 2021. [**Impact Factor:8.171**]
    2. R Toor, Chana I, "Exploring diet associations with Covid-19 and other diseases:a Network Analysis–based approach," *Medical biological engineering computing*, vol. 60, no. 4, pp. 991-1013, April 2022 [**Impact Factor:3.079**]
    3. R Toor, I Chana, "DIDACE: Literature Mining and Exploration of Disease-Diet Associations," *Journal of Information Science and Engineering*, vol. 38, no. 1, Jan 2022 [**Impact Factor:0.541**]

- SCI Indexed Journal [Under Review]

    1. R Toor, I Chana "CloudMenu: Cloud based Network Analysis for Disease-Diet Associations and Recommendations," *Mobile Networks and Applications* [Under Review] [Impact Factor:3.077]
    2. R Toor, I Chana "DAPNEML: Disease-Diet Associations Prediction in a NEtwork using a Machine Learning based approach," *Information Systems Frontiers* [Under Review] [Impact Factor:5.261]

- Scopus-Indexed Journal

    1. R Toor and I Chana "Network Analysis of Disease-Diet Associations: A

Healthcare Perspective," *International Journal of Biology Today's World, Special Issue On: Role of IT in Bioinformatics and Neuroinformatics*, 2017

- Conference/Symposium

1. R Toor and I Chana "Cloud based Network Analysis Model for Predicting Disease-Diet Associations," *6th Indian Symposium on Computer Systems (IndoSys) organized by Indian Institute of Science (IISc), Bangalore*, 19-20th July 2019. [Poster Presentation]