REALTIME GENETIC SEQUENCE VERIFICATION WITH DISEASE MAPPING AND POLYMORPHISM

Submitted in partial fulfillment of the
requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

Abhishek Reddy Vaddepally (Reg.No - 39110007)
Eswara Sree Ram Ala (Reg.No – 39110039)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

APRIL - 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A‖ grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Abhishek Reddy Vaddepally (3911007) and Eswara Sree Ram Ala (39110039)** who carried out the Project Phase-1 entitled **"REALTIME GENETIC SEQUENCE VERIFICATION WITH DISEASE MAPPING AND POLYMORPHISM"** under my supervision from December 2022 to April 2023.

Internal Guide

Dr.SANKARI M

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on 24.04.23

Internal Examiner

External Examiner

DECLARATION

I, **Abhishek Reddy Vaddepally (Reg.No- 39110007) and Eswara Sree Ram Ala (39110039),** hereby declare that the Project Phase-1 Report entitled **“REALTIME GENETIC SEQUENCE VERIFICATION WITH DISEASE MAPPING AND POLYMORPHISM”** done by me under the guidance of **Dr.Sankari M** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

Abishek

DATE: 24-04-2023

PLACE: Chennai **SIGNATURE OF THECANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D**, **Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.Sankari M,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-1 project work.

I wish to express my thanks toall Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

In the Existing system, Gene dependency networks often undergo changes with respect to different disease states. Understanding how these networks rewire between two different disease states is an important task in genomic research. In the Proposed system, a new differential network inference model which identifies gene network rewiring by combining gene expression and gene mutation data. Similarities and differences between different data types are learned via a group bridge penalty function. In the Modification process, project is to recommend Diet pattern or any other natural drugs which can be recommended to those people who is expected to get into the disease in a course of time by verifying the mutated Genes.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Characterizing gene dependency networks can help to understand the causes and mechanisms of diseases. It is well established that gene dependency networks often undergo changes with respect to different disease states. The development and progression of cancers often involve aberrations in gene networks. Therefore, inferring the differential networks between two disease states from high throughput data, which can reveal the structure changes of gene networks, is of great interest in the field of bioinformatics. Differential networks are often inferred from gene expression data based on Gaussian Markov networks. By assuming that the gene expression measurements follow a multivariate Gaussian distribution, the conditional dependencies among genes can be directly determined by the nonzero elements of the precision matrix (the inverse covariance matrix). Two genes are conditionally independent given the other genes if and only if the corresponding element of the precision matrix is zero. Therefore, the state-specific gene networks can be modeled by the corresponding precision matrices, and the differential network between two states can be modeled as the difference between the two state-specific precision matrices. There are two main types of approaches to estimate the precision matrix difference.

The most straightforward one is to separately estimate the state-specific precision matrices first and then compute their difference (called indirect estimation methods). This type of methods often applies sparse penalties to the individual precision matrices to deal with the high dimensional cases where the number of genes is larger than the number of observations. This would be suboptimal when the individual precision matrices are not sparse but only their difference is sparse. The other type of approaches is to directly estimate the precision matrix difference without estimating the individual state-specific precision matrices (called direct estimation methods). Compared with the indirect estimation methods, the direct methods have an advantage of using much smaller sample size to achieve competitive performance.

1.1 DOMAIN INTRODUCTION

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, duration, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on. So, we can implement big data in our project because every employ has instructed information so we can make analysis on this data.

A. Characteristics of Big Data

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. The term “big data” is relatively new in IT and business. However, several researchers and practitioners have utilized the term in previous literature. For instance, referred to big data as a large volume of scientific data for visualization. Several definitions of big data currently exist.” Meanwhile and defined big data as characterized by three Vs: volume, variety, and velocity. The terms volume, variety, and velocity were originally introduced by Gartner to describe the elements of big data challenges. IDC also defined big data technologies as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high velocity capture, discovery, and/or analysis.” specified that big data is not only characterized by the three Vs mentioned above but may also extend to four Vs, namely, volume, variety, velocity, and value This 4V definition is widely recognized because it highlights the meaning and necessity of big data.

Volume
Value
Big Data
Variety
Velocity

Fig 1.1 Four V's of Big Data

Volume - Refers to the amount of all types of data generated from different sources and continue to expand. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data challenge produced an interesting result similar to that in the examination of the predictability of human behavior patterns or means to share data based on human mobility and visualization techniques for complex data.

Variety-Refers to the different types of data collected via sensors, smart phones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data.

Velocity -Refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources

Value- is the most important aspect of big data; it refers to the process of discovering huge hidden values from large datasets with various types and rapid generation

B. Classification of Big Data

Big data are classified into different categories to better understand their characteristics shows the numerous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging, and (v) data processing.

Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. Most popular is the relational database that come in a large number of varieties. As the result of the wide variety of data sources, the captured data differ in size with respect to redundancy, consistency and noise, etc.

C. Big Data Storage System

The rapid growth of data has restricted the capability of existing storage technologies to store and manage data. Over the past few years, traditional storage systems have been utilized to store data through structured RDBMS. However, almost storage systems have limitations and are inapplicable to the storage and management of big data. A storage architecture that can be accessed in a highly efficient manner while achieving availability and reliability is required to store and manage large datasets

Several storage technologies have been developed to meet the demands of massive data. Existing technologies can be classified as direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN). In DAS, various hard disk drives (HDDs) are directly connected to the servers. Each HDD receives a certain amount of input/output (I/O) resource, which is managed by individual applications. Therefore, DAS is suitable only for servers that are interconnected on a small scale. Given the aforesaid low scalability, storage capacity is increased but expandability and upgradeability are limited significantly. NAS is a storage device that supports a network. NAS is connected directly to a network through a switch or hub via TCP/IP protocols. In NAS, data are transferred as files. Given that the NAS server can indirectly access a storage device through networks, the I/O burden on a NAS server is significantly lighter than that on a DAS server. NAS can orient networks, particularly scalable and bandwidth-intensive networks. Such networks include high-speed networks of optical-fiber connections. The SAN system of data storage is independent with respect to storage on the local area network (LAN). Multipath data switching is conducted among internal nodes to maximize data management and sharing. The organizational systems of data storages (DAS, NAS, and SAN) can be divided into three parts: (i) disc array, where the foundation of a storage system provides the fundamental guarantee, (ii) connection and network subsystems, which connect one or more disc arrays and servers, and (iii) storage management software, which oversees data sharing, storage management, and disaster recovery tasks for multiple servers.

D. Hadoop Background

Hadoop is an open-source Apache Software Foundation project written in Java that enables the distributed processing of large datasets across clusters of commodity. Hadoop has two primary components, namely, HDFS and Map Reduce programming framework. The most significant feature of Hadoop is that HDFS and Map Reduce are closely related to each other; each are co-deployed such that a single cluster is produced. Therefore, the storage system is not physically separated from the processing system.

HDFS is a distributed file system designed to run on top of the local file systems of the cluster nodes and store extremely large files suitable for streaming data access. HDFS is highly fault tolerant and can scale up from a single server to thousands of machines, each offering local computation and storage. HDFS consists of two types of nodes, namely, a name node called “master” and several data nodes called “slaves.” HDFS can also include secondary name nodes. The name node manages the hierarchy of file systems and director namespace (i.e., metadata). File systems are presented in a form of name node that registers attributes, such as access time, modification, permission, and disk space quotas. The file content is split into large blocks, and each block of the file is independently replicated across data nodes for redundancy and to periodically send a report of all existing blocks to the name node.

Map Reduce is a simplified programming model for processing large numbers of datasets pioneered by Google for data-intensive applications. The Map Reduce model was developed based on GFS is adopted through open-source Hadoop implementation, which was popularized by Yahoo. Apart from the Map Reduce framework, several other current open-source Apache projects are related to the Hadoop ecosystem, including Hive, Hbase, Mahout, Pig, Zookeeper, Spark, and Avro. Twister provides support for efficient and iterative MapReduce computations. An overview of current Map Reduce projects and related software is shown in MapReduce allows an inexperienced programmer to develop parallel programs and create a program capable of using computers in a cloud. In most cases, programmers are required to specify two functions only: the map function (mapper) and the reduce function (reducer) commonly utilized in functional programming. The mapper regards the key/value pair as input and generates intermediate key/value pairs. The reducer merges all the pairs associated with the same (intermediate) key and then generates an output. Summarizes the process of the map/reduce function.

E. Map Reduce in Clouds

Map Reduce accelerates the processing of large amounts of data in a cloud; thus, Map Reduce, is the preferred computation model of cloud providers. Map Reduce is a popular cloud computing framework that robotically performs scalable distributed applications and provides an interface that allows for parallelization and distributed computing in a cluster of servers. The approachis to apply scientific computing problems to the MapReduce framework where scientists can efficiently utilize existing resources in the cloud to solve computationally large-scale scientific data.

Currently, many alternative solutions are available to deploy MapReduce in cloud environments; these solutions include using cloud MapReduce runtimes that maximize cloud infrastructure services, using MapReduce as a service, or setting up one′s own MapReduce cluster in cloud instances. Several strategies have been proposed to improve the performance of big data processing. Moreover, effort has been exerted to develop SQL interfaces in the MapReduce framework to assist programmers who prefer to use SQL as a high-level language to express their task while leaving all of the execution optimization details to the backend

F. Research Challenges

Although cloud computing has been broadly accepted by many organizations, research on big data in the cloud remains in its early stages. Several existing issues have not been fully addressed. Moreover, new challenges continue to emerge from applications by organization. In the subsequent sections, some of the key research challenges, such as scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy and legal issues, and regulatory governance, are discussed.

G. Scalability

Scalability is the ability of the storage to handle increasing amounts of data in an appropriate manner. Scalable distributed data storage systems have been a critical part of cloud computing infrastructures The lack of cloud computing features to support RDBMSs associated with enterprise solutions has made RDBMSs less attractive for the deployment of large-scale applications in the cloud. This drawback has resulted in the popularity of NoSQL

A NoSQL database provides the mechanism to store and retrieve large volumes of distributed data. The features of NoSQL databases include schema-free, easy replication support, simple API, and consistent and flexible modes. Different types of NoSQL databases, such as key-value column-oriented, and document-oriented, provide support for big data. Shows a comparison of various NoSQL database technologies that provide support for large datasets.

H. Availability

Availability refers to the resources of the system accessible on demand by an authorized individual. In a cloud environment, one of the main issues concerning cloud service providers is the availability of the data stored in the cloud. For example, one of the pressing demands on cloud service providers is to effectively serve the needs of the mobile user who requires single or multiple data within a short amount of time. Therefore, services must remain operational even in the case of a security breach. In addition, with the increasing number of cloud users, cloud service providers must address the issue of making the requested data available to users to deliver high-quality services. Lee et al. introduced a multi-cloud model called “rain clouds” to support big data exploitation. “Rain clouds” involves cooperation among single clouds to provide accessible resources in an emergency. Schroeck et al. predicted that the demand for more real time access to data may continue to increase as business models evolve and organizations invest in technologies required for streaming data and smart phones.

I. Data Integrity

A key aspect of big data security is integrity. Integrity means that data can be modified only by authorized parties or the data owner to prevent misuse. The proliferation of cloud-based applications provides users the opportunity to store and manage their data in cloud data centers. Such applications must ensure data integrity. However, one of the main challenges that must be addressed is to ensure the correctness of user data in the cloud.

J. Transformation

Transforming data into a form suitable for analysis is an obstacle in the adoption of big data. Owing to the variety of data formats, big data can be transformed into an analysis workflow in two ways

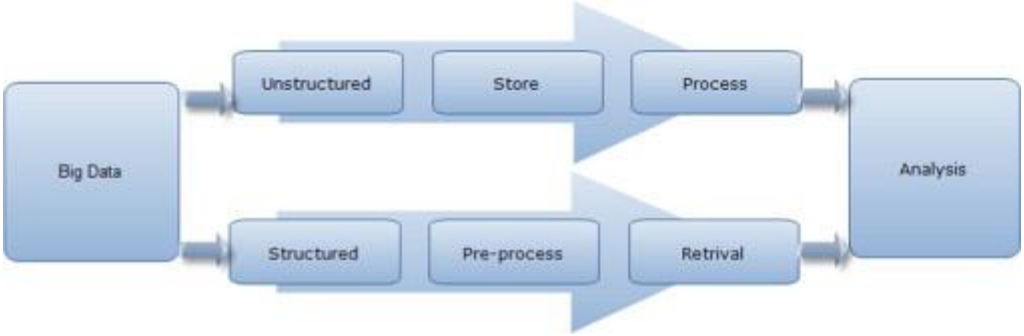

Fig 1.2 Transformation process

K. Data Quality

In the past, data processing was typically performed on clean datasets from well-known and limited sources. Therefore, the results were accurate. However, with the emergence of big data, data originate from many different sources; not all of these sources are well-known or verifiable. Poor data quality has become a serious problem for many cloud service providers because data are often collected from different sources. For example, huge amounts of data are generated from smart phones, where inconsistent data formats can be produced as a result of heterogeneous sources. The data quality problem is usually defined as “any difficulty encountered along one or more quality dimensions that render data completely or largely unfit for use”. Therefore, obtaining high-quality data from vast collections of data sources is a challenge. High-quality data in the cloud is characterized by data consistency. If data from new sources are consistent with data from other sources, then the new data are of high quality.

L. Heterogeneity

Variety, one of the major aspects of big data characterization, is the result of the growth of virtually unlimited different sources of data. This growth leads to the heterogeneous nature of big data. Data from multiple sources are generally of different types and representation forms and significantly interconnected; they have incompatible formats and are inconsistently represented.

In a cloud environment, users can store data in structured, semi-structured, or unstructured format. Structured data formats are appropriate for today's database systems, whereas semi-structured data formats are appropriate only to some extent. Unstructured data are inappropriate because they have a complex format that is difficult to represent in rows and columns. According to Kocarev and Jakimoski, the challenge is how to handle multiple data sources and types.

M. Privacy

Privacy concerns continue to hamper users who outsource their private data into the cloud storage. This concern has become serious with the development of big data mining and analytics, which require personal information to produce relevant results, such as personalized and location-based services. Information on individuals is exposed to scrutiny, a condition that gives rise to concerns on profiling, stealing, and loss of control.

N. Legal/Regulatory Issues

Specific laws and regulations must be established to preserve the personal and sensitive information of users. Different countries have different laws and regulations to achieve data privacy and protection. In several countries, monitoring of company staff communications is not allowed. However, electronic monitoring is permitted under special circumstances. Therefore, the question is whether such laws and regulations offer adequate protection for individuals 'data while enjoying the many benefits of big data in the society at large.

O. Governance

Data governance embodies the exercise of control and authority over data-related rules of law, transparency, and accountabilities of individuals and information systems to achieve business objectives. The key issues of big data in cloud governance pertain to applications that consume massive amounts of data streamed from external sources. Therefore, a clear and acceptable data policy with regard to the type of data that need to be stored, how quickly an individual needs to access the data, and how to access the data must be defined.

Big data governance involves leveraging information by aligning the objectives of multiple functions, such as telecommunication carriers having access to vast troves of customer information in the form of call detail records and marketing seeking to monetize this information by selling it to third parties.

Moreover, big data provides significant opportunities to service providers by making information more valuable. However, policies, principles, and frameworks that strike stability between risk and value in the face of increasing data size and deliver better and faster data management technology can create huge challenges.

Cloud governance recommends the use of various policies together with different models of constraints that limit access to underlying resources. Therefore, adopting governance practices that maintain a balance between risk exposure and value creation is a new organizational imperative to unlock competitive advantages and maximize value from the application of big data in the cloud.

P. Open Research Issues

Numerous studies have addressed a number of significant problems and issues pertaining to the storage and processing of big data in clouds. The amount of data continues to increase at an exponential rate, but the improvement in the processing mechanisms is relatively slow. Only a few tools are available to address the issues of big data processing in cloud environments. State-of-the-art techniques and technologies in many important big data applications (i.e., MapReduce, Dryad, Pregel, PigLatin, MangoDB, Hbase, SimpleDB, and

Cassandra) cannot solve the actual problems of storing and querying big data. For example, Hadoop and Map Reduce lack query processing strategies and have low-level infrastructures with respect to data processing and management. Despite the plethora of work performed to address the problem of storing and processing big data in cloud computing environments, certain important aspects of storing and processing big data in cloud computing are yet to be solved. Some of these issues are discussed in the subsequent subsections.

Q. Data Staging

The most important open research issue regarding data staging is related to the heterogeneous nature of data. Data gathered from different sources do not have a structured format. For instance, mobile cloud-based applications, blogs, and social networking are inadequately structured similar to pieces of text messages, videos, and images. Transforming and cleaning such unstructured data before loading those into the warehouse for analysis are challenging tasks. Efforts have been exerted to simplify the transformation process by adopting technologies such as Hadoop and MapReduce to support the distributed processing of unstructured data formats. However, understanding the context of unstructured data is necessary, particularly when meaningful information is required. MapReduce programming model is the most common model that operates in clusters of computers; it has been utilized to process and distribute large amounts of data.

R. Distributed Storage Systems

Numerous solutions have been proposed to store and retrieve massive amounts of data. Some of these solutions have been applied in a cloud computing environment. However, several issues hinder the successful implementation of such solutions, including the capability of current cloud technologies to provide necessary capacity and high performance to address massive amounts of data, optimization of existing file systems for the volumes demanded by data mining applications, and how data can be stored in such a manner that they can be easily retrieved and migrated between servers.

S. Data Analysis

The selection of an appropriate model for large-scale data analysis is critical. Talia pointed out that obtaining useful information from large amounts of data requires scalable analysis algorithms to produce timely results. However, current algorithms are inefficient in terms of big data analysis. Therefore, efficient data analysis tools and technologies are required to process such data. Each algorithm performance ceases to increase linearly with increasing computational resources. As researchers continue to probe the issues of big data in cloud computing, new problems in big data processing arise from the transitional data analysis techniques. The speed of stream data arriving from different data sources must be processed and compared with historical information within a certain period of time. Such data sources may contain different formats, which makes the integration of multiple sources for analysis a complex task.

T. Data Security

Although cloud computing has transformed modern ICT technology, several unresolved security threats exist in cloud computing. These security threats are magnified by the volume, velocity, and variety of big data. Moreover, several threats and issues, such as privacy, confidentiality, integrity, and availability of data, exist in big data using cloud computing platforms. Therefore, data security must be measured once data are outsourced to cloud service providers. The cloud must also be assessed at regular intervals to protect it against threats. Cloud vendors must ensure that all service level agreements are met. Recently, some controversies have revealed how some security agencies use data generated by individuals for their own benefit without permission.

Therefore, policies that cover all user privacy concerns should be developed. Traditionally, the most common technique for privacy and data control is to protect the systems utilized to manage data rather than the data itself; however, such systems have proven to be vulnerable. Utilizing strong cryptography to encapsulate sensitive data in a cloud computing environment and developing a novel algorithm that efficiently allows for key management and secure key exchange are important to manage access to big data, particularly as they exist in the cloud independent of any platform. Moreover, the issue with integrity is that previously developed hashing schemes are no longer applicable to large amounts of data. Integrity verification is also difficult because of the lack of support, given remote data access and the lack of information on internal storage.

CHAPTER 2

LITERATURE SURVEY

Paper 1: Developing an Index for Detection and Identification of Disease Stages (2016)

Authors: Davoud Ashourloo, Ali Akbar Matkan, Alfredo Huete, Hossein Aghighi, and Mohammad Reza Mobasheri

Abstract: Spectral data have been widely used to estimate the disease severity (DS) levels of different plants. However, such data have not been evaluated to estimate the disease stages of the plant. This study aimed at developing a spectral disease index (SDI) that is able to identify the stages of wheat leaf rust disease at various DS levels. To meet the aim of the study, the reflectance spectra (350–2500 nm) of infected leaves with different symptom fractions and DS levels were measured with a spectroradiometer. Then, pure spectra of the different disease symptoms at the leaf scale were analyzed, and a new function was developed to find the wavelengths most sensitive to disease symptom fraction. The reflectance spectra with highest sensitivity were found at 675 and 775 nm. Finally, the normalized difference of DS and the ratio ρ675/ρ775 was used as a new SDI to discriminate three different levels of the disease stage at the canopy level. The suggested SDI showed a promising performance to improve the detection disease stages in precision plant protection

Inference: The experiment was performed on Plants for it Disease Prediction. It is implemented to identify three different stages of WLR disease. In stage 1, the leaf begins to show symptoms with yellow and orange colors predominant. In stage 2, all of the symptoms are fully developed and can be seen with similar proportion of colors. Stage 3 includes dry leaves in which the predominant symptoms show as brown color

Disadvantages: It is not for Human Disease Diagnosis, It is not implemented on Genome based diagnosis. It is not a Prediction / Early detection of Disease Detection process. Big data is not Used.

Paper 2: IDENTIFYING THE CANDIDATE GENES FOR ALZHEIMER'S DISEASE BASED ON THE REJECTION REGION OF T TEST (2017)

Authors: GUI-QIONG ZHU1,PEI-HUI YANG2

Abstract: Modern medical research has proved that almost all diseases are related to genes except the injuries. Based on the Alzheimer's disease (AD) gene expression data, this paper uses

T test method to identify the significant differences in gene expression between normal people and three states of Alzheimer's disease patients. In this paper, 90 differentially expressed genes were identified, and 5 of which have been confirmed to be associated with Alzheimer's disease by other references, 10 of which are associated with nerve cell tissue and signaling. This result indicates that the identified genes (at least some of them) are likely to be related to Alzheimer's

Disease

Inference: Gene Sequence Analysis towards Disease Diagnosis (Alzheimer's disease) using K means & PCA Algorithms.

Disadvantages: Just a diagnosis Tool and not working towards Prevention mechanism to others. No Realtime Medical Data is used. Big data is not Used.

Paper 3: A new method for disease-related gene prioritization (2017)
Author: Yaogong Zhang

Prioritizing genes according to their association with a disease allows researchers to explore genes in more informed ways. Although some useful algorithms have been developed, they are based on single gene importance, gene interaction networks, or gene modules with little consideration of relative gene importance in the context of modules. In this paper, we propose to prioritize genes considering both individual genes and their affiliated modules, and utilize Gene Ontology (GO) based fuzzy measure value as well as known disease genes as heuristics. The performance of our method is comprehensively validated by using both simulated and real datasets. Results show that our method outperforms other methods in terms of disease-related gene prioritization. This work will aid researchers in the understanding of the genetic architecture of complex diseases, and improve the accuracy of diagnosis and the effectiveness of therapy.

Inference: This Paper, prioritize genes considering both individual genes and their affiliated modules, and utilize Gene Ontology (GO) based fuzzy measure value as well as known disease genes as heuristics.

Disadvantages: Polymorphism is not considered in analysis, Bigdata is not used. Both Realtime & Simulated data set is used for their advantage, so Result may not be considered Genuine also.

Paper 4: CRLEDD: Regularized Causalities Learning for Early Detection of Diseases Using Electronic Health Record (EHR) Data (2021)

Authors: Jiang Bian , Sijia Yang, Haoyi Xiong, Licheng Wang , Yanjie Fu, Zeyi Sun , Zhishan Guo , and Jun Wang.

The availability of Electronic Health Records (EHR) in health care settings has provided tremendous opportunities for early disease detection. While many supervised learning models have been adopted for EHR-based disease early detection, the ill-posed inverse problem in the parameter learning has imposed a significant challenge on improving the accuracy of these algorithms. In this paper, we propose CRLEDD – Causality-Regularized Learning for Early Detection of Disease, an algorithm to improve the performance of Linear Discriminant Analysis (LDA) on top of diagnosis-frequency vector data representation. While most existing regularization methods exploit sparsity regularization to improve detection performance, CRLEDD provides a unique perspective by ensuring positive semi-definiteness of the sparsified precision matrix used in LDA which is different from the regular regularization method (e.g.,L2regularization).To achieve this goal, CRLEDD employs Graphical Lasso to estimate the precision matrix in the ill-posed settings for enhanced accuracy of LDA classifiers. We perform extensive evaluation of CRLEDD using a large-scale real-world EHR dataset to predict mental health disorders (e.g., depression and anxiety) of college students from 10 universities in the U.S. We compare CRLEDD with other regularized LDA and downstream classifiers**.**

Inference: Given EHR data represented as vectors and graphs, researchers have proposed to predict the target disease through supervised learning, using downstream classifiers.

Disadvantages: It is not applied on genome Sequence. It is applied only on PHR Medical data.

Paper 5: Diagnosis of Kashin-Beck disease and other common joint diseases via a gene model (2021)
Author: Yanan Zhang

Kashin-Beck disease and common joint disease display similar early clinical features, result in the difficulty of identification of diagnosis and treatment by observing the clinical manifestations and physical signs. In this paper, we use two sets of gene expression data from cartilaginous tissues and a health tissues, collected from 16 samples. We propose a gene model to achieve the diagnosis of Kashin-Beck disease and other common joint diseases using the expression. In the model, we try to filter various noises caused by improper operation, aging of equipment and so on by wavelet transform. Bhattacharyya distance and Pearson correlation coefficient are used to remove irrelevant genes and redundant genes respectively and acquire the rest as feature genes. In order to deal with the comprehensive data, we choose support vector machine (SVM) to achieve identification. The experiment results demonstrate that the gene model proposed in this paper offers a set of feature genes for further biological interpretation and play an important role in the diagnosis.

Inference: Gene model to achieve the diagnosis of Kashin-Beck disease and other common joint diseases using the expression

Disadvantages: Big Data is not used in this Project. Not spoken about Large Data set. This project is not aimed at Early Discovery of Disease and not working towards Prevention process.

CHAPTER 3
REQUIREMENT ANALYSIS

3.1 FEASIBILITY STUDIES/RISK ANALYSIS OF THE PROJECT

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- OPERATIONAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

OPERATIONAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Hardware specification:

Operating system	:	Windows 7/ 8.1
HDD	:	500GB
RAM	:	2-4GB
Processor	:	Core i3, i5

Software specification:

Front End	:	JDK 7/8
Back End	:	MySQL
IDE	:	NetBeans 8.1 / 8.2
Tool	:	Hadoop

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

The implementation of the project is to extract the Patient's gene sequence coding (Exons) from the given dataset and compare with the standardized Amino-acid sequence. The gene sequence is principally classified into two categorizes namely Introns & Exons. Introns are the Non-Coding sequence which is followed by Exons. Gene Sequence contains repeated of 4 Nucleotides namely Adenine (A), Thymine (T), Guanine (G), Cytosine (C). The proposed system is to recommend Diet pattern or any other natural drugs which can be recommended to those people who is expected to get into the disease in a course of time by verifying the mutated Genes.

This paper infer that identifying disease based on gene using SVM algorithm. Through this system apart from disease identification we also suggest the drug based on the disease.

4.1 SELECTED METHODOLOGY OR PROCESS MODEL

MODULES:

- A modular design reduces complexity, facilities change (a critical aspect of software maintainability), and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity that is software is divided into separately named and addressable components called modules that are integrated to satisfy problem requirements.

- Modularity is the single attribute of software that allows a program to be intellectually manageable. The five important criteria that enable us to evaluate a design method with respect to its ability to define an effective modular design are: Modular decomposability, Modular Comps ability, Modular Understandability, Modular continuity, Modular Protection.
- The following are the modules of the project, which is planned in aid to complete the project with respect to the proposed system, while overcoming existing system and also providing the support for the future enhancement.

1. USER INTERFACE DESIGN
2. DATA SET MAINTENANCE
3. OUTLAYER REMOVAL
4. DISEASE PREDICTION
5. DRUG SUGGESTION

1.USER INTERFACE DESIGN:

User interface design, in this module User will be updating the Genome dataset of the Patient to the HDFS Server for Processing. To develop our application we use netbeans as IDE and MSQL as a back end. All inputs and output will put and get through this IDE only.

2.DATA SET MAINTENENCE:

The Server will monitor the entire Database. The database contains the Patient’s Genome dataset along with Disease dataset for Comparison. Also the Server will store the entire information in their database. Also the Server has to establish the connection to communicate with the Users. Admin will add all the different types of gene datasets like parent gene, normal gene, mutated gene.

3.OUTLAYER REMOVAL:

A gene mutation is a permanent alteration in the DNA sequence that makes up a gene, such that the sequence differs from what is found in most people. Mutations range in size; they can affect anywhere from a single DNA building block (base pair) to a large segment of a chromosome that includes multiple genes. The mutated gene is considered as equal to normal gene and it will be removed from the datasets.

4.DISEASE PREDICTOIN:

In this module we identify the disease by implementing SVM algorithm. we will identify the disease type that means, the gene which will given as a input, that will send to the trained dataset for comparison and identify the disease type whether it is gene based disease or infection based disease. if carrier-based disease found system will predict disease name.

5.DRUG SUGGESTION:

In this module after disease detection system will notify the possibilities of disease. And system will suggest the drug to that infected person. we suggest both English and herbal medicine as per user's requirement.

4.2 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

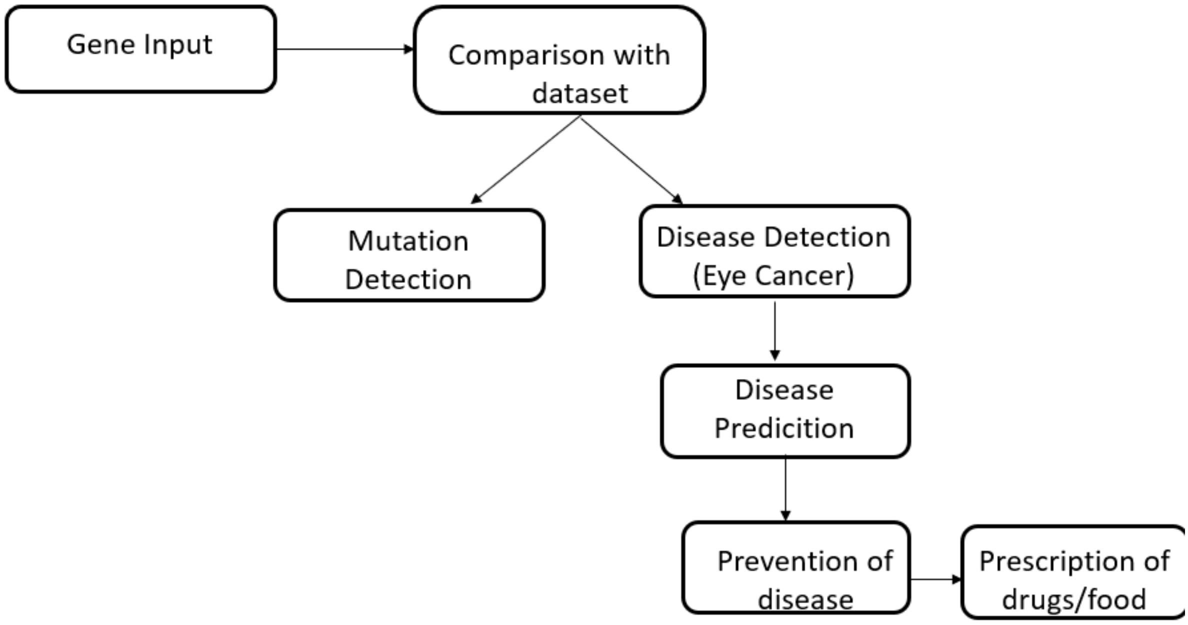

Fig 4.1: System Architecture

The overall architecture describes about user will upload gene sequence, after that those sequence will analyze using SVM algorithm in hadoop analyzer. Hadoop analyzer will translate the gene into short form of every gene sequence. After the translation it will form like a huge amount short form format which will predict the disease type and drugs and food suggestion to user.

.

Chapter 5
CONCLUSION

Here we can identify disease based on gene using SVM algorithm. We can also find out whether it is polymorphism. Through this system apart from disease identification we also suggest the drug based on the disease.

REFERENCES:-

[1] K. M. Boycott, M. R. Vanstone, D. E. Bulman and A. E. MacEnzie, “Rare-disease genetics in the era of next-generation sequencing: discovery to translation”, in Nature Reviews Genetics, vol. 14(10), pp. 681–691, 2013.

[2] C. M. Condit, P. J. Achter, I. Lauer and E. Sefcovic, “The changing meanings of “mutation:” A contextualized study of public discourse”, in Human Mutation, vol. 19(1), pp. 69–75, 2002.

[3] W. Raghupathi and Viju Raghupathi. “Big Data Analytics in Healthcare: Promise and Potential.”, in Health Information Science and Systems, 2:3, 2014. doi:10.1186/2047-2501-2-3.

[4] D. Howe et al., “Big data: The future of biocuration”, Nature, vol. 455, pp. 47-50, 2008.

[5] T. B. Murdock et al., “The Inevitable Application of Big Data to Health Care”, JAMA, vol. 309(13), pp. 1351-1352, 2013.

[6] F. Celesti et al., “Big data analytics in genomics: The point on Deep Learning solutions”, in 2017 IEEE Symposium on Computers and Communications (ISCC), pp. 306–309, 2017.

[7] D. Laney, “3D data management: Controlling data volume, velocity and variety”, in META Group Research Note, February 2001).

[8] Y. Demchenko, Cee de Laat and P. Membrey, “Defining architecture components of the Big Data Ecosystem”, in 2014 International Conference on Collaboration Technologies and Systems (CTS), pp. 104–112, 2014

[9] A. Splendiani, M. Donato and S. Drăghici, “Ontologies for Bioinformatics”, in Springer Handbook of Bio-/Neuroinformatics, pp. 441–461, 2014.

[10] N. W. Paton et al.,“Conceptual modelling of genomic information”, in Bioinformatics,vol.16(6),pp.548–57,2000.