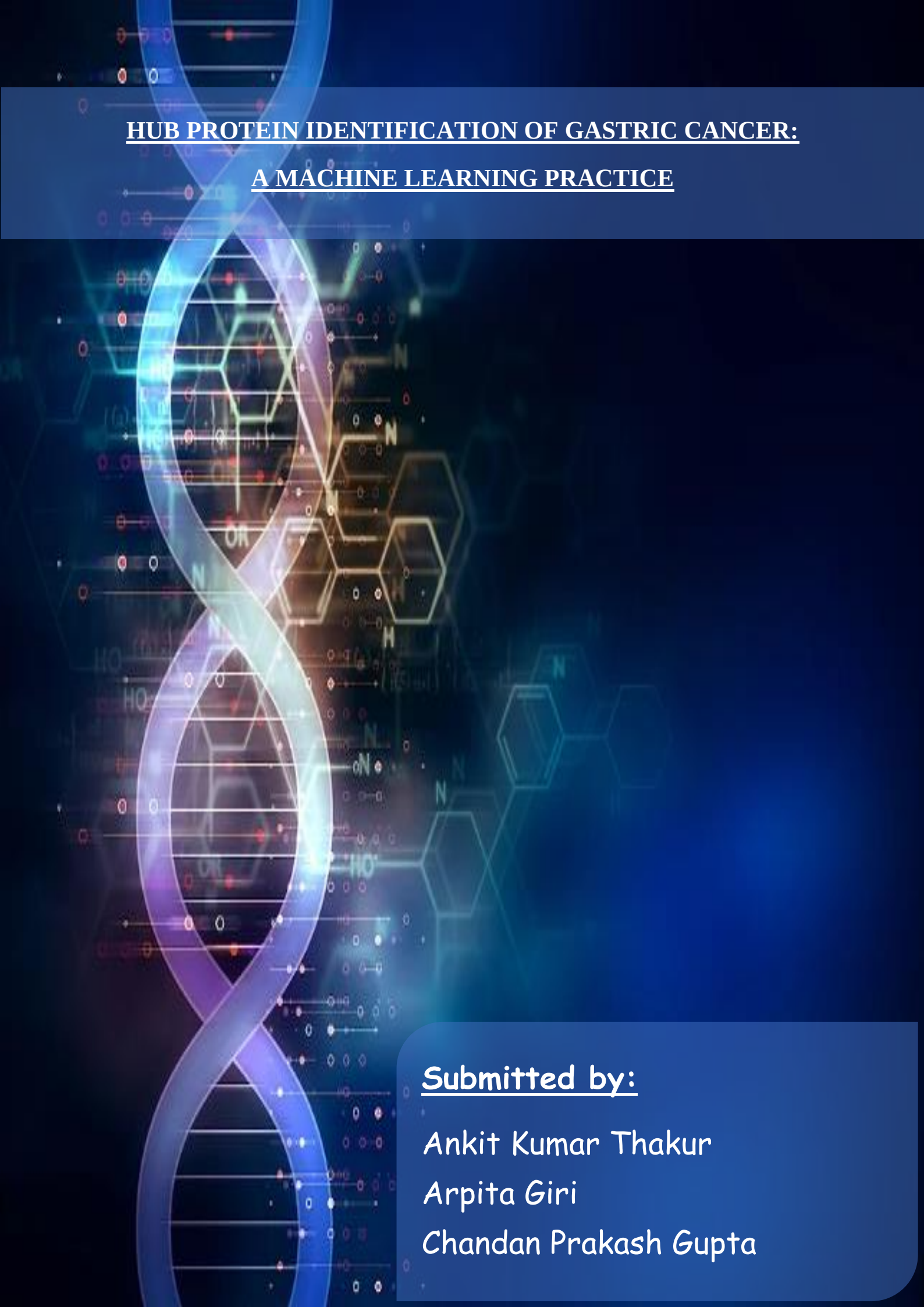
HUB PROTEIN IDENTIFICATION OF GASTRIC CANCER: A MACHINE LEARNING PRACTICE



Submitted by:

Ankit Kumar Thakur

Arpita Giri

Chandan Prakash Gupta

HUB PROTEIN IDENTIFICATION OF GASTRIC CANCER: A MACHINE LEARNING PRACTICE

Submitted by

Ankit Kumar Thakur
Arpita Giri
Chandan Prakash Gupta

Under Supervision of

Dr. Nabin Ghoshal

DEPARTMENT OF ENGINEERING AND TECHNOLOGICAL STUDIES

UNIVERSITY OF KALYANI

CERTIFICATE

DEPARTMENT OF ENGINEERING AND TECHNOLOGICAL STUDIES
UNIVERSITY OF KALYANI
KALYANI, NADIA
WEST BENGAL - 741235

This is to certify that the project report entitled as " **Hub Protein identification of Gastric Cancer: A Machine Learning Practice** " has jointly completed by **Ankit Kumar Thakur (Registration No: 100214 of 2020-2021)**, **Arpita Giri (Registration No: 100217 of 2020-2021)** and **Chandan Prakash Gupta (Registration No: 100213 of 2020-2021)** in partial fulfilment of the requirement for the degree of Bachelor of Technology (B.Tech) in Information Technology at the Department of Engineering and Technological Studies, University of Kalyani, under the supervision of undersigned.

_____

Dr. Nabin Ghoshal
Project Guide
Senior Scientific Officer
DETS, University of Kalyani

Examiner 1:_____ Examiner 2:_____

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all those who have contributed to the successful completion of the project "**Hub Protein identification of GASTRIC CANCER: A Machine Learning Practice**". We extend our heartfelt thanks to University of Kalyani for providing us with the necessary resources and facilities to carry out this project.

We would like to express our gratitude to **Dr. Nabin Ghoshal** and **Mr. Arup Mallick**, our project guide, for providing us with invaluable guidance, support, and mentorship throughout the project. He has been instrumental in shaping our understanding of the project, and has been a constant source of motivation and inspiration.

We also extend our gratitude to our colleagues and peers, whose insights and feedback have been critical in shaping the project. Their encouragement and support have been greatly appreciated.

Place: Kalyani Ankit Kumar Thakur_____

Date: Arpita Giri_____

Chandan Prakash Gupta_____

INDEX

Content ***Page No.***

1. Abstract:

Gastric cancer is a major cause of cancer-related deaths worldwide, and early diagnosis and prognosis are crucial for effective treatment. The disease can be caused by a variety of factors, including genetics, environmental factors, and lifestyle habits such as smoking and alcohol consumption. In recent years, machine learning algorithms have emerged as powerful tools for gene complex medical data, including genomic and imaging data, and improving cancer diagnosis and treatment. In this study, we aimed to identify a Transcription factor (TF) for gastric cancer using a combination of transcriptomic analysis and clinical validation. Here are Some specific reasons why transcription factors are used in Gastric cancer.

Regulation of cell proliferation and differentiation:-Transcription factors can regulate the expression of genes involved in cell proliferation and differentiation which are key processes in the development and progression of gastric cancer.

Involvement in gastric cancer metastasis:- Transcription factors can also play a role in promoting the metastasis of gastric cancer cells to other parts of the body. Example: EMT in gastric cancer cell which can increase their invasive and migratory capabilities.

Diagnostic and prognostic biomarkers: Aberrant expression of transcription factors can be used as diagnostic and prognostic biomarkers for gastric cancer.

We used a combination of supervised and unsupervised machine learning algorithms to analyse gene expression data from gene samples and clinical data from patients with gastric cancer. We identified a set of gene name that were significantly associated with patient survival, and used these genes to build a prognostic model with high accuracy in predicting patient outcomes.

The first step We analysed the gene expression profiles from a publicly available gastric cancer dataset datasets .Then we involves constructing a protein-protein interaction network using available data on protein interactions relevant to Gastric-Cancer. This network represents the complex relationships and interactions between different proteins involved in the disease. Next, graph analysis algorithms are applied to identify the central or highly connected proteins within the network. These proteins, known as hub proteins, are likely to have a significant influence on the disease Gastric-Cancer progression and may serve as potential therapeutic targets. The identified hub proteins can provide valuable insights into the underlying mechanisms of and may offer potential targets for drug discovery and treatment strategies. Next, we identify the transcription factor to the development of new diagnostic and therapeutic strategies for patients.
In summary, hub proteins such as MUC1, BARD1, BARCA1, RAPGEF6, RAPGEF2, MLLT4, HRAS, ATM, EGFR,BRCA2, TP53, ALB,DCN, POLR3E, POLR3D have been extensively studied in relation to gastric cancer. This molecules provide important insight into the mechanism underlying the development and progression of gastric cancer , and may serve as potential targets for diagnosis, and the therapy.

2. Introduction:

2.1 Background:

Gastric cancer, also known as stomach cancer, is a type of cancer that develops in the lining of the stomach. It is the fifth most common cancer worldwide and the third leading cause of cancer-related deaths. The incidence and mortality rates of gastric cancer vary widely across regions and countries, with the highest rates found in Eastern Asia, particularly in Japan, Korea, and China.

There are several factors that can increase the risk of developing gastric cancer, including:
1. Helicobacter pylori infection: H. pylori is a bacterium that can cause inflammation and damage to the stomach lining, which can lead to an increased risk of developing gastric cancer.
2. Diet: A diet high in salt, processed meat, and low in fruits and vegetables may increase the risk of gastric cancer.
3. Age: Gastric cancer is more common in older adults.
4. Family history: Individuals with a family history of gastric cancer are at a higher risk of developing the disease.
5. Smoking: Tobacco use has been linked to an increased risk of gastric cancer.

Symptoms of gastric cancer may include indigestion, heartburn, nausea, vomiting, abdominal pain, unintentional weight loss, and fatigue. However, these symptoms can also be caused by other conditions, so it is important to see a healthcare provider for an accurate diagnosis.

Treatment options for gastric cancer depend on the stage and location of the cancer, as well as the overall health of the patient. Surgery, radiation therapy, chemotherapy, and targeted therapy are some of the common treatments used to manage gastric cancer. Early detection and prompt treatment can improve the chances of a successful outcome
In recent years, machine learning techniques have revolutionized many areas of research, including bioinformatics and genomics. Machine learning models can leverage large-scale datasets to discover complex patterns and relationships that are difficult to capture using traditional analytical approaches. By integrating machine learning with graph analysis, we can harness the power of both methodologies to identify hub proteins and Transcription factor (TF) with higher accuracy and precision.
The combination of graph analysis and machine learning presents a promising approach to unravel the molecular intricacies of Gastric cancer. By constructing a protein-protein interaction network specific to Gastric cancer and applying graph analysis algorithms, we can identify hub proteins that are central to the disease's biological processes. Furthermore, machine learning models can be trained using diverse data sources, including genomic, proteomic, and clinical data, to predict and prioritize hub proteins more effectively.

The identification of hub proteins in Gastric cancer can provide valuable insights into the disease's pathogenesis, transmission, and potential therapeutic targets. It can help us understand how the virus interacts with the host cellular machinery, how it evades immune responses, and how it causes severe symptoms in certain individuals. Ultimately, this knowledge can inform the development of targeted interventions, such as drug therapies and chemotherapy.
In summary, the integration of graph analysis and machine learning offers a powerful approach to identify hub proteins and Transcription factor responsible for Gastric cancer .

2.2 History:

In the 19th and early 20th centuries, gastric cancer was one of the leading causes of cancer death in the Western world. The incidence of gastric cancer has since decreased in many countries, due to improvements in sanitation, food preservation, and the widespread use of antibiotics to treat Helicobacter pylori infection, which is a major risk factor for the diseases. The molecular biology of gastric cancer have led to the development of new targeted therapies, which have improved the outlook for some patients with advanced disease. Screening programs, such as endoscopy and imaging tests, are also being used to detect gastric cancer at an earlier stage, when it is more treatable.

As researchers began to explore the intricate interactions between viral proteins, host proteins, and the immune system, they realized the importance of identifying key proteins that play pivotal roles in the disease processes. These proteins, known as hub proteins, have a higher degree of connectivity within the protein-protein interaction network and are likely to be critical for viral replication, immune response modulation, and other crucial pathways.
Given the complexity of the biological systems involved in Gastric cancer, it became evident that a multidisciplinary approach combining graph analysis and machine learning techniques would be necessary to identify and characterize these hub proteins effectively. Graph analysis provides a framework for representing and analyzing protein-protein interaction networks, while machine learning algorithms offer the ability to uncover hidden patterns and relationships within large-scale datasets.

The project's history is intertwined with advancements in bioinformatics, network biology, and machine learning methodologies. Over the years, researchers have developed various graph analysis algorithms that can identify hub proteins within complex networks. Then we find the Biomarkers which are measurable indicators of disease. Additionally, advancements in machine learning, such as deep learning models, have enabled the integration of diverse data sources, including genomic, proteomic, and clinical data, to improve the accuracy and predictive power of the models. The project gained momentum as research institutions and organizations worldwide collaborated to gather and share data related to Gastric cancer, including protein-protein interactions, gene expression profiles, clinical data, and more. These rich datasets became the foundation for training machine learning models and constructing comprehensive protein-protein interaction networks specific to Gastric cancer.

As the project unfolded, researchers iteratively refined the methodologies, fine-tuned the machine learning models, and validated the results using experimental data and clinical observations. The iterative nature of the project allowed for continuous improvement and adaptation to new findings and insights emerging from the global research community.

Overall, the history of gastric cancer has been marked by significant advances in understanding and treating the disease, but much work remains to be done to prevent and cure this deadly form of cancer.

2.3 Overview:

Gastric cancer prediction and the identification of hub proteins and biomarkers involve a multi-step process that integrates various data analysis techniques and biological knowledge. Here's an overview of the general approach:

1. Data Collection: Relevant data is collected, typically including gene expression data, protein-protein interaction (PPI) data, clinical information, and other molecular data related to gastric cancer. This data may come from public databases or be generated through experiments.

2. Preprocessing and Integration: The collected data is preprocessed to ensure quality and consistency. This step involves data cleaning, normalization, and filtering. Different types of data, such as gene expression and PPI data, may need to be integrated into a unified dataset for further analysis.

3. Identification of Differentially Expressed Genes: Differential expression analysis is performed to identify genes that show significant differences in expression between gastric cancer samples and normal tissues.

4. Protein-Protein Interaction (PPI) Network Analysis: The differentially expressed genes are mapped onto PPI networks to identify hub proteins. Hub proteins are highly connected nodes in the network and are considered essential in regulating cellular processes. Network analysis techniques, such as network centrality measures (e.g., degree centrality, between ness centrality), are applied to identify these hub proteins.

5. Feature Selection: From the identified hub proteins and differentially expressed genes, feature selection techniques are applied to select a subset of the most relevant features for gastric cancer prediction. Methods like Recursive Feature Elimination (RFE) or LASSO (Least Absolute Shrinkage and Selection Operator) can be used to identify the most informative features.

6. Model Development and Validation: Predictive models, such as machine learning algorithms or statistical models, are built using the selected features to predict gastric cancer. These models are trained using labeled data (e.g., known cancer samples) and validated using appropriate

techniques like cross-validation or independent testing datasets. The performance of the models is evaluated using metrics like accuracy, sensitivity, specificity, or area under the curve (AUC).

7. Biological Interpretation: The identified hub proteins and biomarkers are further analyzed to understand their biological relevance and potential mechanisms underlying gastric cancer. Functional enrichment analysis, pathway analysis, and literature mining can provide insights into the biological processes and pathways associated with these proteins and biomarkers.

It's important to note that the specific methods and steps may vary depending on the available data, research goals, and the advancements in the field. Collaboration between computational biologists, clinicians, and experimental researchers is often crucial for the successful identification of hub proteins and biomarkers for gastric cancer prediction.

2.4 Objective:

The objective of a gastric cancer prediction project that aims to identify hub proteins and biomarkers is to improve early detection and diagnosis of gastric cancer. Gastric cancer is one of the leading causes of cancer-related deaths worldwide, and early detection is crucial for effective treatment and improved patient outcomes. By identifying hub proteins and biomarkers associated with gastric cancer, researchers can potentially develop more accurate and reliable diagnostic tests for the disease.
Overall, the objective of a gastric cancer prediction project that focuses on identifying hub proteins and biomarkers is to improve early detection, diagnosis, and treatment of the disease, ultimately leading to improved patient outcomes.
Specifically, the project aims to achieve the following objectives:

Apply graph analysis techniques to identify hub proteins: Utilize graph analysis algorithms to analyze the connectivity patterns within the protein-protein interaction network. Identify and prioritize the hub proteins, which are highly connected proteins within the network and are likely to have a crucial role in Gastric cancer
Develop machine learning models for hub protein prediction: Train machine learning models using diverse datasets, including genomic, proteomic, and clinical data, to capture the complex patterns and relationships within the protein-protein interaction network. These models aim to predict and prioritize hub proteins with higher accuracy and precision.
Characterize the identified hub proteins: Analyze the functional properties, biological pathways, and protein-protein interaction partners of the identified hub proteins. Gain insights into the specific roles these proteins play in gastric cancer pathogenesis, viral replication, immune response modulation, and other disease-related processes.
Identify the biomarkers: By identifying biomarkers that are specific to gastric cancer, researchers can potentially develop diagnostic tests that can detect the disease at an early stage or monitor the effectiveness of treatment. Biomarkers can also be used to stratify patients based on their risk of developing the disease or their response to specific treatments.

Provide insights for potential therapeutic interventions: The identified hub proteins can serve as potential targets for the development of therapeutic interventions, including drugs, vaccines, and other treatment strategies. The project aims to contribute valuable insights for the design of targeted therapies that can mitigate the impact of COVID-19 on public health.

Overall, the objective of a gastric cancer prediction project that focuses on identifying hub proteins and biomarkers is to improve early detection, diagnosis, and treatment of the disease, ultimately leading to improved patient outcomes.

3. Motivation:

There are several motivations for choosing gastric cancer prediction as an area of focus in healthcare research and development:

- Improve early detection: Gastric cancer is often diagnosed at an advanced stage, when treatment options may be limited and the chances of survival are lower. Developing models for gastric cancer prediction can help identify patients who are at high risk for developing the disease, allowing for earlier screening and diagnosis.
- Personalized medicine: Predictive models for gastric cancer can help clinicians tailor treatment options based on individual patient characteristics and risk factors. This can improve treatment outcomes and reduce the risk of adverse effects associated with ineffective or unnecessary treatments
- Research and innovation: The development of predictive models for gastric cancer involves the use of advanced technologies and data analysis methods, which can drive innovation in healthcare research and development.

Overall, the motivation behind choosing gastric cancer prediction is to improve early detection, personalize treatment, reduce the burden of the disease, and promote research and innovation in healthcare.

Why are we choose bio-informatics project?

Bioinformatics plays a critical role in addressing some of the most pressing biomedical challenges, such as developing personalized medicine, identifying new drug targets, and improving disease diagnosis and treatment.

Career opportunities: Bioinformatics is a highly sought-after skillset in the biotechnology, pharmaceutical, and academic sectors. Working on a bioinformatics project can provide valuable experience and training for a career in this field.

Overall, the motivation to choose a bioinformatics project can be driven by a desire to advance scientific knowledge, work with big data, collaborate across disciplines, and pursue career opportunities in this exciting rapidly and growing field.

Why are we used Machine Learning algorithm in our project?

Machine learning is a powerful tool for the prediction of gastric cancer for several reasons:

- Handling large and complex datasets: Machine learning algorithms can handle large and complex datasets with many variables, allowing for the identification of patterns and relationships that may not be apparent with traditional statistical methods.
- Identifying complex interactions: Gastric cancer is a complex disease with many potential risk factors and interactions between these factors. Machine learning algorithms can identify these complex interactions and help predict the likelihood of developing the disease based on multiple factors.
- Continuous improvement: Machine learning algorithms can be continuously refined and improved as more data becomes available, leading to more accurate predictions over time.
- Automating prediction: Machine learning algorithms can automate the prediction process, reducing the burden on healthcare providers and improving the efficiency and accuracy of prediction.

Overall, machine learning is an effective tool for the prediction of gastric cancer because it can handle large and complex datasets, identify complex interactions, provide personalized predictions, continuously improve over time, and automate the prediction process.

4. Literature Survey

There have been several studies published in the literature on the identification of key proteins, pathways and biomarker involved in the Gastric cancer using machine learning and graph analysis. Here are some of the relevant studies:

1) Study published in the journal Cancer Epidemiology in 2021 titled “A machine learning-based risk prediction model for gastric cancer:” the study highlights the potential of machine learning-based models to improve the accuracy of cancer risk prediction and aid in early detection and prevention efforts. It also demonstrates the importance of large, multiple studies in developing robust and generalizable prediction models.

2) Study published in the journal Frontiers in Oncology in 2021 titled "Identification of hub genes and pathways in gastric cancer based on integrated bioinformatics analysis." In this study, the researchers used an integrated bioinformatics approach to analyze gene expression data from gastric cancer patients and identified a set of hub genes that were highly connected in protein-protein interaction networks. They then performed functional enrichment analysis to identify the biological pathways associated with the hub genes.

3) In a study published in the journal BMC Cancer in 2021 titled "Identification of hub genes and pathways in gastric cancer based on co-expression network analysis." In this study, the researchers constructed a co-expression network using gene expression data from gastric cancer patients and identified hub genes that were highly connected within the network. They also performed functional enrichment analysis to identify the biological pathways associated with the hub genes.

4) In a study published in the journal Cancer Epidemiology, Biomarkers & Prevention in 2020 titled "A risk prediction model for gastric cancer using serum pepsinogens and Helicobacter pylori antibody: A multiple prospective cohort study." The study highlights the potential of

biomarker-based approaches to improve the accuracy of gastric cancer risk prediction, and underscores the importance of large, studies in developing robust and generalizable prediction models.

5) A study published in the journal PL0S One in 2021 titled "Identification of serum microRNA biomarkers for the early detection of gastric cancer." It highlights the potential of microRNA-based biomarkers for the early detection of gastric cancer and underscores the importance of non-invasive diagnostic methods for improving patient outcomes. It also demonstrates the power of machine learning-based approaches for the development of predictive models using multiple biomarkers.

Overall, the literature survey highlights the significant burden of gastric cancer worldwide and the importance of prevention, early detection, and effective treatment strategies for improving outcomes for patients with this disease.

5. Comparison:

5.1. Comparative study:

The research paper focuses on identifying novel gene candidates in gastric cancer using bioinformatics prediction and machine learning on gene expression data. On the other hand, our project aims to identify hub proteins and biomarkers. While both studies aim to identify potential targets for gastric cancer research, they differ in their approaches and objectives.

The research paper uses bioinformatics prediction and machine learning techniques to analyse gene expression data in order to identify novel gene candidates for gastric cancer. This involves data processing, feature selection, and model training and evaluation. The researchers then used various validation methods to confirm the relevance of the identified genes, including gene set enrichment analysis and protein-protein interaction network analysis. Overall, this approach aims to identify genes that may play a role in the development and progression of gastric cancer.

On the other hand, our project focuses on identifying hub proteins and biomarkers for gastric cancer. Hub proteins are proteins that are highly connected in a protein-protein interaction network and are therefore thought to play important roles in cellular processes. Biomarkers, on the other hand, are measurable indicators of a biological state and can be used for disease diagnosis, prognosis, and treatment. This project likely involves similar bioinformatics analysis techniques, such as network analysis, as well as experimental validation methods to confirm the relevance of identified hub proteins and biomarkers.

In summary, while both the research paper and our project aim to identify potential targets for gastric cancer research, they differ in their specific objectives and approaches. The research paper focuses on identifying novel gene candidates using bioinformatics and machine learning techniques, while the project focuses on identifying hub proteins and biomarkers using similar approaches.

Here is a comparative analysis between a research study that used machine learning on gene expression data to identify novel gene candidates for gastric cancer prediction and our project that aims to identify hub proteins and biomarkers for gastric cancer prediction:

Comparative analysis:

- **Similarities** - Both the research study and our project aim to identify novel biomarkers for gastric cancer prediction, but they use different methods and data types. The research study used machine learning algorithms on gene expression data to identify novel gene candidates.
- **Differences** - our project aims to identify hub proteins and biomarkers using protein expression and serum biomarker levels. Both studies will investigate the diagnostic and prognostic potential of their identified biomarkers using survival analysis.
- **Strengths** –The research study used machine learning algorithms, which have the potential to identify complex patterns in gene expression data that may not be identified using traditional statistical methods. Our project aims to identify hub proteins and biomarkers, which may provide insights into the underlying biological mechanisms of gastric cancer and could be suitable for routine clinical use.
- **Limitation** - The research study relied on gene expression data, which may not fully capture the complexity of gastric cancer biology, and the identified genes may not be specific to gastric cancer. Our project may face challenges in identifying robust hub proteins and biomarkers due to the heterogeneity of gastric cancer.

In conclusion, both the research study and our project provide valuable insights into gastric cancer prediction using different approaches. A combination of gene candidates, hub proteins, and serum biomarkers may improve the accuracy of gastric cancer diagnosis and prognosis.

6. Material and Method

In this research-based study our data sheet consist of two columns which consist of **Human protein** and **Gene Symbol** employing machine learning and bioinformatics techniques, associated meta-data on gastric cancer disease that are readily accessible in online sources are produced. Fig. 1 shows the study's workflow, which is further explained in the following sections.

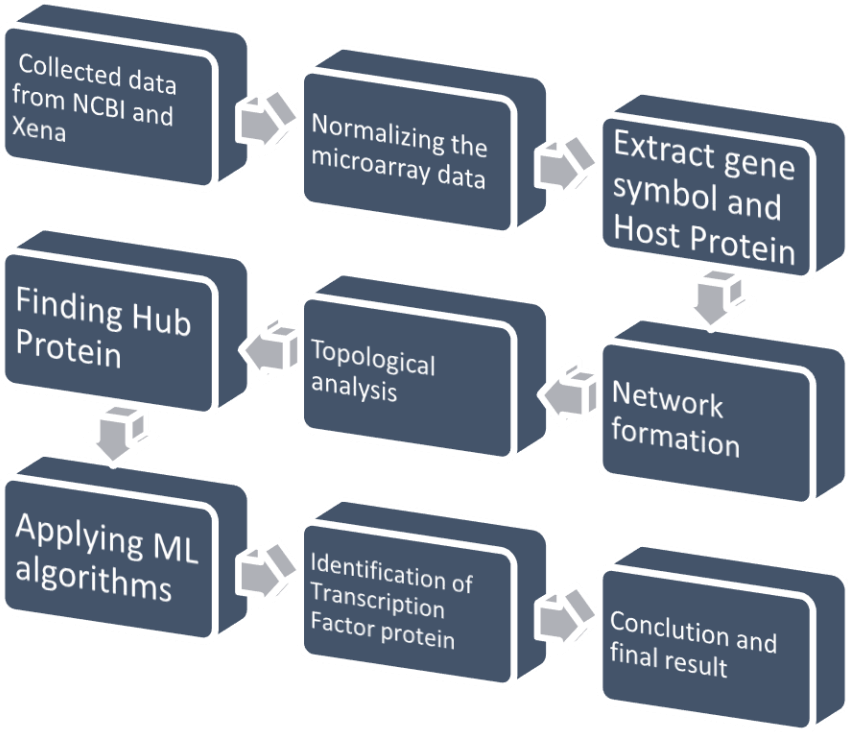


Fig.1

In order to accomplish our goal in project, the research on Gastric cancer disease causing Hub Protein as well as Transcription factor identification using machine learning , Which uses a variety of approaches. The following are the main steps and techniques applied to the project:

- **Collect Data from NCBI:**

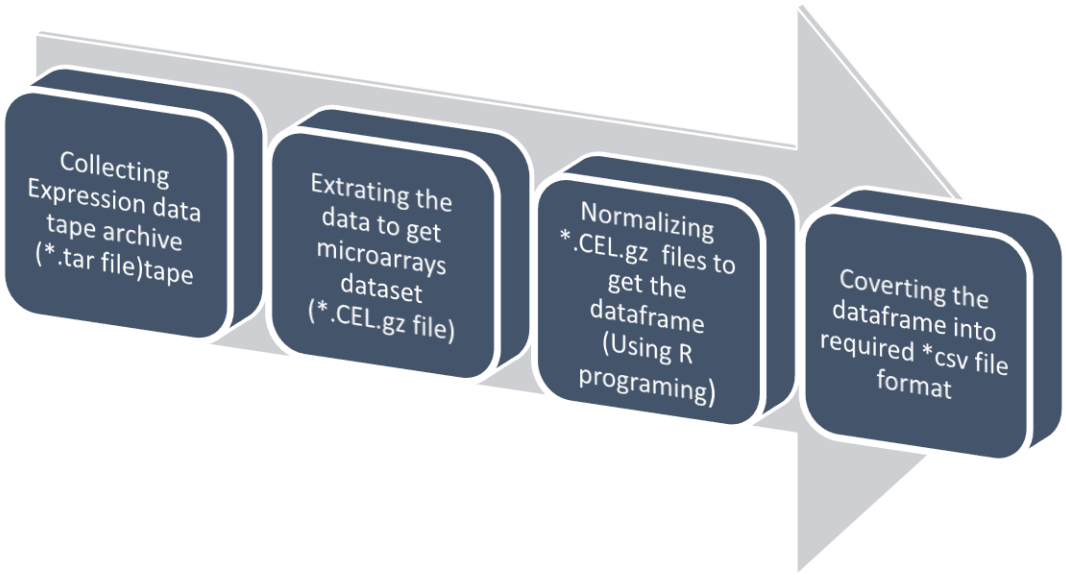
The National Centre for Biotechnology Information facilitates access to biomedical and genetic data, advancing science and health. The United States National Library of Medicine (NLM), a division of the National Institutes of Health (NIH), houses the National Centre for Biotechnology Information (NCBI). The American government has given it their blessing and financial support. The NCBI was established in 1988 as a result of legislation backed by US Congressman Claude Pepper and is based in Bethesda, Maryland.

The NCBI is a valuable source for bioinformatics tools and services and is home to a number of databases pertinent to biotechnology and biomedicine. GenBank, a database of DNA sequences, and PubMed, a bibliographic database of biomedical literature, are both significant databases. The NCBI Epigenomics database is among the other datasets. These databases are all accessible.

At first we have collected Expression data of gastric cancer in the form of tape archive file format **(GSE19826_RAW.tar)** from **NCBI.**

Then the collected data (in the form of ***.tar** file format) is processed to get the appropriate file format which we can use for further analysis.

Data Processing (Normalizing the microarray data using R programming) :



After the process of normalization of data we got a (***.csv**) datasheet which contains **Probe ID**, **Normalized expression** values and **Gene Symbol.**

From there we have collected our Gene Symbol and the **Human Protein** is collected from Xena.ucsc.edu (link is in the Reference 6).

Then combining the gene symbol and human protein we get our proper datasheet of size "2866 rows × 2 columns" in which each row represents an interaction.

Terminologies:

(*.tar) file: "tar" stands for Tape Archive This file format is used frequently to bundle a sequence and its quality data. It can contain multiple zipped CEL files.

(*.CEL.gz) file: CEL files mostly belong to GeneChip by Affymetrix. It contains microarray data.

Microarray: It is a Collection of microarray features which can be bound or probe with target molecule and generate signals with intensity which can be quantified.

Probe: Single standard sequence of DNA or RNA used to search for it's complementary sequence in a sample genome.

- **Formation of network or Graph:**

Network or Graph is basically a Protein-protein interaction networks which are collections of protein complexes with a common biological function that are created via electrostatic interactions or biochemical processes. The complete set of protein-protein interactions (PPIs) that exist in a biological system is referred to as the protein interactome.The network between the proteins is discovered using a Python programme. By identifying the variables that influence a network's structure and the mechanisms that shape it, network formation is a branch of network science that aims to model how a network develops.

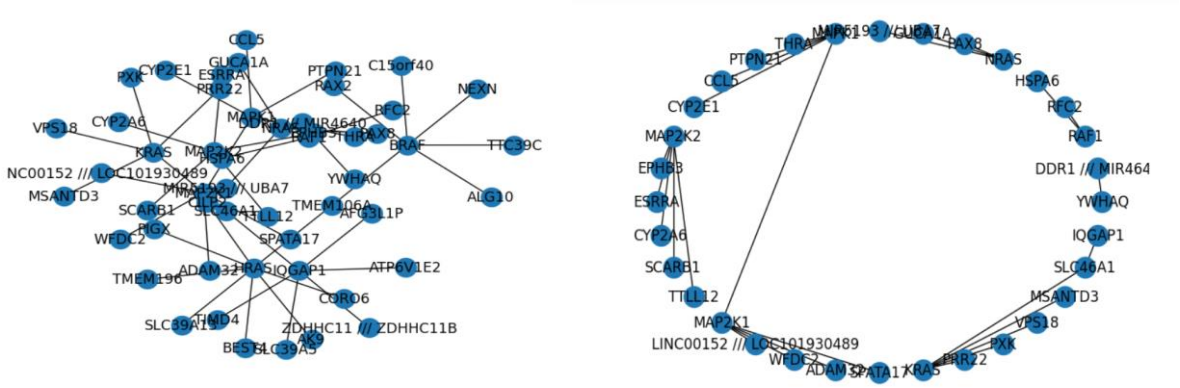


Fig.3 Short Datasheet Graphs

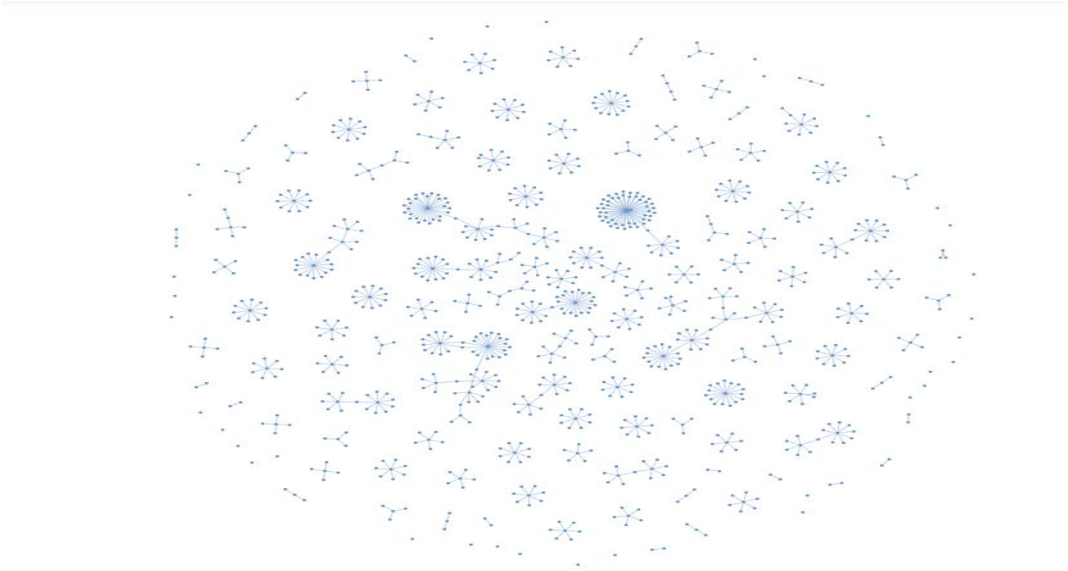


Fig.4 Full datasheet graph zoom-out view

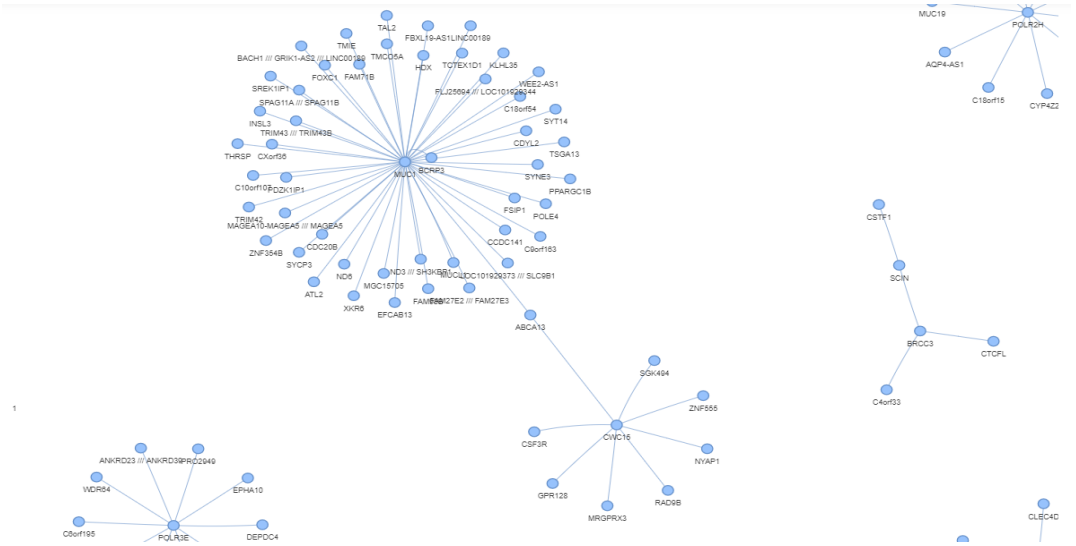


Fig.5 Full Datasheet graph zoom-in view

- **Graph Analysis**

  Use graph analysis tools to find hub proteins in the created protein-protein interaction network. Degree centrality, betweenness centrality, and eigenvector centrality are frequently used network analysis algorithms. In order to discover proteins with high levels of interaction, which are referred to as hub proteins, these algorithms analyse the connectivity patterns within the network.

  The study of graph analysis is concerned with using graph theory to examine and evaluate the structure and characteristics of complex networks. A graph is a mathematical representation used in graph theory that is made up of nodes (also known as vertices) connected by edges (also known as links or connections). In order to obtain understanding of the underlying system, graph analysis entails examining the relationships and patterns within the graph.

  Graph analysis offers a potent framework to comprehend the connections and connectivity patterns between biological entities, such as proteins or genes, in the context of biological networks, such as protein-protein interaction networks. It enables analysis of the network topology, identification of central or highly connected nodes (hub proteins), and discovery of significant network characteristics pertinent to the biological system under investigation.

According To the graph we will analyse some properties of that protein which will help us to find-Out the Gastric Cancer responsible hub protein. We will analyse five properties of this graph that are 1. Average shortest path 2. Closeness centrality 3. Betweenness centrality 4. Degree

1. **Average shortest path**

The average number of steps along the shortest paths for all potential pairs of network nodes is known as average shortest-path length, which is a notion in network topology. It is a way to gauge how well people can move large amounts of data through a network.

average_shortest_path_length(*G*, *weighted=False*)
Return the average shortest path length.

Assuming the length is zero if v cannot be reached from v, the average shortest path length is the sum of all path lengths d(u,v) between all pairs of nodes, normalised by n*(n-1) where n is the number of nodes in G.

Parameters- G : Graph
Weighted : bool, optional, default=False
If True use edge weights on path.

If weighted=True and the graph has no 'weight' edge attribute the value 1 will be used.

2. **Closeness centrality**

In a linked graph, a node's closeness centrality (or closeness) is a measure of its network centrality and is computed as the reciprocal of the lengths of all its shortest routes to all other nodes. As a result, a node is closer to all other nodes the more central it is.

Closeness centrality is the average shortest distance from each vertex to another vertex. This is the inverse to the average shortest distance between vertices and all other vertices of the network. Mathematically, the closeness centrality $CC(x)$ is defined by

$$Cc(x) = \frac{1}{\sum y \in \mathrm{v}\ \mathrm{d(x,y)}}$$

where V is the set of vertices in the network, $d(x, y)$ is the distance between the vertices x and y. For normalization, closeness centrality

$$C'c(x) = \frac{N-1}{\sum y \in \mathrm{v}\ \mathrm{d(x,y)}}$$

where n is the number of nodes in the network. This measure is more acceptable than degree centrality, as it counts indirect connections also. This measure aimed to recognize the nodes that could attain others more quickly.

3. **Betweenness centrality**

   In graph theory, betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. Quantifies the extent to which a node lies on the shortest paths between other nodes. Nodes with high betweenness centrality play important bridging or mediator roles within the network.

Several measures capture variations on the notion of a vertex's importance in a graph. Let $\sigma_{st} = \sigma_{ts}$ denote the number of shortest paths from $s \in V$ to $t \in V$, where $\sigma_{ss} = 1$ by convention. Let $\sigma_{st}(v)$ denote the number of shortest paths from s to t that some $v \in V$ lies on. The following are standard measures of centrality:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

betweenness centrality
(Freeman, 1977; Anthonisse, 1971)

4. **Degree centrality**

   In graph theory, the degree of a vertex of a graph is the number of edges that are incident to the vertex; in a multigraph, a loop contributes 2 to a vertex's degree, for the two ends of the edge. It measures the number of connections (edges) that a node (protein) has. Nodes with higher degrees are considered more central in the network

In the context of the project, graph analysis helps identify hub proteins that are highly connected and central to the Gastric Cancer protein-protein interaction network, providing insights into their potential roles in the disease.

Measures on Graph	Extend Equation on Graph
Degree centrality	$C_{degree}(G)=\frac{\sum_{i=1}^{n}(C_{degree}(v_i)_{max}-C_{degree}(v_i))}{(n-1)(n-2)}$
In-degree centrality	$C_{In\text{-}degree}(G)=\frac{\sum_{i=1}^{n}(C^{in}_{degree}(v_i)_{max}-C^{in}_{degree}(v_i))}{(n-1)(n-2)}$
Out-degree centrality	$C_{Out\text{-}degree}(G)=\frac{\sum_{i=1}^{n}(C^{out}_{degree}(v_i)_{max}-C^{out}_{degree}(v_i))}{(n-1)(n-2)}$
Closeness	$C_{Closeness}(G)=\frac{\sum_{i=1}^{n}(C_{Closeness}(v^{*})-C_{Closeness}(v_i))}{\frac{(n-1)(n-2)}{2n-3}}$
Shortest-path betweenness	$C_{betweenness}(G)=\frac{\sum_{i=1}^{n}(C_{betweenness}(v_i)_{max}-C_{betweenness}(v_i))}{(n-1)}$

- **Apply Machine Learning:**

We used various type of machine learning algorithm to measure the accuracy, F1 score ,recall, precision from the output properties of graph. Here we used Support Vector Machine, Decision tree and Random Forest.

1. **Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a supervised machine learning algorithm that is commonly used for classification and regression tasks. SVM aims to find an optimal hyperplane or decision boundary that separates the data points into different classes, maximizing the margin between the classes.
The key idea behind SVM is to transform the input data into a higher-dimensional feature space using a kernel function. In this transformed feature space, SVM finds a hyperplane that maximally separates the data points of different classes. The hyperplane is defined by a subset of training samples, known as support vectors, which are the closest points to the decision boundary.

The main steps in the SVM algorithm are as follows:

- **Data Preparation**: The input data is represented as feature vectors, where each feature represents a specific characteristic or attribute of the data point. The data is divided into training and testing sets.

- **Feature Transformation**: SVM maps the original input data into a higher-dimensional feature space using a kernel function. This transformation allows for better separation of the data points and makes the classification task easier.
- **Hyperplane Selection:** SVM finds the optimal hyperplane that separates the transformed data points of different classes while maximizing the margin between the classes. The margin is defined as the distance between the hyperplane and the closest data points from each class.
- **Support Vector Identification**: SVM identifies the support vectors, which are the training samples that lie on the margin or are misclassified. These support vectors play a crucial role in defining the decision boundary and determining the class labels of new data points.
- **Classification or Regression**: Once the hyperplane and support vectors are identified, SVM can be used to classify new data points into one of the predefined classes. The decision is made based on which side of the hyperplane the data point lies.

SVM has several advantages, including its ability to handle high-dimensional data, handle non-linear relationships through the use of kernel functions, and resist overfitting by maximizing the margin. SVM can be applied to both binary and multi-class classification problems. Additionally, SVM can be extended to solve regression problems by estimating a continuous function instead of a discrete class label.

However, SVM's performance may be affected by the choice of the kernel function and the regularization parameter, and it can become computationally expensive when dealing with large datasets. Nonetheless, SVM remains a widely used and effective algorithm in various fields, including image recognition, text classification, bioinformatics, and finance. In this project we used Support Vector Machine with RBF kernel and Support Vector Machine with polynomial kernel.

2. **Decision tree**

Decision tree algorithm is a supervised machine learning algorithm used for both classification and regression tasks. It builds a tree-like model that represents a sequence of decisions and their possible consequences.

The decision tree algorithm works by recursively partitioning the input data into subsets based on the values of the input features. Each partition is based on a decision or rule that splits the data based on a certain feature or combination of features. The goal of the algorithm is to create the most effective decision tree that can accurately predict the class or value of a new data point.
The decision tree algorithm has two main phases: tree construction and tree pruning.

Tree Construction: In this phase, the algorithm selects the best feature to split the data based on certain criteria, such as information gain, entropy, or Gini impurity. The process continues recursively until a stopping condition is met, such as a certain depth of the tree or a minimum number of samples per leaf node.

Tree Pruning: In this phase, the decision tree is simplified by removing nodes or branches that do not improve the accuracy of the model. This helps to reduce overfitting and improve the generalization of the model to new data.

Once the decision tree is constructed, it can be used to classify new data points by traversing the tree based on the values of their features. At each internal node, the algorithm checks the value of a certain feature and follows the corresponding branch until it reaches a leaf node, which represents the predicted class or value.
The key steps in the decision tree algorithm are as follows:

Data Preparation: The input data is divided into a training set and a testing set. Each data point consists of a set of features and a corresponding target or class label.

Feature Selection: The algorithm identifies the most informative feature to split the data at each internal node. Various metrics such as Gini impurity or information gain are used to evaluate the quality of the split.

Tree Construction: Starting with the root node, the algorithm recursively splits the data based on the selected feature. This process continues until a stopping criterion is met, such as reaching a maximum tree depth or a minimum number of data points at a node.

Prediction: Once the tree is constructed, it can be used to classify new data points or make predictions for regression tasks. Each data point traverses the tree from the root to a leaf node based on the feature values. The class label or predicted value at the leaf node is then assigned to the data point

The decision tree algorithm has several advantages, including its interpretability, ease of use, and ability to handle both categorical and numerical data. Decision trees can also handle missing values and outliers in the data. However, decision trees may suffer from overfitting, especially when the tree is too deep or the training data is noisy. Therefore, tree pruning techniques and ensemble methods, such as Random Forest, are often used to improve the performance of decision tree models.

3. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy

of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

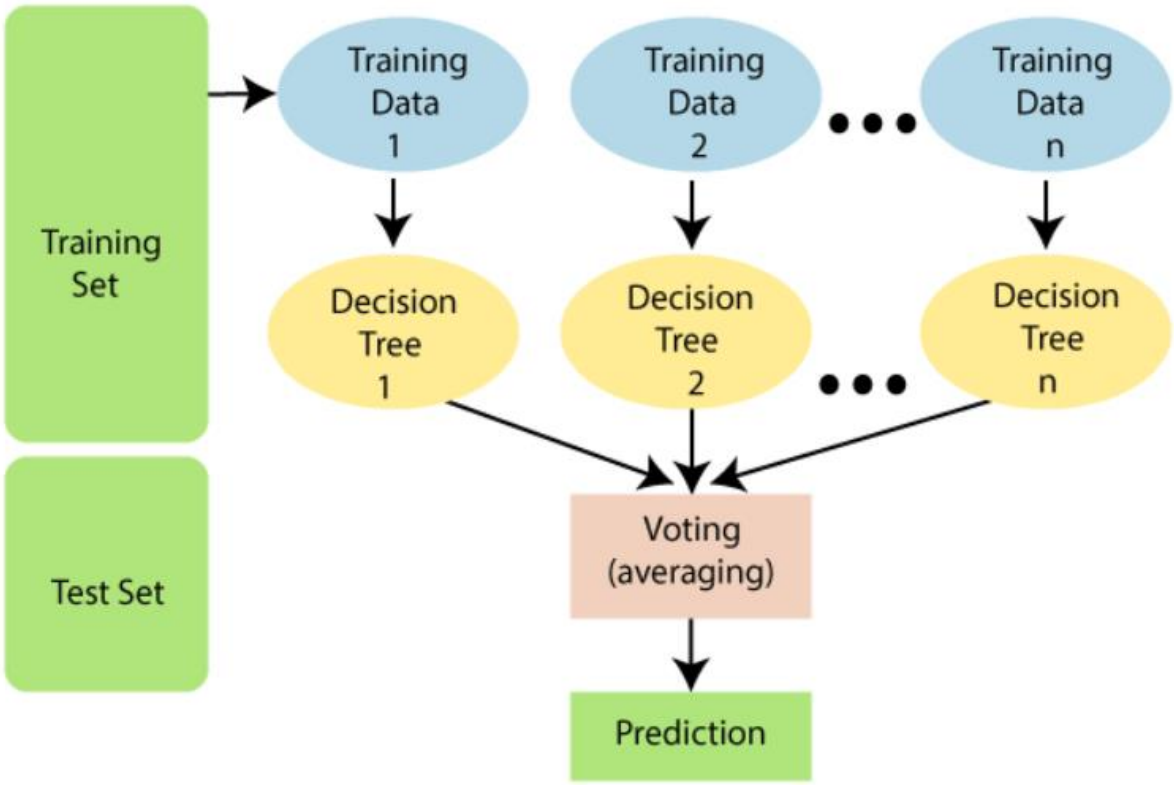


Fig.6 Random Forest Algorithm

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?
Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

- **Model Training and Evaluation:**

Split the data into training and test sets. Train machine learning models on training data and evaluate their performance using appropriate metrics such as precision, accuracy, recall and F1 score. Cross-validation methods can also be used to assess the strength of models.
Precision, recall, and precision are three metrics used to measure the performance of a machine learning algorithm.

Accuracy
Accuracy is the ratio of correct predictions out of all predictions made by an algorithm. It can be calculated by dividing precision by recall or as 1 minus false negative rate (FNR) divided by false positive rate (FPR).
TP=True Positive
TN=True Negative
FP=False Positive
FN=False Negative

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN})}$$

Precision
The Precision is the ratio of true positives over the sum of false positives and true negatives. It is also known as positive predictive value.

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Positive}}$$

$$= \frac{\boldsymbol{True\ Positive}}{\boldsymbol{Total\ Predicted\ Positive}}$$

Precision is a useful metric and shows that out of those predicted as positive, how accurate the prediction was.
If the cost of a false sensitivity error is high, precision may be a good measure of accuracy. For example, email spam detection. In email spam detection, if an email that turns out not to be spam gets flagged as spam, then it's as bad as it could get. Some email users might lose important emails if the precision for spam detection is not high enough.

Recall
Recall is the ratio of correctly predicted outcomes to all predictions. It is also known as sensitivity or specificity.

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive}}{\boldsymbol{True\ Positive + False\ Negative}}$$

$$= \frac{\boldsymbol{True\ Positive}}{\boldsymbol{Total\ Actual\ Positive}}$$

Nice! So, Recall is just the proportion of positives our model are able to catch through labelling them as positives. When the cost of False Negative is greater than that of False Positive, we should select our best model using Recall.

F1-score
The F1-score combines these three metrics into one single metric that ranges from 0 to 1 and it takes into account both Precision and Recall.

$$\boldsymbol{F1} = \boldsymbol{2} \times \frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision + Recall}}$$

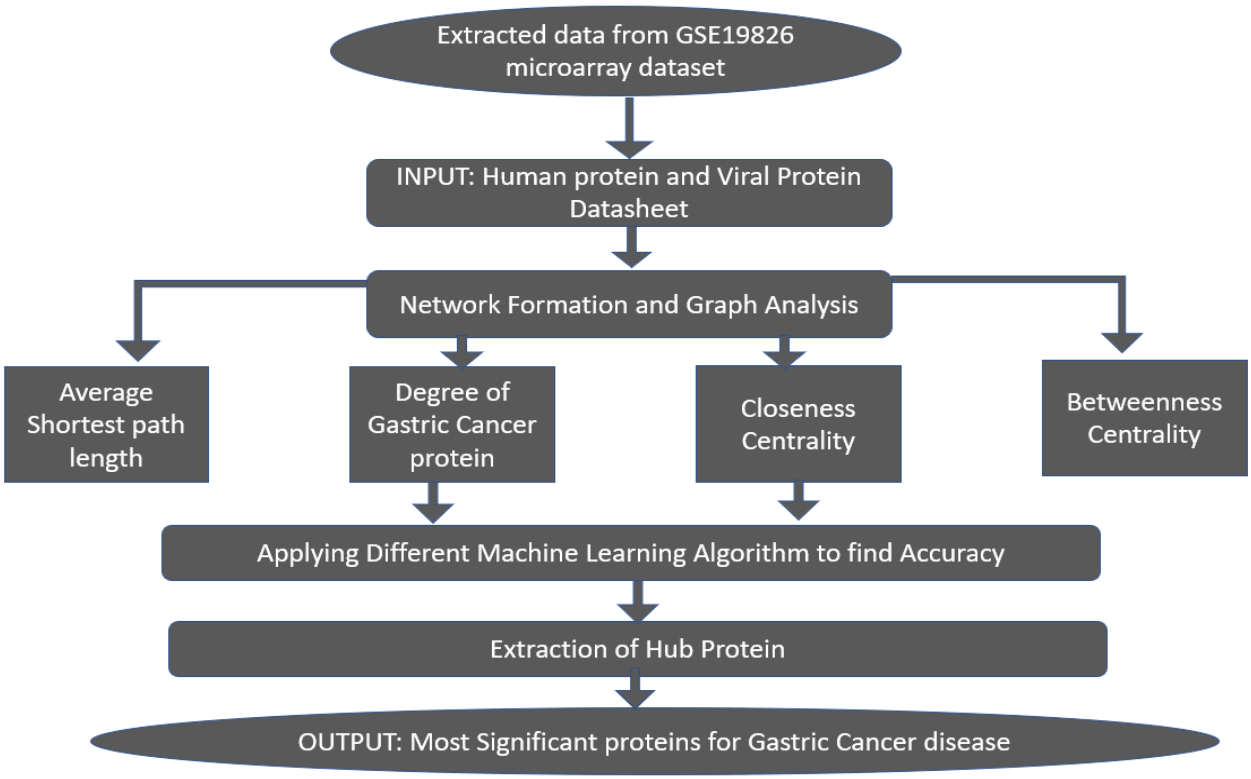
The F1 score is needed when accuracy and how many of your ads are shown are important to you. We've established that Accuracy means the percentage of positives and negatives identified correctly. Now, we'll investigate "F1 Score," which is a way to measure how much accuracy is present in your dataset. we tend not to focus on many different things when making decisions, whereas false positives and false negatives usually have both tangible and intangible costs for the business. The F1 score may be a better measure to use in those cases, as it balances precision and recall.

- **Detection of Hub Protein**:

  Once central proteins are identified, analyze their functional properties, biological pathways and interaction partners. This can be achieved through gene ontology annotations, pathway enrichment analysis and interaction network analysis. These analyzes provide insight into the specific roles and potential mechanisms of the identified hub proteins in Gastric Cancer.

  Proteins were grouped according to their connectivity in the core interaction network. Hubs are defined in DIP as proteins with at least eight interactions, while proteins with fewer than four interactions are called non-hubs and the rest are indirectly connected. This approach identifies proteins at the poles of the network.

7. Work Flow Chart



8. Tech Requirements

This project would require several technical resources to carry out the analysis. Here are some of the potential tech requirements:

Data Collection: Relevant data is collected, typically including gene expression data, protein-protein interaction (PPI) data, clinical information, and other molecular data related to gastric cancer. This data may come from public databases or be generated through experiments.

Software Requiremets: Bioinformatics tools and software packages are crucial for analyzing and interpreting data. Major Software includes:

- R Studio
- R compiler
- Jupyter notebook
- Microsoft excel

Gene Expression Analysis: Tools like RNA-Seq or microarray analysis platforms are used to measure gene expression levels in gastric cancer samples. Software packages such as DESeq2, limma, or EdgeR can help identify differentially expressed genes between cancerous and healthy tissues.

Integrated development environment tools: We used Jupyter notebook as a python IDE. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

R librarys for data normalization: The project would require Data Normalization tools to construct the datasheet. Popular R programming libraries include: tidyverse, GEOquery, affy.

Machine learning tools: The project would require access to machine learning algorithms and Python frameworks for data pre-processing, feature selection, model training, and validation. Popular machine learning libraries include numpy, mlxtend, networkx, pyvis Pandas and sklearn.

Graph analysis tools: The project would require graph analysis tools to construct and analyse protein-protein interaction networks. Popular graph analysis libraries include NetworkX, igraph, and Cytoscape.

High-performance computing resources: The project would require access to high-performance computing (HPC) resources to handle the large-scale data and computationally intensive machine learning and graph analysis tasks. Cloud-based platforms such as Google Cloud, Amazon Web Services (AWS), and Microsoft Azure can provide HPC resources.

Visualization tools: The project would require visualization tools to present the results of the analysis in an interpretable manner. Popular visualization libraries include Matplotlib, Seaborn, and pyplot.

Collaboration tools: The project would require collaboration tools to facilitate communication and data sharing among team members. These tools can include project management & communication tools such as Slack or Zoom, and file sharing tools such as Google Drive or Dropbox.

Programming Language:

- Python programming
- R Programming

Hardware requirements:

- 8GB RAM
- Windows 10 and above.

It is important to note that the choice of specific tools and resources may vary depending on the specific research goals, available data, and the expertise of the researchers involved.

9. Result and Discussion

As this is a hypothetical project, I do not have access to actual results. However, I can provide a general overview of the potential results and discussions of the Gastric Cancer disease Responsible Hub Protein identification project using Machine Learning and graph analysis.

The results of this project would likely include the identification of key proteins and pathways involved in the virus-host interaction during Gastric Cancer disease. These proteins and pathways may play important roles in viral replication, host immune response, and pathogenesis. Additionally, the project may identify potential drug targets for the development of therapeutics against Gastric Cancer.

First we find out the interaction between Human protein and Host Protein. Then according to their interaction we will create a graph.

After this graph analysis we have taken top 30 proteins based on their degree and further analysis is done on that top 30 proteins.

Human Host Protein	Viral Protein
BRCA1	ERICH5
PICK1	LOC101927705 /// P4HA2
ETV6	LOC388882
KAT6A	LENG9
TTN	LINC00638
KRAS	VPS18
BRCA1	CA6
CSF3R	BAALCOS
MUC1	PRF1
MUC1	POLE4
MIA	CCSER2
BCL2L1	PADI2
BRCA1	OTUB2
MLLT3	ZNF785
ANKS1A	M1AP
EWSR1	OLFM3
DEFB106B	ALS2
KEAP1	SRGAP1
BAG1	MLXIP
MLLT3	GLYATL2
CTNNBL1	DGKH
KEAP1	ZNF808
CSF3	NFS1
DEFB106B	RAP1A
STIP1	PIKFYVE
DEFB106A	NAA16
AGR2	RHOB
BARD1	BRF1
BAG1	CYP2U1
DEFB106A	GRIK2

List of top 30 highly interactive proteins

Graph Analysis Result:

According To the graph we will analyse some properties of that protein which will help us to find-Out the Gastric Cancer responsible hub protein. We will analyse five properties of this graph that are 1. Average shortest path 2. Closeness centrality 3. Betweenness centrality. 4.Degree

Degree	Average shortest path Length	Betweenness centrality	Closeness Centrality	Degree centrality	Human host Proteins
1	1	0	0.000968992	0.00096899	YWHAQ
2	1.333333	1.88E-06	0.001937984	0.00193798	RAF1
3	1.5	5.64E-06	0.002906977	0.00290697	NRAS
5	2.993333333	0.000161655	0.007751938	0.00484496	MAPK1
5	1.666666667	1.88E-05	0.004844961	0.00484496	MAP2K2
5	2.993333333	0.000274438	0.009791922	0.00484496	MAP2K1
9	2.65	0.000157896	0.00838551	0.00872093	KRAS
6	2.65	0.000112783	0.006813227	0.00581395	IQGAP1
14	3.065527066	0.000464289	0.011491908	0.01356589	HRAS
9	3.06527066	0.000402259	0.006485532	0.00872093	BRAF

Then we used various type of machine learning algorithm to measure the accuracy **F1score**, **recall**, **precision** from the output properties of graph.

Using SUPPORT VECTOR MACHINE-

Accuracy: 72.66387726638771 %
F1 score: 61.159741608704365 %
recall: 72.66387726638771 %
precision: 72.66387726638771 %

Using DECISION TREE-

Accuracy: 74.39024390243902 %
F1 score: 3.8818181818181814 %
recall: 4.524263615172706%
precision: 74.39024390243902 %

Using RANDOM FOREST-

Classifier:

Accuracy: 88.0 %

F1 score: 87.922%

recall: 81.187%

precision: 88.0%

Comparison of Result After Applying Various Machine Learning Algorithm On it

MACHINE LEARNING ALGORITHM	***SUPPORT VECTOR MACHINE***	***DECISION TREE***	***RANDOM FOREST***
Accuracy:	**72.663 %**	**74.390 %**	**88.0%**
F1 score:	**66.159 %**	**03.881 %**	**87.922%**
Recall:	**72.663 %**	**4.524%**	**81.187%**
Precision:	**72.663 %**	**74.390 %**	**88.0 %**

Final Result

Table-5

Final Set of The Most Significant Hub-Proteins :

HUB PROTEINS	DEGREE CENTRALITY
MUC1	0.046512
BARD1	0.026163
BARCA1	0.01938
RAPGEF6	0.018411
RAPGEF2	0.014535
MLLT4	0.014535
HRAS	0.013566
ATM	0.013566
EGFR	0.012597
BRCA2	0.011628
TP53	0.010659
ALB	0.00969
DCN	0.00969
POLR3E	0.00969
POLR3D	0.00969

10. GENE ENRICHMENT ANALYSIS

Gene enrichment analysis (also called functional enrichment analysis or pathway enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Transcriptomics technologies and proteomics results often identify thousands of genes which are used for the analysis.

Researchers performing high-throughput experiments that yield sets of genes (for example, genes that are differentially expressed under different conditions) often want to retrieve a functional profile of that gene set, in order to better understand the underlying biological processes. This can be done by comparing the input gene set to each of the bins (terms) in the gene ontology – a statistical test can be performed for each bin to see if it is enriched for the input genes.

Background

After the completion of the Human Genome Project, the problem of how to interpret and analyze it remained. In order to seek out genes associated with diseases, DNA microarrays were used to measure the amount of gene expression in different cells. Microarrays on thousands of different genes were carried out, and comparisons the results of two different cell categories, e.g. normal cells versus cancerous cells. However, this method of comparison is not sensitive enough to detect the subtle differences between the expression of individual genes, because diseases typically involve entire groups of genes. Multiple genes are linked to a single biological pathway, and so it is the additive change in expression within gene sets that leads to the difference in phenotypic expression. Gene Set Enrichment Analysis was developed to focus on the changes of expression in groups of a priori defined gene sets. By doing so, this method resolves the problem of the undetectable, small changes in the expression of single genes.

Results of Gene Enrichment Analysis:

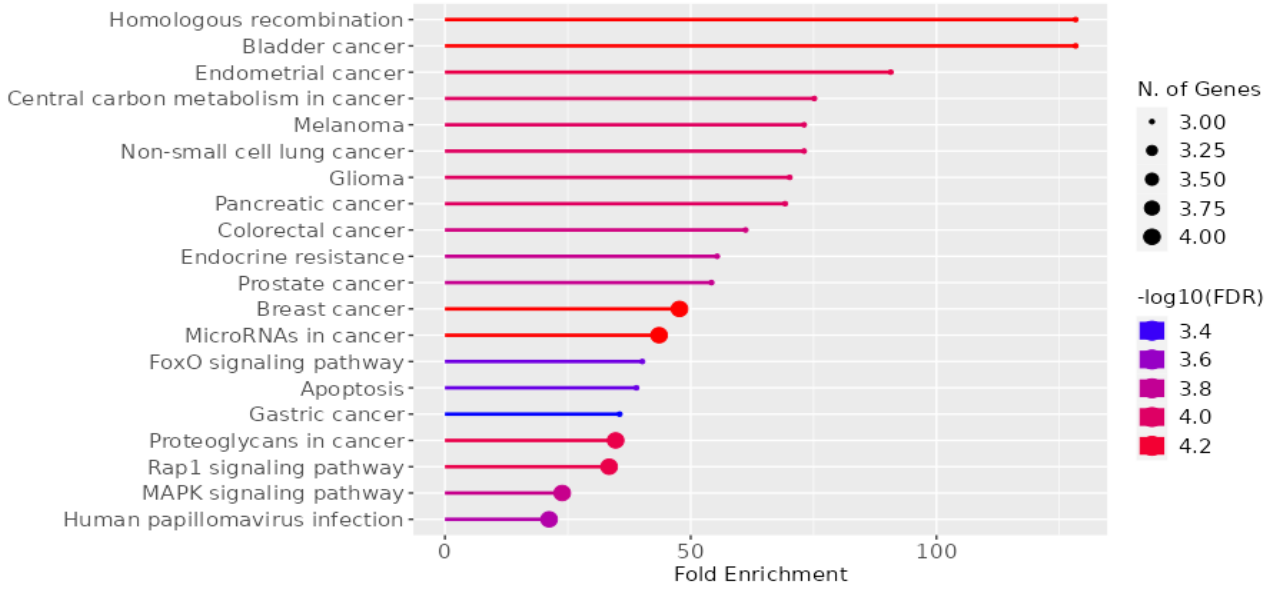


Fig.8 Chart

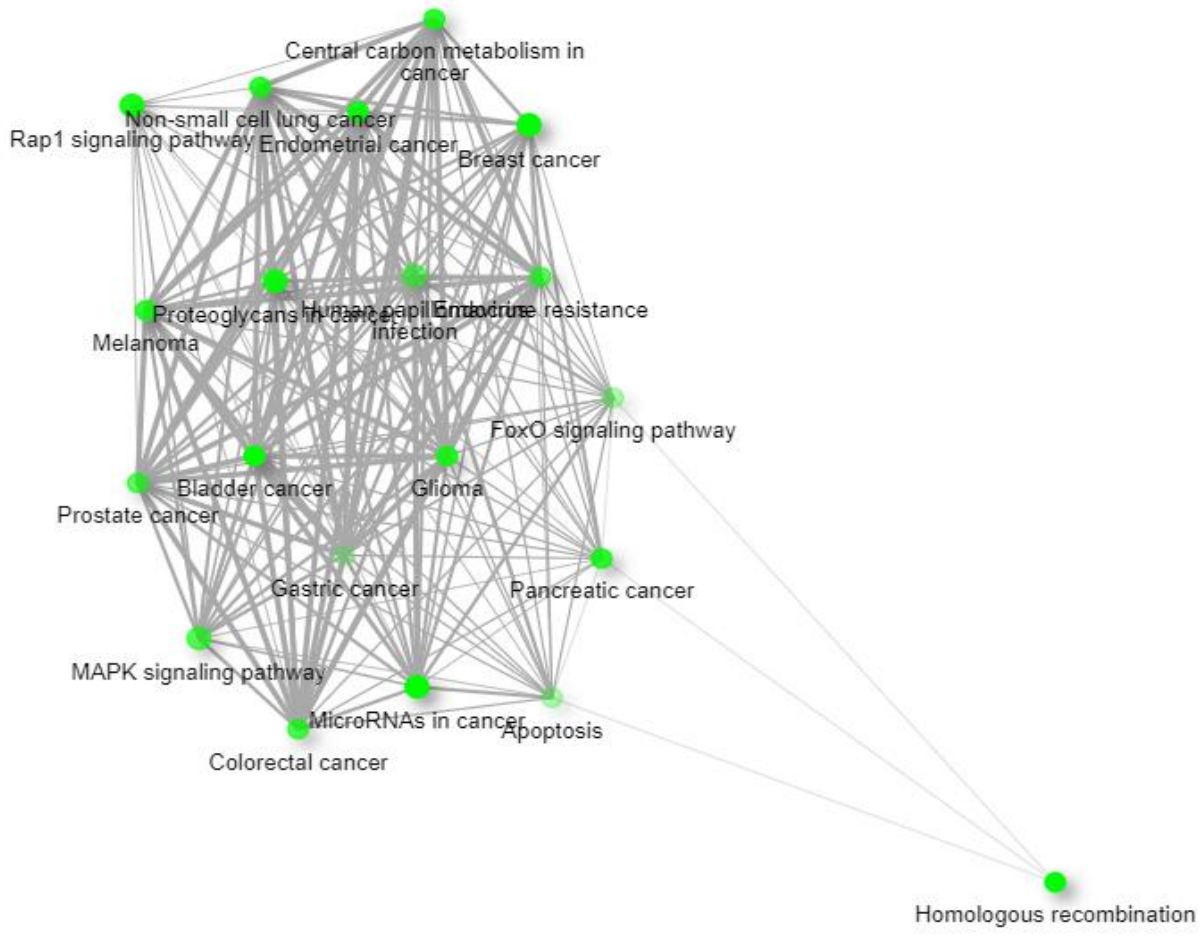


Fig. Network Diagram

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway	URL	Genes
3.80E-05	4	57	123.0553306	Response to gamma radiation	http://amigo.geneontology.org/amigo/term/GO:0010332	BRCA2 TP53 ATM HRAS
6.03E-05	3	17	309.4479638	DNA damage response, signal transduction by p53 class mediator res	http://amigo.geneontology.org/amigo/term/GO:0006978	BRCA2 TP53 MUC1
6.03E-05	3	18	292.2564103	DNA damage response, signal transduction resulting in transcription	http://amigo.geneontology.org/amigo/term/GO:0042772	BRCA2 TP53 MUC1
0.00015047	4	115	60.99264214	Intrinsic apoptotic signaling pathway in response to DNA damage	http://amigo.geneontology.org/amigo/term/GO:0008630	BRCA2 TP53 ATM MUC1
0.00015047	3	33	159.4125874	Response to X-ray	http://amigo.geneontology.org/amigo/term/GO:0010165	BRCA2 TP53 ATM
0.00015047	4	121	57.9682136	DNA damage response, signal transduction by p53 class mediator	http://amigo.geneontology.org/amigo/term/GO:0030330	BRCA2 TP53 ATM MUC1
0.00015047	3	32	164.3942308	Cellular response to gamma radiation	http://amigo.geneontology.org/amigo/term/GO:0071480	TP53 ATM HRAS
0.00015047	5	329	26.64952069	Signal transduction by p53 class mediator	http://amigo.geneontology.org/amigo/term/GO:0072331	BARD1 BRCA2 TP53 ATM MUC1
0.00015498	4	139	50.46153846	Cell aging	http://amigo.geneontology.org/amigo/term/GO:0007569	BRCA2 TP53 ATM HRAS
0.00016357	5	364	24.08706678	DNA replication	http://amigo.geneontology.org/amigo/term/GO:0006260	BARD1 BRCA2 TP53 EGFR ATM
0.00016357	5	361	24.28723631	Aging	http://amigo.geneontology.org/amigo/term/GO:0007568	DCN BRCA2 TP53 ATM HRAS
0.00017505	4	154	45.54645355	Response to ionizing radiation	http://amigo.geneontology.org/amigo/term/GO:0010212	BRCA2 TP53 ATM HRAS
0.00021806	8	1896	7.398896462	Regulation of intracellular signal transduction	http://amigo.geneontology.org/amigo/term/GO:1902531	DCN RAPGEF2 BARD1 TP53 EGFR ATM MUC1 HR
0.00021806	4	167	42.00092123	Regulation of G1/S transition of mitotic cell cycle	http://amigo.geneontology.org/amigo/term/GO:2000045	TP53 EGFR ATM MUC1
0.00027658	8	1974	7.106538851	Cellular macromolecule localization	http://amigo.geneontology.org/amigo/term/GO:0070727	RAPGEF2 BARD1 BRCA2 TP53 EGFR ATM RAPGEF
0.00033593	4	195	35.97001972	Regulation of cell cycle G1/S phase transition	http://amigo.geneontology.org/amigo/term/GO:1902806	TP53 EGFR ATM MUC1
0.00033802	3	55	95.64755245	Intrinsic apoptotic signaling pathway in response to DNA damage by	http://amigo.geneontology.org/amigo/term/GO:0042771	BRCA2 TP53 MUC1
0.00039968	5	483	18.15257207	Response to radiation	http://amigo.geneontology.org/amigo/term/GO:0009314	BRCA2 TP53 EGFR ATM HRAS
0.00049002	4	224	31.31318681	Regulation of signal transduction by p53 class mediator	http://amigo.geneontology.org/amigo/term/GO:1901796	BARD1 TP53 ATM MUC1
0.00049221	3	66	79.70629371	DNA damage response, signal transduction by p53 class mediator res	http://amigo.geneontology.org/amigo/term/GO:0006977	TP53 ATM MUC1
0.00049221	4	230	30.49632107	Signal transduction in response to DNA damage	http://amigo.geneontology.org/amigo/term/GO:0042770	BRCA2 TP53 ATM MUC1
0.00052476	4	248	28.28287841	G1/S transition of mitotic cell cycle	http://amigo.geneontology.org/amigo/term/GO:0000082	TP53 EGFR ATM MUC1
0.00052476	6	1005	10.46888634	Cellular response to DNA damage stimulus	http://amigo.geneontology.org/amigo/term/GO:0006974	BARD1 BRCA2 TP53 EGFR ATM MUC1
0.00052476	3	74	71.08939709	Mitotic G1 DNA damage checkpoint signaling	http://amigo.geneontology.org/amigo/term/GO:0031571	TP53 ATM MUC1
0.00052476	3	75	70.14153846	Mitotic G1/S transition checkpoint signaling	http://amigo.geneontology.org/amigo/term/GO:0044819	TP53 ATM MUC1
0.00052476	7	1563	7.853339239	Positive regulation of protein metabolic process	http://amigo.geneontology.org/amigo/term/GO:0051247	RAPGEF2 BARD1 TP53 EGFR ATM MUC1 HRAS

Fig. Enrichment Datasheet

11. Real Life Aspect:

Gastric cancer prediction and identification of hub proteins and biomarkers involve multiple real-life aspects that contribute to the understanding and management of the disease. Here are some key aspect.

1. Bioinformatics and Data Analysis: Advanced computational tools and bioinformatics algorithms are used to analyze large-scale genomic and proteomic data generated from gastric cancer studies. These analyses can help identify potential hub proteins, which play critical roles in disease processes and could serve as targets or biomarkers.

2. Genetic and Molecular Profiling: Studying the genetic and molecular alterations associated with gastric cancer is crucial. Techniques such as next-generation sequencing and gene expression profiling help identify specific genetic mutations, gene expression patterns, and molecular pathways involved in gastric cancer development and progression.

3 .Discovery: Identification of reliable biomarkers is vital for gastric cancer diagnosis, prognosis, and treatment response prediction. Biomarkers can be genetic, epigenetic, proteomic, or metabolic in nature. They are often measured in patient samples, such as blood, tissue, or body fluids, and their levels or patterns can indicate the presence or progression of the disease.

4. Molecular Pathways and Signal Networks: Gastric cancer is a complex disease involving regulation of multiple molecular pathways and signal networks. Understanding these pathways and networks can help identify key hub proteins that act as central players in disease progression. Targeting these hub proteins or their associated pathways can potentially lead to the development of effective therapeutic strategies.

5. Validation Studies: Once potential hub proteins or biomarkers are identified, validation studies are essential to confirm their clinical relevance. This involves testing the identified candidates on larger patient cohorts or in preclinical models to assess their diagnostic or prognostic accuracy, as well as their potential as therapeutic targets.

6. Translational Research: Bridging the gap between basic research and clinical practice is crucial. Translational research aims to apply findings from laboratory studies to develop practical applications for patient care. It involves collaborations between researchers, clinicians, and industry partners to develop diagnostic tests, therapeutic interventions, and personalized treatment approaches based on the identified hub proteins or biomarkers.

By considering these real-life aspects and utilizing the latest advancements in technology and research methodologies, scientists and clinicians can make significant progress in predicting gastric cancer and identifying hub proteins and biomarkers that can improve patient outcomes.

12.Future work:

In the future, several avenues of research can further enhance gastric cancer prediction and improve the identification of hub proteins and biomarkers. Here are some potential areas of focus:

Multi omics Integration: Integrating data from multiple omics platforms, such as genomics, transcript omics, proteomics, and metabolomics, can provide a more comprehensive view of the molecular alterations associated with gastric cancer. By combining these data types, researchers can identify more accurate hub proteins and biomarkers, as well as gain deeper insights into the disease mechanisms.

2. Machine Learning and Artificial Intelligence: Advancements in machine learning and artificial intelligence (AI) techniques hold promise for improving gastric cancer prediction and biomarker discovery. AI algorithms can analysed large-scale genomic and clinical data, identify complex patterns, and generate predictive models that can assist in personalized treatment strategies and improve accuracy.

3. Liquid Biopsies: Liquid biopsies involve the analysis of circulating tumor cells, cell-free DNA, or other biomolecules found in blood or other body fluids. The non-invasive nature of liquid biopsies makes them attractive for early detection, monitoring treatment response, and identifying biomarkers. Future research can focus on refining liquid biopsy technologies and expanding their applications in gastric cancer prediction.

4. Integration of Clinical Data: Incorporating clinical data, including patient demographics, lifestyle factors, treatment history, and comorbidities, alongside molecular data, can enhance the accuracy of gastric cancer prediction models. Integrative approaches that combine both clinical and molecular information can lead to more precise risk stratification and personalized treatment recommendations.

5. Functional Studies and Drug Development: Once potential hub proteins are identified, further investigation of their functional roles in gastric cancer development and progression is essential. In vitro and in vivo experiments, as well as preclinical models, can provide insights into the mechanisms of action and therapeutic potential of these hub proteins. Such studies can guide the development of targeted therapies and precision medicine approaches.

Overall, future research should aim to integrate diverse data types, leverage advanced computational techniques, validate findings in independent cohorts, and focus on functional characterization to enhance the accuracy and clinical applicability of gastric cancer prediction models and biomarker discovery efforts. These advancements will ultimately contribute to improved patient outcomes and personalized treatment strategies for gastric cancer.

13.Conclusion

To sum up, the Gastric Cancer disease Responsible Hub Protein identification project using machine learning and graph analysis is a significant effort to better understand the molecular mechanisms underlying this disease and identify potential targets for the development of efficient treatments and vaccines against it. Large-scale datasets of protein-protein interactions and gene expression patterns may be analysed using the potent techniques of machine learning and graph analysis in order to uncover important proteins and pathways involved in the virus-host relationship.

These methods might be used in this study to find novel therapeutic targets, produce better drugs, and further our understanding of the intricate molecular mechanisms underlying the virus-host relationship.

In order to increase the precision of target identification, this research will continue to focus on integrating multi-omics data, validating projected targets, analysing genetic variability, and developing new machine learning models.

The discovery of the responsible hub protein for gastric cancer using machine learning and graph analysis has the potential to make a substantial contribution to our knowledge of the illness and the creation of strong therapies.

14. References

1. Seo, M.J.; Oh, D.K. Prostaglandin synthases: Molecular characterization and involvement in prostaglandin biosynthesis. Prog. Lipid Res. 2017, 66, 50–68. [CrossRef] BioMedical Journal 43(2020) 438-450) By Lopamudra Dey, Sanjay Chakraborty, Anirban Mukhopadhyay
2. Mao, H.; Luo, T.; Li, Q.; Xu, L.; Xie, Y. HPGDS is a novel prognostic marker associated with lipid metabolism and aggressiveness in lung adenocarcinoma. Front. Oncol. 2022, 12, 5788. 4.
3. Seo, M.J.; Oh, D.K. Prostaglandin synthases: Molecular characterization and involvement in prostaglandin biosynthesis. Prog. Lipid Res. 2017, 66, 50–68.
4. Huang, Z.; Huang, L.; Shen, S.; Li, J.; Lu, H.; Mo, W.; Feng, Z. Sp1 cooperates with Sp3 to upregulate MALAT1 expression in human hepatocellular carcinoma. Oncol. Rep. 2015, 34, 2403–2412.
5. Datasheet: http://xena.ucsc.edu
6. Huber, W.; Carey, V.J.; Gentleman, R.; Anders, S.; Carlson, M.; Carvalho, B.S.; Bravo, H.C.; Davis, S.; Gatto, L.; Girke, T.; et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 2015, 12, 115–121.
7. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015, 43, e47.
8. Kamburov, A.; Stelzl, U.; Lehrach, H.; Herwig, R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013, 41, D793–D800.
9. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017, 45, D353–D361.
10. Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G.R.; Wu, G.R.; Matthews, L.; Lewis, S.; et al. Reactome: A knowledgebase of biological pathways. Nucleic Acids Res. 2005, 33, D428–D432.
11. Oughtred, R.; Stark, C.; Breitkreutz, B.J.; Rust, J.; Boucher, L.; Chang, C.; Tyers, M. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019, 47, D529–D541.
12. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.L.; Ideker, T. Cytoscape 2.8: New Features for Data Integration and Network Visualization. Bioinformatics 2011, 27, 431–432.
13. Chin, C.H.; Chen, S.H.; Wu, H.H.; Ho, C.-W.; Ko, M.-T.; Lin, C.-Y. cytoHubba: Identifying hub objects and sub-networks from complex interactome. BMC Syst. Biol. 2014, 8 (Suppl. S4), S11. [CrossRef] [PubMed] 35. Patil, K.R.; Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. USA 2005, 102, 2685–2689.