

Pravir Kumar

THESIS PLAG CHECK

 Nishita Singh

Document Details

Submission ID**trn:oid:::27535:138775682****Submission Date****May 13, 2026, 12:09 PM GMT+5:30****Download Date****May 13, 2026, 12:16 PM GMT+5:30****File Name****THESIS PLAG CHECK.docx****File Size****1.2 MB****60 Pages****13,625 Words****82,058 Characters**

2% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

-  **23 Not Cited or Quoted 2%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 2%  Internet sources
- 0%  Publications
- 1%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 23 Not Cited or Quoted 2%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 2% Internet sources
- 0% Publications
- 1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.daffodilvarsity.edu.bd:8080	<1%
2	Student papers	Higher Education Commission Pakistan on 2026-03-02	<1%
3	Internet	isjem.com	<1%
4	Internet	www.grin.com	<1%
5	Publication	Nuriel-Ohayon, Meital, Hadar Neuman, and Omry Koren. "Microbial Changes duri...	<1%
6	Student papers	University of Liverpool on 2025-01-24	<1%
7	Internet	assets-eu.researchsquare.com	<1%
8	Internet	wjso.biomedcentral.com	<1%
9	Internet	www.frontiersin.org	<1%
10	Internet	mdpi-res.com	<1%

11	Internet	pubmed.ncbi.nlm.nih.gov	<1%
12	Student papers	National Institute Of Technology, Tiruchirappalli on 2026-05-12	<1%
13	Student papers	University of Hertfordshire on 2025-09-18	<1%
14	Publication	Xu-Wen Wang, Tong Wang, Yang-Yu Liu. "Artificial intelligence for microbiology a...	<1%
15	Internet	bmcgenomics.biomedcentral.com	<1%
16	Internet	link.springer.com	<1%
17	Internet	www.bio-conferences.org	<1%
18	Internet	www.biorxiv.org	<1%
19	Internet	www.mdpi.com	<1%

COMPUTATIONAL APPROACHES FOR OVARIAN CANCER DRUG REPURPOSING AND GUT MICROBIOME-BASED COLORECTAL DISEASE CLASSIFICATION

DIKSHA TRIPATHI

ABSTRACT

The thesis focuses on the computational approaches for diagnosis of disease and drug discovery. The work combines classification of colorectal disease based on gut microbiome using machine learning and transcriptomics drug repurposing for the treatment of ovarian cancer. In the colorectal disease study, the analysis of gut microbiome data which was collected from stool sample is done to identify differences between healthy individuals and colorectal disease patients. There are various bioinformatics tools such as SRA Toolkit, Kraken2, Bowtie2 and Bracken which has been used to process the publicly available metagenomics datasets. Preprocessing is followed by CLR normalization and disease classification with the help of different machine learning models such as Logistic Regression, SVM, Random Forest, XGBoost, LightGBM, and KNN. The best performance with high accuracy and ROC-AUC score is shown by Logistic Regression combined with SVM(RBF). With the help of results, we can say that gut microbiome profiles may serve as biomarkers for colorectal disease detection.

In the ovarian cancer study, we applied approaches for drug repurposing for identification of potential EPCAM inhibitor molecules. GEO2R and Enrichr is used for Differential gene expression and enrichment analyses. Drug prediction and screening is done by using tools such as LINCS L1000, SwissADME. Further procedure is carried out for molecular docking analysis using AutoDock Vina. There were several compounds which was shortlisted, AT-7519 showed the best binding affinity against the EPCAM protein and lead to indication that AT-7519 has the potential for the treatment of ovarian cancer and can be used as repurposed drug candidate.

Overall, the findings of the studies suggest the importance of gut microbiome profiles which can act as a potential biomarker for detection of colorectal disease. On the other hand, transcriptomics drug repurposing can be used to identify drug candidates for the treatment of ovarian cancer. Thus, the main highlights of the thesis are use of machine learning, metagenomics, transcriptomics and molecular docking for in-silico drug discovery and precision medicine.

Keywords: Machine Learning, ADME Profiling, Molecular Docking, Gut microbiome, SVM, Random Forest

TABLE OF CONTENTS

Title

Candidate's Declaration

Certificate



Acknowledgement

Abstract

CHAPTER-1

INTRODUCTION

CHAPTER-2

LITERATURE REVIEW

2.1 Introduction to Cancer **and** Computational Biology

2.2.1 Overview of Colorectal disease

2.2.2. Global Burden of colorectal cancer

2.2.3 Risk Factors and Disease Progression

2.2.4 Limitations of Conventional Diagnostic Methods

2.3.1 Human Gut microbiome

2.3.2 Gut Microbial Dysbiosis

2.3.3 Gut Microbiome as Biomarker

2.3.4 Specific Microbial Species associated with colorectal cancer

2.3.5 Mechanisms by which Gut microbiota contributes to colorectal cancer

2.4 Metagenomics and microbiome profiling

2.4.1 Metagenomics

2.4.2 16S rRNA and Shotgun Sequencing

2.4.3 OTU Analysis

2.4.4 Challenges in Microbial Data

2.5 Machine Learning in Disease Classification

2.5.1 Introduction to Machine learning

2.5.2 Supervised Learning

2.5.3 Logistic Regression

2.5.4 Support Vector Machine

2.5.5 Random Forest

2.5.6 XGBoost

2.6 Drug Repurposing and Computational Approaches

2.6.1 Introduction to Drug Repurposing

2.6.2 Advantages of drug repurposing

2.6.3 Drug Repurposing in Cancer Research

2.6.4 Computational Approaches in Drug Repurposing

2.6.5 Transcriptomics and Gene expression analysis

2.6.6 Differentially Expressed Genes(DEGs)

2.6.7 Pathway Enrichment Analysis

2.6.8 Connectivity Map (CMap) Analysis

2.6.9 LINCS Database

2.6.10. Molecular Docking

2.6.11. Challenges and limitations of Drug Repurposing

CHAPTER- 3

METHODOLOGY

3.1 Microbiome Data generation and representation

3.2.1 Metagenomics basis of the dataset

3.2.2 Dataset generation and preprocessing

3.3.1 Selection of classifiers

3.3.2 Model evaluation strategy

3.4 Transcriptomics data collection

3.4.1 Differential Gene Expression Analysis

3.4.2 Gene Set Enrichment and Drug Prediction Analysis

3.4.3 Drug Repurposing Strategy and Candidate Selection

3.4.4 Pharmacokinetics and Drug-Likeness Screening

3.4.5 Protein and Ligand Preparation

3.4.6 Molecular Docking Analysis

3.4.7 Visualization and Interaction Analysis

3.4.8 Comparative Analysis and Candidate Prioritization

CHAPTER -4

RESULT AND DISCUSSION

CHAPTER-5

FUTURE SCOPE

CHAPTER-6

CONCLUSION AND SOCIAL IMPACT

CHAPTER-7

REFERENCES

CHAPTER-1

INTRODUCTION

Cancer is reason for most of the death which occurs worldwide and affects a large number of public's challenging their health[1], [2]. There is continuous growth in medical science and technologies of healthcare, but successful treatment of various cancer and their diagnosis at early stage is difficult. Cancer is generally detected when the disease is progressed to third or final stage and this leads to the reduction of effective treatment and affects the health of the patient in negative manner. These challenges lead to increased focus on advanced computational approaches by researchers which might improve diagnosis of the disease, biomarker discovery as well as therapeutic development at faster pace.

Computational approaches and bioinformatics are emerging as an important research area in biomedical research over the few years[3]. Next generation sequencing, transcriptomics, metagenomics, artificial intelligence as well as machine learning are few technologies which have been the key reasons to analyse biological data in large scale with greater speed. The results obtained with the help of these technologies are also very accurate. Experimental methods which were used traditionally for drug discovery and diagnosis were expensive[4]. Not only they were expensive but also time consuming, and took many years in clinical validation and laboratory testing. Poor therapeutic response by many drugs becomes the key reason for failure of drugs. These limitations can be overcome with the help of computational methods and time as well as cost can be reduced too. The efficiency of prediction and analysis is also improved.

Disease classification, biomarker identification, and drug discovery are now widely done with the help of machine learning and computational approaches[5], [6], [7]. These methods are considered useful when sequencing microbiome data and performing transcriptomics gene expression profiles as these are complex biological datasets. The hidden patterns are identified in data and models developed by computational approaches help most of the researchers in understanding the mechanisms of disease and therapeutic strategies are developed more precisely.

Colorectal cancer and ovarian cancer are one of the major healthcare concerns in different cancers as their prevalence is increasing, diagnosis is generally done at very late stage[1], [2], [8]. Colon and rectum are affected in colorectal diseases which majorly includes concerns such as colorectal cancer, inflammatory bowel diseases, adenomas and disorders in gastrointestinal tracts. Among all the cancers, colorectal cancer is generally common and diagnosis of the colorectal cancer is done at later stage which leads to ineffective treatment. There are several traditional methods to treat colorectal cancer which includes colonoscopy and biopsy. Those traditional methods were too pricy, invasive and also inconvenient. Early disease detection is done by researchers by mainly focusing on techniques which are non-invasive computational approach.

Strong connection is diagnosed in between colorectal disease and human gut microbiomes in recent studies[6], [9], [10]. There are trillions of microorganisms in gut microbiome which includes bacteria, fungi, viruses, and archae which reside in the gastrointestinal tract. An important role is performed by these microbial communities such as in digestion, regulation of immune system, metabolism and the overall health of intestine is maintained. Dysbiosis is caused by disturbances in composition of microbial communities and this is very closely related with progression of colorectal disease[8], [11]. Inflammation, growth of tumor and development of disease can occur if there are changes in microbial diversity and in its abundance.

Stool samples can be studied directly from microbial communities which are generally complex due to technologies advancements such as metagenomics sequencing. Identification of diseases whose patterns are synchronised with microbes are done by Operational Taxonomic Unit using microbiome profiling. Though, datasets if microbiome are complex and various microbial features are contained which makes the analysis very difficult if performed manually. Thus, the role of computational approaches and machine learning is one of the most important features in microbe research as they identify biological patterns which are meaningful and also process large- scale datasets. Promising results have been shown by various algorithms such as Logistic Regression, Support Vector Machine, Random Forest and XG Boost on the basis of microbial abundance profiles as healthy and diseased individuals[5], [12], [13]. Accurate and non-invasive diagnostics systems is being developed by computational ways for classification of colorectal disease.

Computational methods are also useful in therapeutics development and modern drug discovery. If we compare modern approaches with traditional methods, the reach of drugs in the market by traditional methods requires time as well as money. Toxicity, low efficacy and poor pharmacokinetics is the reason for many newly developed compound's failures. One of the approaches to solve the later

problem is drug repurposing. In drug repurposing, new therapeutic applications are identified for already approved drugs[4]. Cut down of cost as well as time is done as safety profiles for already approved drugs are known.

In cancer research, one of the computational approaches is Transcriptomics-based drug repurposing which helps in identification of drug candidates. Gene expression patterns which are associated with diseased and healthy conditions are studied under Transcriptomics. At the time of disease progression, those genes that are abnormally regulated are identified and it is generally done by the help of Differential gene expression analysis. Prediction of small molecules who are capable to reverse disease associated expression patterns is done by molecular signatures. Identification of therapeutic candidate is improved by integrating computational screening and transcriptomic analysis.

Symptoms in ovarian cancer remains hidden at early stage which is one of the reasons why it is called “silent killer. Deadliest cancer among gynecological malignancies is known to be ovarian cancer. Most of the cases where patient is suffering from ovarian cancer is progressed at third or fourth stage where survival rate tends to be low and rate of recurrence is very high. Ovarian cancer progression and metastasis is generally contributed by overexpression of cell-surface biomarkers. When signalling pathway and gene expression is dysregulated, it also leads to the cancer progression. Epithelial Cell Adhesion Molecule (EPCAM) is one of the important biomarkers which contributes in proliferation of tumor, progression of cancer and invasion[14]. Treatment of ovarian cancer is being given as an opportunity while targeting EPCAM with the help of computational drug repurposing.

Transcriptomic-driven drug repurposing is done these days with the help of various computational tools such as GEO2R, LINCS L1000, SwissADME, Enrichr, AutoDock Vina[15], [16], [17], [18], [19]. These approaches allow various methodologies to be performed in low cost such as molecular docking, pathway enrichment analysis, differential gene expression. Protein-ligand interactions and compounds identification is done with the help of molecular docking at favourable criteria such as binding affinity as well as at stable condition.

This thesis mainly focuses on computational methods with the help of integration of machine learning, microbiome analysis, use of transcriptomics and also techniques involved in molecular docking.

Objectives of the thesis are:

Objective 1

Gut microbiome-based machine learning models for classification of colorectal disease driven by metagenomics OTU datasets.

Objective 2

Drug repurposing in ovarian cancer identified EPCAM as a potential inhibitor.

CHAPTER-2

LITERATURE REVIEW

2.1 Introduction to Cancer and Computational Biology

Cancer is deadly disease which affects major people across the world and remains one of the reasons for the death per year. It generally occurs when uncontrollably those cells which are normal grows and the reason behind the growth lies on factors such as genetic mutations, environmental factors, and disturbances in signalling pathways occurring normally. Cancer patients have increased in large number in past few years, and the reason behind the increase are lifestyles, smoking, no physical activity, lifestyle which is unhealthy and poor diet[1], [17]. Diagnosis and treatment of cancer have been improved over the past few years, but treatment of cancer is tough as the disease is generally diagnosed at later stage.

Histopathological examination, clinical testing, Imaging techniques are few ways in which cancer was examined and treated traditionally. The methods which were used traditionally were costly and time consuming though patient care and management of disease were improved. At early stage, diagnosis of disease is rare and also explanation of disease at molecular level was difficult with traditional method. Success rate of treating cancer patient is low as disease is diagnosed at late-stage Drug discovery and development of new drugs is extensive process as it requires clinical trials at various stages before approval and utilization in market. There are chance of failure for approval of drugs at later stage as toxicity level is too high.

Researchers are focused on methods and approaches which depends computationally to secure fast results and efficiently for drug discovery due to above limitations. Interdisciplinary fields such as computational biology and bioinformatics are emerging and they are connecting link between

computer science along with biological sciences, mathematics and statistics for analysis of biological data. Traditional methods were unable to secure data regarding biological information. This limitation is tackled due to growth of knowledge in the field of next-generation sequencing, omics and due to growth of Artificial intelligence. Disease progression which is driven by interaction of molecules, variation in genes, pathways associated with disease is being understood by researchers due to techniques in computational biology.

Machine Learning, transcriptomics, metagenomics, molecular docking and network biology are applied in cancer research and these techniques are computational methods. These methods allow analysis of complex datasets and also help to know the pattern of disease which generally cannot be depicted through traditional ways. Biological data is used for the classification of disease using machine learning techniques. Those genes which are expressed abnormally in diseased tissues are identified with the help of transcriptomics analysis. Computational methods such as drug repurposing and molecular docking help in identification of drugs in fast and cheaper ways[4].

Targeted therapy and personalised diagnosis are supported by computational methods in precision medicine recently. Computational approaches along with biological datasets is the reason for the improvement in prediction of disease, biomarker discovery and drug development. Thus, time and cost related issues are solved with the growth of computational biology.

Thus, two studies are focused in this thesis with the former using techniques such as machine learning in classification of colorectal disease and latter focused on drug repurposing for ovarian cancer which is mainly transcriptomics focused. Improved disease diagnosis, drug discovery and biomarker identification is generally occurring as computational approach highlights the growth.

2.2.1 Overview of Colorectal disease

There are various disorders such as colorectal polyps, inflammatory bowel disease, diverticular disease, colorectal cancer and other conditions related to gastrointestinal tracts which causes colorectal diseases[9]. There are various types of disorders and conditions but colorectal cancer remains most dangerous and deadly. Mutations at genetic level as well as changes at cellular level abnormally remains one of the primary reasons for the occurrence of cancer in colon inner lining and rectum. Detection of disease at later stage is one of the reasons for conversion of adenomatous polyps into malignant tumors.

Genetic factors, environmental conditions, lifestyle, diet choices are key reasons for development and progression of colorectal disease. There are various factors such as obesity, smoking, consumption of alcohol, physical patterns that are low, eating patterns which are unhealthy, inflammation of intestine at chronic level which are mainly associated with progression of colorectal cancer. This type of cancer generally occurs in older people more as compared to young generation so age acts as major factor. There is strong link in between progression of colorectal disease and gut microbe communities which have been studied in recent years. Development of tumor is generally supported by changing the composition of microorganism in gut and its diversity which leads to metabolic imbalance and inflammation.

At early stage, colorectal cancer is asymptomatic due to which diagnosis of disease generally occurs at later stage of cancer. The diagnosis of disease at early stage increases the chance of patient's survival and treatment is also done properly if diagnosed at early stage but in most of the cases diagnosis and symptoms occurs very late followed by weight loss, bleeding in rectum, fatigue.

Conventional methods involve colonoscopy, biopsy, sigmoidoscopy, imaging-based techniques. Direct examination of the colon and its identification as well as removal of polyps which are abnormal is done in colonoscopy and this feature makes it more reliable approach among all of these methods mentioned. Though this approach is reliable but it is costly, invasive and takes time and patients complains about being uncomfortable most of the times. Due to poor patient compliance, cost and accessibility which is limited screening program is not practical always.

Colorectal disease identification is generally approached computationally and researches are focused on the treatments which are non- invasive. Analysis of the disease with the help of stool samples is increasing being done due to growth in field of computational methods, microbiome research and metagenomics. Closed relationship with progression of colorectal disease is the major reasons for gaining major attention in human gut microbiome. Machine learning techniques in combination with microbial composition is the reason for promising results while comparing healthy and non-healthy patients.

2.2.2. Global Burden of colorectal cancer

Colorectal cancer is considered to be most dangerous among all the cancers which occurs across the world[1]. As the number of patients are increasing continuously day by day so it is one of the reasons for concerns. Healthcare system is facing serious burden due to increase in number of patients rapidly

as treatment requires costly medical care which are in long run as well requires invasive treatment and costly medical condition. Cancer is considered one of the diseases which not only affects patient physically but also ruins the financial condition and affects families emotionally.

The variation in disease pattern changes due to different lifestyle, food habits, healthcare facilities from one region to another region. In countries which are developed, there are various factors such as consuming processed food, smoking, obesity, life pattern. consumption of alcohol, no exercise all of these factors lead to rise in number of colorectal cancer patients. Though those countries which are not developed and are in urban areas are also witnessing rapid increase in colorectal disease due to changing lifestyles. Disease burden is generally risen due to changing life style pattern. These days physical activity is least and food habit is shifting towards western diet so this leads to overall increase in risk factor of colorectal cancer.

Colorectal cancer is closely associated with age as a major risk factor which is one of the reasons why patients in older age is suffering with this disease majorly. This does not mean that individuals who are in younger age are not suffering from this disease as in recent diagnosis most of the patients are young. The main reasons for this occurrence is generally dietary patterns which are unhealthy, stress, obesity, lack of exercise and composition of gut microbiome which are altered. This pattern which is generally rising these days in modern cancer research contributes to the colorectal cancer.

Mortality is not the only major concern which lies with this disease, patients' financial condition and emotionally patient is impacted. Surgery, chemotherapy, radiation-therapy and monitoring for long term is done in treatment of colorectal cancer and it leads to stress and financial costs in huge amount. Increase risk of mortality is added to lack of healthcare infrastructure in adequate manner and it is generally seen in patients belonging to low- and middle-class families.

There is gradual increase in number of patients of colorectal cancer in India over these years. If we compare the overall incidence with western countries main reason for the increased risk in this disease is due to life style pattern, dietary changes and also urbanization which is rapidly increasing these days is also one of the main reasons for increasing risk of disease. Regular screening may reduce the chance of mortality rate in India but due to lack of awareness and no screening program at regular interval patients generally is diagnosed at late stage and eventually treatment and chance of survival is at high risk.

Worldwide burden of colorectal cancer can be reduced if diagnosed at early stage. Early disease identification and prognosis of patient can occur due to advancements in computational approaches for diagnosis and other facilities such as dietary modifications, campaigns generating awareness are also helpful. There are various invasive methods to treat colorectal cancer but it is costly, time consuming so diagnosis with the help of machine learning and bioinformatics have been proved to be promising and also gained attention.

2.2.3 Risk Factors and Disease Progression

Association of several factors such as genetics, diet, gut health, environmental conditions lead to the development of colorectal cancer. Abnormal changes occur at regular time interval if closely observed at the linings of the cells in the colon and rectum. A close relationship is interlinked with the way polyps is formed by growth of small non- cancerous tissues and eventually disease occurs. These polyps which are non-cancerous at early stage is converted into cancerous if not treated properly at early stage. Abnormal cellular activity and the accumulations of mutation at genetic level is the reason for the progression of the disease.

Increased risk of colorectal cancer is associated with day-to-day life and diet patterns. Those who consumes alcohol, have diet pattern unhealthy, no or low physical activity, consumptions of food which are highly processed and also does smoking are prone to risk of disease due to their habits. Negative impact of diet which are lower in fibre and rich in fats affect the gut health and promote inflammation[6], [10], [17]. Excess body fat also reduces the metabolic pathways and increases the risk of development of tumor.

Genetic disorders and history of family add to the chance of risk in development of colorectal cancer. Researchers found the pattern associated with development is that age may be the reason for colorectal cancer but in recent studied individuals at young age are also prone to colorectal cancer. This shift in youngster occurs due to western life style, poor diet, consumption of alcohol, smoking and thus composition of gut microbiome is altered. Disease risk increases with hereditary conditions such as DNA damage and oxidative stress.

Human gut microbe is the key to be focused on for understanding colorectal disease and it is studied generally in recent years. The balance of immune system, maintaining digestion, metabolism of the body is maintained by diverse microorganism present in gut microbiome. Dysbiosis, which refers to disturbances in the composition of microorganism is mainly responsible for the development of cancer

and inflammation of intestinal linings[10]. Growth of tumor and inflammation of intestinal cells occur due to harmful toxins and compounds which are inflammatory in nature are produced by microbial species residing in intestinal tract. Differences have been observed in between healthy individuals and patients of colorectal diseases when microbial abundance is compared in between them and conclusion is depicted on the basis of the results that says strong relationship in between the imbalance of gut microbes and progression of colorectal cancer.

There are sequential steps in the progression of colorectal cancer which involves formation of polyps, growth of tumor, invasion of tissues and gradually metastasis occurs to different parts of the body organs. Higher survival rate and treatment is done at better efficacy if diagnosed at early stage. As symptoms related to colorectal cancer is absent at early stage so diagnosis does occur only at later stage. Thus, mechanism of colorectal disease progression and risk factors are to be studied in detail to understand the strategies for prevention, diagnosis at early stage and approaches for treatment.

2.2.4 Limitations of Conventional Diagnostic Methods

Diagnosis of colorectal diseases with the help of techniques such as colonoscopy, biopsy, fecal occult blood testing, sigmoidoscopy, MRI scans. Colonoscopy is said to be most effective among of these procedures as it helps in examination of inner lining of colons and its abnormal tissues and polyps by doctors[6], [9]. Further examination can be followed by biopsy for confirmation. Early-stage detection of the colorectal disease with the help colonoscopy is difficult and this adds to the limitation of colonoscopy.

Invasive procedure involved in colonoscopy causes anxiety, fear, discomfort and it can be considered as the biggest limitations of colonoscopy. Many individuals are not comfortable and find it inconvenient to clean the bowel and follow dietary rules before the beginning of procedure. Complications such as infection, bleeding, perforation of the wall may occur but it is uncommon. Delay of diagnosis and discomfort generally occurs in the patient of colorectal cancer due to which early treatment is difficult.

High cost and facilities of healthcare which is limited adds to the challenges related to diagnosis by conventional method. Advanced medical equipment, specialist who are properly trained and good infrastructures are required for the conventional methods. Thus, it is difficult to maintain such high facilities in rural areas. Limitation in medical facilities and no awareness program by government results in poor screening and ultimately failure of early detection and risking life.

In different conventional methods, one of the methods which is faecal occult blood testing is cost effective as well as less invasive but they are most of the time inaccurate. Failure of early-stage disease identification leads to trust issues when finding the disease at early stage. There are other ways as well such as CT scans and MRI but they fail to diagnose the disease at early stage as they are not detecting abnormalities at microscopic level.

Structural and histological changes are being focused in traditional methods to detect the disease. Early detection of the disease is possible with the help of alteration in microbes and molecular biomarkers which has been studied recently. Thus, stools samples are emerging as a possible invasive approach for the identification of disease by analysing microbial signatures.

Diagnosis of colorectal disease is possible these days due to advancements in computational approaches, machine learning and metagenomics. Efficient analysis of complex microbiome datasets and identification of hidden disease pattern is possible with the help of machine learning algorithms. Early disease prediction is possible due to computational methods as compared to traditional methods where former is non-invasive, at lower cost and scalability which is generally improved. Thus, machine learning integrated with microbiome is generally considered these days for classifying colorectal disease and assessing the risk.

2.3.1 Human Gut microbiome

Gut microbiome consists of various populations of microorganisms and they are highly complex. Number of microorganisms such as bacteria, fungi, viruses and archaea resides naturally in the digestive system. There are various types of microorganisms which resides in gut microbiome which helps to fight and protect the gut from harmful pathogens by maintaining the health of gut and contribution in digestion, metabolism, nutrient absorption, and immune regulation[10].

Factors such as age, diet, lifestyle, environmental exposure all these contributes to the composition of the gut microbiome. Gastrointestinal tract has various regions and different microorganism do resides within it on the basis of availability of oxygen, pH and concentration of nutrients. Among all the microorganisms, bacterial communities are the ones which are studied extensively in the gut microbiome.

Colonization of harmful pathogens is prevented by the gut microbiome microorganisms which are in balanced condition normally. The main role of these microorganisms are maintaining the intestinal

barrier integrity, metabolism of dietary fibres and synthesis of short fatty chain acids. If the balance is disturbed it generally leads to negative harmful impact and disease may occur.

Understanding the human gut microbiome and its connection with human disease is improved due to numerous advanced technologies such as sequencing technologies and metagenomics. It has been found by most of the researchers that altering the diversity of microbes and its abundance are strongly associated with disorders related to gastrointestinal tracts, obesity, inflammation as well as serious diseases such as colorectal cancer. Progression of disease generally occurs by promoting the changes in gut microbiome[9], [10].

During these days, study of gut microbiome is taken seriously as there are various theories which supports initiation of tumour and its progression due to microbial species. If composition of species is compared between healthy individuals and colorectal disease patients' studies suggest differences. There are few microorganisms which produces toxins and inflammatory compounds which results in degradation of intestinal linings. In. the similar manner, there are few microorganisms which can improve intestinal health and suppress progression of disease.

As strong connection has been studied between gut microbiome and colorectal disease so they might act as a disease biomarker. Disease associated microbial patterns are identified with the help of stool-based microbiome analysis in collaboration with computational methods[9].

2.3.2 Gut Microbial Dysbiosis

Gut microbial dysfunction means disturbance in the balance of composition of microorganism and their function compared normally with the microorganisms in gastrointestinal tracts[6], [10]. Proper immune function and intestinal health is well taken care of due to microorganisms present in balanced state and these microorganisms support good gut health in healthy conditions. If the balance is disturbed, this leads to various diseases related to gastrointestinal tracts. Aging, alcohol consumption and various other factors contribute to the dysbiosis.

Colorectal disease progression is closely associated with gut microbiome dysbiosis. Health of an individual deteriorates by the usage of excessive antibiotics, alcohol consumption, stress, obesity and environmental exposure. Those foods which are rich in fibres support beneficial microorganism to grow while fatty diet and processed food tend to the growth of harmful microorganisms. Chronic inflammation and damage of intestine occurs if there is imbalance in microbial composition.

Studies are being done to find the closed relationship between microbial dysbiosis occurring in gut as well as progression of colorectal disease. Abnormal growth of microorganisms led to increase in colorectal cancers so the maintenance of gut microbiome is the key to maintain the gut healthy of the individuals. If continuous inflammation of intestinal does occurs it contributes to damage of DNA and formation of tumour. Reduction in number of bacteria which are beneficial leads to weak intestinal barrier function and regulation of immune system and thus, increases the susceptibility to development of disease.

There has been research which have proved the noticeable differences in the composition of healthy individuals and patients of colorectal disease. It has been noticed that bacterial species who are harmful are present in large amount if an individual suffers from colorectal cancer and healthy microbes' number is reduced.

Detailed study of microbial dysbiosis is possible with the advancements in computational methods and metagenomics. Computational techniques support the development of approaches which may help in analysis of complex microbiome datasets and identify the disease.

2.3.3 Gut Microbiome as Biomarker

Diagnosis and prediction of colorectal disease is possible at early stage where gut microbiome acts as a biomarker and this potential of gut microbiome has gained immense popularity[6], [9]. A biological indicator which helps in the identification of disease at the time of progression of disease is considered as biomarker. As composition of microorganisms present in gut microbiome changes in the comparison of healthy individuals with diseased one, researchers have begun to notice the remarkable changes in the stool sample composition for diagnosing the disease at early stage.

Differences in between the healthy individuals and unhealthy individuals' microbial composition shows clearly the abundance and diversity. Some microorganisms which are of great importance and benefits the gut is reduced in number and those who are harmful, there numbers are increased significantly. Though with complex diagnosis methods such as MRI, colonoscopy disease can be helpful generally at late stage but alterations in microbial compositions can be easily read with the help of computational methods such as machine learning and metagenomics.

Advantages of biomarkers on the basis of microbiome is that it can be detected easily with the help of stool sample analysis in non-invasive manner. In traditional methods where colonoscopy and biopsy are done is invasive as well as painful thus, gut microbiome-based analysis reduces the anxiety and

fear among patients. Microbiome based analysis provides details at molecular level which is not possible with conventional methods.

To study microbial communities with high and greater accuracy in detail is possible only due to advancement of sequencing technologies and metagenomics. Identification of microbial abundance patterns and their variations are generally possible due to high through sequencing methods[20]. Researchers use computational methods with microbial profiles to identify the patients suffering from colorectal cancer easily.

There are various computational methods but machine learning is one of the most important techniques that helps in management of complex datasets. There are patterns associated with gut microbiome which can be analysed easily with the help of algorithms such as Logistic regression, Support vector machine, Random Forest and XGBoost. Diagnostic accuracy can be improved with the help of computational methods and helps in predictive system development.

Identification of personalised therapeutics can be done with the help of microbiome biomarkers when monitored and it also helps in understanding the progression of disease and response of the treatment. Investigation of specific biomarkers is continuously done by researchers to find out the microorganism which may act as a biomarker for colorectal disease classification. Gut microbiome is recognised as one of the best approaches for diagnosis of disease non-invasively, in biomarker discovery and for precision medicine.

2.3.4 Specific Microbial Species associated with colorectal cancer

Specific species of microbes which are present in gut has close relationship with colorectal cancer progression and development. Composition of microorganisms changes while comparing gut of healthy individuals with diseased ones. Presence of certain microorganism leads to inflammation of intestinal lining, production of toxins, damage of DNA, and cellular growth abnormally while some organisms which are present may contribute in maintenance of intestinal health and protect from progression of disease.

Presence of *Fusobacterium nucleatum* is majorly associated with colorectal cancer. The presence of this bacterium is abundant in patients suffering from colorectal cancer and it is noticed that progression of disease occurs due to this bacterium and it also suppresses the immune response. Studies shown by researchers suggests that presence of this bacterium helps in survival of tumour and growth of cancer.

Bacteroides fragilis is one of the important microorganisms associated with colorectal diseases and the strain which is most important is enterotoxigenic strain. The presence of this strain is followed by the inflammation of intestinal lining as well as damage to the colon. Abnormal proliferation of cells and formation of tumor is increased due to inflammation in intestine continuous manner. On the comparison of microorganisms in healthy individuals with diseased patients indicates that these microorganisms are higher in number in diseased individuals.

Polyketide synthase (pks) genes carried by *Escherichia coli* strains have correlation with colorectal cancer and its development. Genotoxins which is a harmful substance is produced by these strains and they might induce DNA damage in cells of colon[8], [11]. Cellular division is uncontrolled and cancer progression continues due to mutations of DNA caused by these strains. Studies demonstrated by researchers on colorectal cancer suggest that these strains are found in patients of colorectal cancer and strains are generally present in tumor.

Peptostreptococcus anaerobius is present in patients of colorectal cancer and these are found in abundance compared to the healthy individuals. Production of reactive oxygen species is increased due to the strains of this bacterium and it leads to damage of DNA and inflammation occurs inside the intestine. Cellular signaling pathways are altered and microorganism support in the growth and proliferation of tumor cells.

Microorganisms present in intestine also support the protection roles such as bacteria *Lactobacillus* and *Bifidobacterium* help in reduction of inflammation and maintenance of balance in intestine. Butyrate which is a short chain fatty acid is produced by the bacteria which helps in maintenance of health of bacteria and also inhibit cancerous cells growth[6], [8]. These microorganisms are reduced in patients of colorectal cancer and it indicates that there is imbalance in gut microbiota which leads in contribution to the development of the disease.

Progression of colorectal cancer and its development is closely associated with specific microbial species present in gut. Inflammation, production of toxin, damage of DNA, proliferation of cells abnormally all are induced by harmful microorganisms and intestinal health is maintained by beneficial microorganisms. Prevention and disease treatment can be done if the disease is diagnosed at early stage and there is a close link between gut microbiota and colorectal cancer[10].

2.3.5 Mechanisms by which Gut Microbiota contributes to Colorectal Cancer

Colorectal cancer is initiated and progressed by multiple mechanisms and it significantly contributes to the colorectal cancer. Gut homeostasis is disturbed by the alteration in the composition of microbes in intestine and these conditions favours development of tumor [10], [11]. Progression of cancer occurs by the interaction of cells of intestine and gut which is done by certain microorganisms present in our gut.

Chronic inflammation is involved as mechanisms for the progression of cancer. Stimulation by harmful microbial species causes inflammation in tissues of intestine and further damages the colon's epithelial lining. Inflammatory cytokines are released by persistent inflammation and cellular proliferation and tumor cells survival is also supported by the mediators. Risk of mutation and colorectal cancer progression is increased by long term inflammations.

DNA damage is another important mechanism which is influenced by metabolites and toxins produced from microbes. Genotoxins and reactive oxygen species is produced by strains of certain bacteria and it generally damages the host cell's DNA. Uncontrolled cell division and tumor formation occurs when DNA mutations is accumulated after the regulation of normal cells. Colorectal carcinogenesis is caused by these microorganisms due to DNA instability.

Disturbing intestinal barrier integrity, producing harmful metabolites and immune regulation is affected by gut microbiota which leads to colorectal cancer. Tumor growth is favoured by suppressing anti-tumor immune responses by few microorganisms. Proliferation of abnormal cell, cellular signalling pathways alteration and oxidative stress is induced by harmful microbial metabolites. Intestinal barrier is disturbed and it allows movement of toxins and bacteria across the intestinal tissues and it damages the cells and leads to its inflammation. Thus, all of these leads to progression of colorectal cancer and its initiation.

2.4 Metagenomics and Microbiome Profiling

2.4.1 Metagenomics

In laboratory conditions microorganism are grown by culturing for study but in the case of metagenomics approach is to study genetic material without culturing microorganisms. In human intestine, there are various microorganisms present and metagenomics helps in studying the functional role as well helps to understand the composition [11]. This is one of the most important techniques as it helps to do research on colorectal cancer as traditional methods does not support growth of microorganism.

Identification of harmful and beneficial microorganisms is allowed by metagenomics analysis in colorectal diseases. Interaction between the microorganism and host is also understood by metagenomics. Researchers analyse microbial genes and it helps researchers to study contribution of microbes in progression of cancer, product of toxin, modulation of immune and inflammation.

Understanding the mechanism of disease and its application in colorectal cancer is one of the applications of metagenomics and early diagnosis and prediction of disease is also easily done with the help of metagenomics.

2.4.2 16S rRNA and Shotgun Sequencing

To study bacterial communities in the gut microbiome 16S rRNA sequencing is used and it is widely accepted methods for the study. Bacterial identification is specifically possible due to presence of variable regions in 16S ribosomal RNA gene, and this technique targets to study this gene[12]. This technique is used to determine microbial diversity in biological samples and it is simple and cost effective.

Limitation of 16S rRNA sequencing includes identification of bacteria and providing information limited to metabolic pathways and functional genes. It also finds difficulty in accurately distinguishing closely related microbial species.

In comparison to 16S rRNA shotgun metagenomics sequencing is better and advanced approach to sequence entire genetic material present in a sample. Functional activities of microorganisms and detailed information regarding types of microorganism such as bacteria, viruses, fungi are provided by the shotgun metagenomics[12]. Development of disease is studied by researchers by analysing the microbial genes, their interaction and metabolic pathways.

For microbiome research, both sequencing approaches are important: 16S rRNA and Shotgun Metagenomics. One of the approach helps in identification of microbiomes and analysis of diversity and another approach helps to understand the function of microbial association with colorectal diseases.

2.4.3 OTU Analysis

In microbial analysis, OTU(Operational Taxonomic Units) groups similar sequences of microbiomes. Classification of microorganism is analysed by OTU on the basis of sequence similarity and microbial abundance is also estimated in it.

Comparison of microbial species in patients of colorectal cancer and healthy individuals is studied in OTU. Identification of diversity of microbial species and their changes can be detected by OTU analysis.

OTU analysis helps in studying alpha diversity and beta diversity. Microbial diversity is measured within a single sample by Alpha diversity, and to compare diversity between different samples Beta diversity is used. Valuable information is provided by the analysis and also information regarding colorectal cancer and their microbial imbalance is studied.

2.4.4 Challenges in Microbial Data

Large diversity and variation of microorganism in biological samples is key reason for complexity in analysis of microbial data. Thousands of microbial features are analysed simultaneously which is one of the most important challenges as it shows high dimensional nature of microbiome data.

Variability in samples among different samples occur due to differences in their lifestyle, age, environmental conditions and diet. Results are affected due to variations in microbial composition and their consistency is reduced.

Accuracy of analysis and machine learning models are influenced by the presence of noise, missing values, and imbalanced microbial abundance. Standardization of sequencing methods and different technique for preprocessing of data remains a difficult challenge.

Advancements in bioinformatics, sequencing technologies and machine learning approaches helps to improve microbiome data analysis and in understanding the gut microbiome and their correlation with colorectal disease.

2.5 Machine Learning in Disease Classification

Analysis of large and complex datasets is possible due to computational approaches such as machine learning. Identification of patterns are not easily done by traditional statistical methods and thus machine learning is chosen to be an important tool in health care and biomedical research[5]. Machine learning is used to differentiate between healthy and diseased individuals on the basis of biological features such as microbial abundance, expression of genes, metabolites profiles and clinical data.

To understand the relationship between gut microbiome and colorectal cancer, machine learning plays an important role. If we compare the composition of microbes in healthy and diseased patients, there

composition differs. Prediction of disease and its classification is analysed effectively with the use of machine learning techniques.

Sequencing technologies generates high dimensional microbiome datasets which is handled easily with the help of machine learning models. Disease progression is identified by features that contributes in progression by learning from previously available data. Diagnostic accuracy is improved by such approaches and this supports planning of treatment, early disease detection and personalised medicine.

Research objective and type of data decide which algorithm can be used in machine learning. Labeled datasets are used in disease classification by supervised learning methods in which condition of disease is already known. Efficiency and predictive performance is achieved by machine learning models such as Logistic Regression, SVM, Random Forest, and XGBoost for disease prediction in microbiomes.

Identification of important microbial biomarkers is achieved by machine learning. Stool microbiome data is used in developing non-invasive diagnostic tools and understanding disease mechanism with the help of these biomarkers. Powerful analytical capabilities are provided by machine learning and prediction of accuracy is depended on factors such as feature selection, dataset size quality of data and validation techniques from model.

Thus, machine learning is an important tool in disease classification and biomedical research. Integration of it with microbiome analysis helps in understanding the complex diseases such as colorectal cancer.

2.5.1 Introduction to Machine learning

Prediction or decisions regarding a data is monitored with the help of machine learning and it is not directly programmed. Identification of pattern and their relationship with large datasets are performed with the help of machine learning. As biological data is usually complex so traditional methods were not beneficial to carry out the analysis over computational approaches.

While studying disease classification, machine learning is used for comparing and differentiating between healthy and non-healthy individuals. The existing datasets are used for training algorithms and finally useful in predicting conditions in diseased patients with accuracy in new samples.

For studying microbiome research, computational approaches such as machine learning is very useful as gut microbiome data is composed of large number of features. Identification of microorganisms

associated with diseases include colorectal cancer and helps in assisting biomarker discovery and early diagnosis.

Unsupervised and Supervised learning are types in which machine learning models are divided. Disease prediction is done by supervised learning as it closely works with labeled datasets where disease categories are already known. Studies related to microbiome-based disease classification uses Algorithm **such as Logistic regression, Support Vector Machine (SVM), Random Forest and XGBoost.**

Machine learning is an important tool used to improve disease prediction, providing support to personalised medicine and helping researchers to understand complex diseases like colorectal cancer.

2.5.2 Supervised Learning

Labeled data is used for training models in supervised learning. Healthy or diseased samples is associated with input data. During the training process, algorithm learns what is the relationship between input features and target output.

To predict condition of disease on the basis of biological data supervised learning is used. To train the model, algorithm learns the relationship healthy individuals and colorectal cancer. Accuracy can be predicted after training the model.

There are different types of algorithms used in biomedical research **such as Logistic Regression, Support Vector Machine, Random Forest and XGBoost.** Identification of microbial features with its association with disease progression and improving performance of diagnostics.

Supervised learning can be used for prediction of disease and identification of biomarker in microbiome.

2.5.3 Logistic Regression

Disease classification is generally performed by Logistic Regression and it is used prediction of binary outcomes such as healthy and non-healthy conditions. Relationship between input features and their probability is analysed by use of algorithm[5]. It is interpreted easily, and performance is efficient on biological datasets.

Microbial patterns are identified by logistic regression and classification of samples are done by model for differentiating in different categories.

Contribution of different types of features for disease prediction is done by Logistic Regression. It helps researchers identify different microbial biomarkers linked with colorectal cancer. Its performance decreases while dealing with high complex and linear datasets.

Its simplicity, reliability, and interpretability is widely studied under Logistic Regression.

16 **2.5.4 Support Vector Machine**

Support vector machine is used for prediction and classification of tasks. Hyperplane works by identifying an optimal boundary. SVM'S main objective is to improve classification accuracy by minimizing the distance points of data[5].

Microbial abundance data is the basis for distinguishing healthy individuals and colorectal cancer patients. Complexity of microbiome datasets is major concern for in microbiome and it does contain large number of datasets. Effective performance helps handling high dimensional data by SVM.

Linear and non-linear datasets is performed with the help of kernel functions which is main advantage of SVM. Easy classification of microbiome datasets are possible due to large datasets and it helps in handling the high dimensional biological data. SVM is most reliable for predicting disease and it avoids overfitting.

Better performance of SVM is generally possible with the help of high computational power and tuning properly with parameters. Accuracy and efficiency is high in SVM so it is considered as one of the most effective parameters in identification and classification of colorectal cancer.

2.5.5 Random Forest

For classification and prediction studies, algorithm of supervised learning is performed. For improvement of accuracy and model reliability, multiple decision trees combination is mostly associated. Combined results of several trees are generated than a single tree.

Microbiome abundance is used for the classification of healthy and diseased individuals. Large and complex datasets are efficiently handled and it helps to identify complex microorganisms associated with colorectal cancer[5].

High accuracy is provided by algorithms and reduction of overfitting is possible. If data is incomplete or noisy, use of algorithm of random forest is important as it identifies the microbial features which

are involved in prediction of disease. These advantages helps in disease classification of microbiome with the help of Random Forest.

2.5.6 XGBoost

XGBoost is widely used for the classification and predicting tasks in supervised machine learning algorithms for prediction and classification[5], [13]. Boosting techniques is the basis on which models which are multiple and weak are combined sequentially and this improves the accuracy of overall prediction.

Microbial and abundance data is used for classification of individuals from healthy and diseased samples. High and large complex datasets are efficiently used which helps in identification of important microbial features which are associated with colorectal cancer.

Most important advantage of XGBoost is its high speed and its ability to reduce overfitting. Missing data and biological datasets which are complex can be easily performed by this algorithm. It has strong predictive classification and it is generally used in microbiome-based diseases.

2.6 Drug Repurposing and Computational Approaches

Drug Repurposing is identification of new uses for drugs which are already existing in market. Developing a new drug is a long and time taking procedure so repurposing of drugs for the treatment of different diseases are approved more easily as compared to the new drugs[4]. Reduction of time, cost and risk involved is less as compared to traditional methods.

2.6.1 Introduction to Drug Repurposing

In this approach drugs that are approved already is identified for treatment of another disease. It has become one of the important strategies as it reduces time, risk as well as cost in comparison to traditional medicine.

Computational and bioinformatics approaches are used for the identification of potential anti-cancer drugs with the help of drug repurposing[4]. As safety profiles and pharmacological informations are already known for existing drugs so it makes conditions suitable. For discovery of potential drugs and improving treatment drug repurposing methods are efficient. Thus, drug repurposing is efficient and effective approach to develop new cancer therapies and in advancement of precision medicine.

2.6.2 Advantages of drug repurposing

Drug Repurposing is advantageous in comparison to traditional methods. Safety profiles, clinical information and pharmacokinetics of exiting drugs are already known so the risk of failure is reduced while developing drugs. This reduces time and it is less costly as compared to development of new drugs all form the beginning.

Many preclinical and safety studies are already completed so it adds advantage in drug repurposing by reducing time and cost. This lowers time duration while testing and developing new drugs.

In drug repurposing, identification of potential therapeutic agents for deadly diseased such as cancer is effective. Computational methods and bioinformatics tools improves drug repurposing by pathways and gene expression analysis for prediction of effective drugs. Drug repurposing is effective for treating disease and development of drugs are accelerated.

2.6.3 Drug Repurposing in Cancer Research

Existing drugs are used as new treatment options for cancer research. There are few drugs which has potential to treat cancer by affecting pathway involved in progression of cancer, growth of cells, and metastasis.

Identification of alternative therapeutic compounds might improve treatment outcomes. Established drugs already have clinical data and its safety is already tested so they reduces time and cost in evaluation than developing a new drug for the treatment of drug[4].

Techniques such as gene expression analysis, analysis of pathway, transcriptomics helps in identification of drugs which can reverse molecular changes associated with drugs. Efficiency of drug discovery is improved and it is supported by the therapies which is personalised and their development plays major roles in drug repurposing.

Anti-cancer drugs and its advancement in treatment of cancer is effective due to drug repurposing and it reduces cost as well as time.

2.6.4 Computational Approaches in Drug Repurposing

Modern drug repurposing studies uses computational methods. For identification of existing drugs various tools such as bioinformatics, biological databases, and machine learning is effective and it is

possible against different diseases. Large scale data are processed quickly and effectively with the help of computational approaches.

Analysis of gene expression profiles, interactions of protein, pathways of molecules and disease associated biomarkers are possible due to development of computational drug repurposing. Comparison of drug-induced pattern of gene expression with therapeutics potential disease related molecular changes is identified.

Computational drug repurposing involves techniques such as transcriptomics analysis, analysis of pathway enrichment, Connectivity Map analysis. Identification of drugs based on relationship between drug-disease relationship is approachable due to mentioned techniques.

Efficiency of identifying potential anti-cancer therapies is possible due to computational approaches.

2.6.5 Transcriptomics and Gene expression analysis

Studying RNA transcripts which is produced by those genes which are under specific biological conditions is Transcriptomics. Gene activity is understood by researchers while studying transcriptomics and helps in identification of gene expression which are closely associated with diseases[15], [16]. To study disease mechanism and to discover potential therapeutics, researchers widely uses gene expression analysis.

To identify abnormal molecular changes, transcriptomics analysis performs comparison between healthy and diseased tissues. Progression of cancer, metastasis and resistance treatment is contributed by above changes.

Microarray and RNA sequencing are the technologies which helps in analysis of gene expression patterns by comparing transcriptomics analysis in cancer research. Drug repurposing is supported by identification of gene expression patterns by prediction of drugs who are capable to reverse disease associated molecular signatures.

For identification and studying therapeutic strategies transcriptomics and gene expression analysis is important.

2.6.6 Differentially Expressed Genes(DEGs)

Significant differences between healthy and diseased samples are evident at expression levels in Differential Expressed Genes(DEGs). One of the most important steps in transcriptomics analysis is identification of DEGs as these are genes which can be associated with development of disease and its progression.

Upregulated and Downregulated genes are identified with the help of DEGs in tumor tissues compared to normal tissues. Biological processes such as cell proliferation, apoptosis, metastasis, and immune response is influenced by altered genes.

Microarray or RNA sequencing technologies generated by bioinformatics tools and statistical methods uses gene expression datasets for identification of DEGs. Pathway analysis, discovery of biomarker, computational drug repurposing uses identified DEGs.

2.6.7 Pathway Enrichment Analysis

Sets of genes which are associated with biological pathways are used for analysis for pathway enrichment. Biological pathways and their molecular mechanisms are understood by researchers on the basis of this analysis.

Cellular pathways significantly affected in tissues of diseased patients are determined by pathway enrichment[21]. Cell cycle regulation, apoptosis, transduction of signals and metastasis in cancer are related to these pathways.

Databases and tools which are commonly used for pathway analysis are KEGG, Gene Ontology(GO) and various other bioinformatics tools. Molecular targets are identified by researchers by studying the results and for understanding biological processes associated with disease.

An important role is played by pathway enrichment analysis in study of transcriptomics and identification of pathways that can be targeted by cancer treatment is also supported by computational drug repurposing.

2.6.8 Connectivity Map (CMap) Analysis

Gene expression patterns is the basis for identification of potential drugs by computational approaches in Connectivity Map Analysis. Disease related molecular targets can be reversed for predicting drug compounds by comparing gene expression signatures with gene expression profiles.

Those drugs which are existing can be identified by CMap which show ant-cancer activity. Those drugs which shows opposite gene expression are considered to be potential drug candidates.

Treatment of cells with different chemical compounds is contained by CMap along with gene expression data. Analysis of these datasets can reveal potential drug-disease relationships and discover drug candidates.

Fast and efficient methods are used for identification of potential drug compounds with the help of computational drug repurposing

2.6.9 LINCS Database

Drug discovery and drug repurposing studies utilize large-scale database for their studies and this database is called Library of Integrated Network-Based Cellular Signatures[15], [16]. After the treatment of different cells of human, gene expression profiles are generated.

Gene expression patterns and their cellular pathways are influenced by different drugs and this is studied by the help of LINCS database. Identification of different potential compounds is done by comparison of disease associated gene expression patterns with LINCS data.

In Ovarian cancer, cumulative effect of tools such as LINCS, CMap and transcriptomics are widely used for prediction of drug-disease relationships and identify drugs which can be useful for disease. Large scale computational methods are supported by database and it improves the study and efficiency of drug repurposing.

Thus, LINCS database can be understood as an important tool in bioinformatics for identification of potential drugs.

2.10.10. Molecular Docking

Studying interaction of ligand and a protein molecule is a technique which is taught in molecular docking. Prediction of binding affinity and their interaction pattern is studied between drug molecule and receptor at molecular level.

Evaluation of potential drug candidate is identified with the help of molecular docking[19]. Effective drugs which can bind to protein are determined through computational drug discovery study. There are various bioinformatics tools which can be used for molecular docking such as Autodock Vina,

EasyDock Vina and DockThor where stable drug-target can be identified. Strong binding interaction is indication for potential drug.

2.10.11. Challenges and limitations of Drug Repurposing

At the time of validation those drugs which were predicted as potential drug candidate might fail during clinical validation. This is one of the major challenges which computational method of drug repurposing faces.

Accuracy of results are dependent on the quality of datasets used along with different models in computational work and bioinformatics tools so incomplete or noisy data may reduce its performance[4], [12].

Though computational approach such as molecular docking may predict possible drug candidate but its validation and clinical studies can confirm the safety and its effectiveness.

Chapter-3

Methodology

The overall thesis design focuses on computational approaches for disease diagnostic and therapeutic discovery. Both the study follows complex raw dataset and further cleaning and preprocessing pipelines are used for the result gain. In this section, the detailed methodology followed by author in each part are explained. The methodology aims for reproducibility of the result obtained in the projects.

Overall study design and workflow This study follows methodology integrating the computational tools with complex and diverse metagenomics data and supervised machine learning based classification. The overall methodology can be explained in simple stages:

1. Gut microbiome data generation
2. Dataset generation and its preprocessing
3. Machine learning model development
4. Output including performance, evaluation and interpretation.

This pipeline aims to generate easy datasets, reproducibility, minimum technical errors results, and our main goal to extract meaningful biological patterns that directly connect to colorectal disease classification. Microbiome data analysis involves steps from high throughput sequenced data of microbiome to derived datasets then perform ML model-based classification to get results and then interpretation.

3.1 Microbiome Data generation and representation

3.2.1 - Metagenomics basis of the dataset

The dataset used in study derived from samples isolated from faecal matter, their DNA extraction, and high throughput sequencing using shotgun sequencing method.

These datasets are directly presented **on the website of NCBI**(<https://www.ncbi.nlm.nih.gov>).

Sequence based microbial analysis could answer two questions-

1. What microbes are present within the sample?
2. What functional role do they play?

3.2.2 Dataset generation and preprocessing

The pipeline includes several key steps including firstly data retrieval of desired raw sequencing reads mostly in FASTAQ format from NCBI website using SRA toolkit. Before any analysis, the quality of

raw sequence reads checked for issue like low quality bases or adapter contamination using FastQ. Raw reads needs to clean by trimming low quality sections and adapter sequence that are compromising with quality of raw data using FASTp[22]. Samples are aligned to the host genome and only unmapped, non-host reads are kept; this refers to host removal using Bowtie2[23]. Taxonomic classification using Kraken2 by comparing reads to a vast dataset of known microbial genomes. This defines the exact microbes present in the sample. To generate the final OTU dataset using this pipeline involves last two steps of abundance estimation using Bracken to estimate the relative abundance of species size more accurately, and in last visualisation using Pavian to get results in charts and plots form to explore complex data, shankey plot may be output of this. Preprocessing of bioprojects that were targeted in this study were PRJNA1237248 and PRJEB6070. Microbial sequenced analysis data refined with health status includes healthy as 0 and colorectal diseases as 1 [24].

TABLE 1: OVERVIEW OF DATASET AFTER PROCESSING

<i>Raw OTU Table</i>	<i>4602 OTUs × 2129 samples</i>
Metadata	2129 samples × 4 metadata fields
Feature Matrix (X _{raw})	2129 samples × 4602 OTU features
Target Vector (y)	2129 labels (Health status)
Class Distribution	Non-Healthy: 1231 Healthy: 898
Prevalence Filter	Applied threshold: ≥20% presence
Features Before Filter	4602 OTU features
Features After Filter	4010 OTU features

CLR Transformation	2129 samples $\times$ 4010 features
Train (80%)/Test Split (20%)	Train: 1703 Test: 426 samples

The OTU dataset had 4602 features that were further filtered out as prevalence filters to 4010 features and after clear noise using CLR transformation that defines compositional nature of microbiome data[25]. CLR transformation remove sample size less than or equal than 20% that gives feature matrix to 2129*4010 OTUs. The class distribution clearly defines 1231 as non- healthy and 898 as healthy.

3.3 ML model development

The complex and diverse dataset that needs a machine learning based framework to analyse and correlate with colorectal disease in front of the ROC curve and confusion matrix. 80% of processed and filtered out dataset were used to train ML models on 1703 samples and the rest 20% used as samples to get credibility of ML models[5].

3.3.1 Selection of classifiers

Multiple supervised machine learning algorithms were analysed and observed to classify colorectal disease status based on given microbiome analysed features. The selected models consist of linear classifiers with regularization, tree-based models, support vector machines, and boosting based approaches[5], [20], [26]. These were chosen due to their widespread uses in microbiome research and ability to handle high dimensional feature spaces. Each model was trained using the same pre-processed feature matrix to give consistent comparison across models.

3.3.2 Model evaluation strategy

Standard classification metrics including accuracy, precision, recall and F1 score were used to evaluate model performance and these metrics impacts in a balanced assessment across both classes and status of model performance. A further examination of misclassification patterns that focused on to give false negative predictions, a critical task in disease classification, was analysed through confusion matrix.

This is the methodology followed for the study of Ubiquitin-Associated Transcriptomic Drug Repurposing. This aims to follow best suitable pathways and software that could give the best analysis and best result.

3.4 Transcriptomic Dataset Collection

Publicly available ovarian cancer gene expression datasets were directly retrieved from **Gene Expression Omnibus (GEO)** database[18]. The transcriptomic expression dataset GEO6008 consists of 103 samples containing all expressed genes in ovarian cancer with corresponding 4 control samples. This dataset consists of diverse histological subtypes allowing differential expression analysis between tumor and normal ovarian samples.



3.4.1 Differential Gene Expression Analysis

Differential gene expression analysis is performed using GEO2R, an online open-source platform provided by the National Centre for Biotechnology Information (NCBI). GEO2R applies the Limma (Linear Models for Microarray Data) statistical framework to analyse differentially expressed genes with diseased and control group samples[16], [18].

Based on log fold change values and adjusted p-values, these analyses are statistically classified in upregulated and downregulated [15]. Upregulated genes were selected for downstream analysis, as they are more likely to represent oncogenic drivers of ovarian cancer.

3.4.2 Gene Set Enrichment and Drug Prediction Analysis

The upregulated genes were further processed for enrichment using the online tool Enrichr, a gene set enrichment analysis platform[21]. It is used to identify biological pathways, functional annotations, and drug perturbation signatures associated with the selected gene set. Drug predictions were performed using the **LINCS L1000 Chem Perturbation Consensus Signature (CPCS)** database, which identifies the pre-existing small molecule that could reverse the disease associated with transcriptional signatures. [16] High enrichment scorer drugs were selected for further screening.

3.4.3 Drug Repurposing Strategy and Candidate Selection

The shortlisted drugs selected from LINCS1000 analysis are already approved for some other disease earlier and well-studied. Compounds with unclear chemical identity and redundancy were removed and top ranked compounds were selected for further pharmacokinetic evaluation.

3.4.5 Pharmacokinetic and Drug-Likeness Screening

Pharmacokinetics and Drug-Likeness performed using an online tool SwissADME[27]. Screening performance was based on Lipinski's Rule, PAINS alerts, Brenk, BBB permeability. Compounds meet basic drug-likeness were selected and further docking performed.

3.4.6 Protein and Ligand Preparation

The 3D structure of EPCAM protein was prepared for docking by removing water molecules, adding polar hydrogen atoms, and assigning appropriate atomic charges[28]. Selected ligand's

structure retrieved from public chemical databases and prepared by optimizing geometry, adding hydrogens, prepared by optimizing geometry, adding hydrogens[29]. All proteins and ligands were converted into PDBQT format before docking. Costunolide was used as a reference ligand[14], [19].

6 3.4.7 Molecular Docking Analysis

Docking was performed by using AutoDock Vina to predict the binding affinity of each ligand with protein. A grid box was defined to know the active and binding regions of EPCAM. Docking simulations performed using default Vina parameters. Ligands with higher binding affinity than control were selected for further visualisation.

3.4.8 Visualization and Interaction Analysis

Visualized docked protein-ligand complexes using PyMOL to analyse binding orientation and pocket occupancy. Two dimensional interactions representing hydrogen bonds and hydrophobic interactions were generated using Biovia. Key interacting residues involved in ligand stabilisation within EPCAM binding pocket were identified and compared with different compounds.

3.4.9 Comparative Analysis and Candidate Prioritization

Comparative analysis of reference ligand Costunolide and other compounds with respect to binding affinities were performed. Compounds having higher binding affinities and favoured stabilisation profiles were selected as potential inhibitors of EPCAM[30]. Computational predictions identified these as potential inhibitors, providing a basis for future experimental validations.

CHAPTER-4

RESULT AND DISCUSSION

The thesis contains two important computational research studies of gut microbiome based colorectal disease classification using machine learning and the other one is transcriptomics driven drug repurposing for ovarian cancer. The studies show some highly important parts of bioinformatics, artificial intelligence, computational biology and precision medicine.

The gut microbiome study focused on stool microbiome samples containing datasets were collected to identify microbial patterns linked with colorectal disease. Preprocessing includes data cleaning, prevalence filtering and CLR transformation. The final processed dataset contains 2129 samples and 4010 OTU features. The six machine learning models are used for modelling. These are Logistic regression+SVM(RBF), XGBoost +SMOTE, LightGBM+SMOTE, Tuned SVM, Random Forest and KNN. Logistics regression+SVM(RBF) showed best performance with accuracy achieved 80.7%. The ROC-AUC score achieved 0.892 and the model shows balanced precision recall and F1 score.

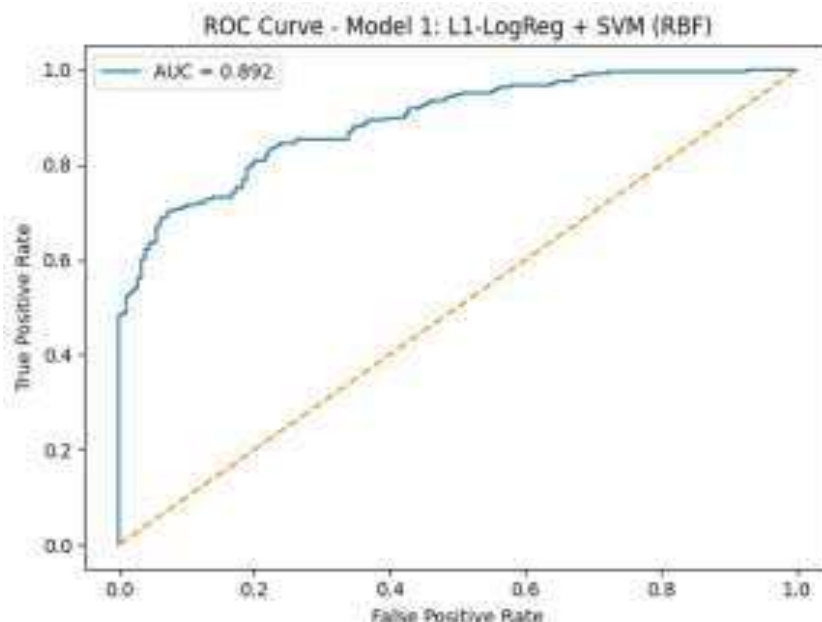


Figure 1- ROC curve for Logistic regression +SVM(RBF) model, represents discrimination between healthy and diseased sample with AUC of 0.892.

The low false negative predictions through confusion matrix observed signifies effective disease classification ability. Low false negative prediction is important because missed disease cases delay treatment and increase disease severity[5].

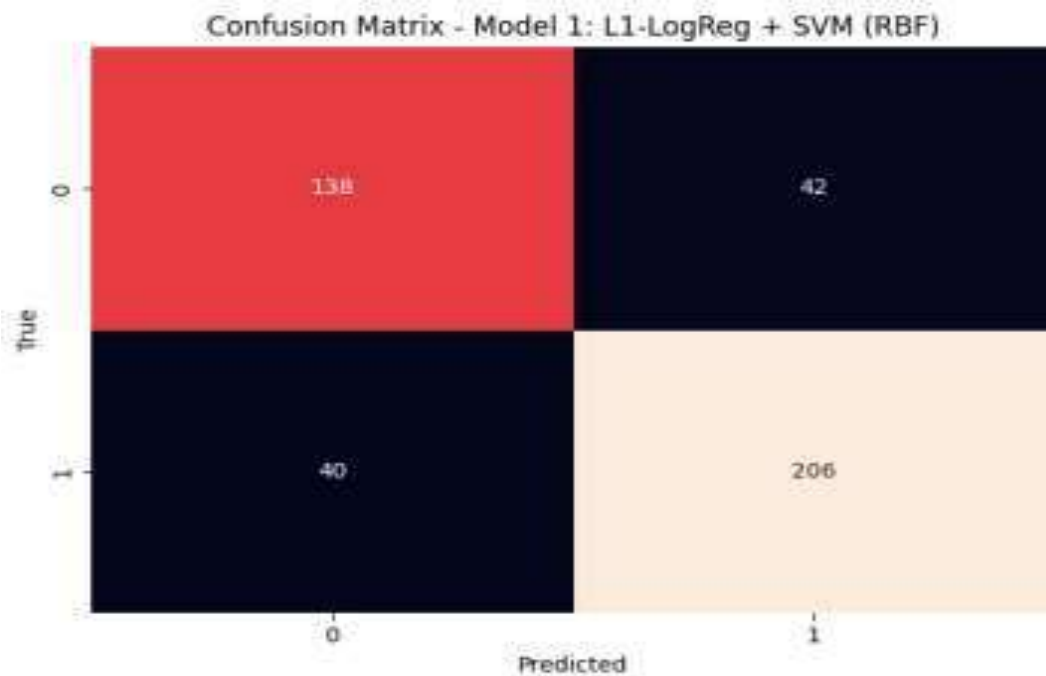
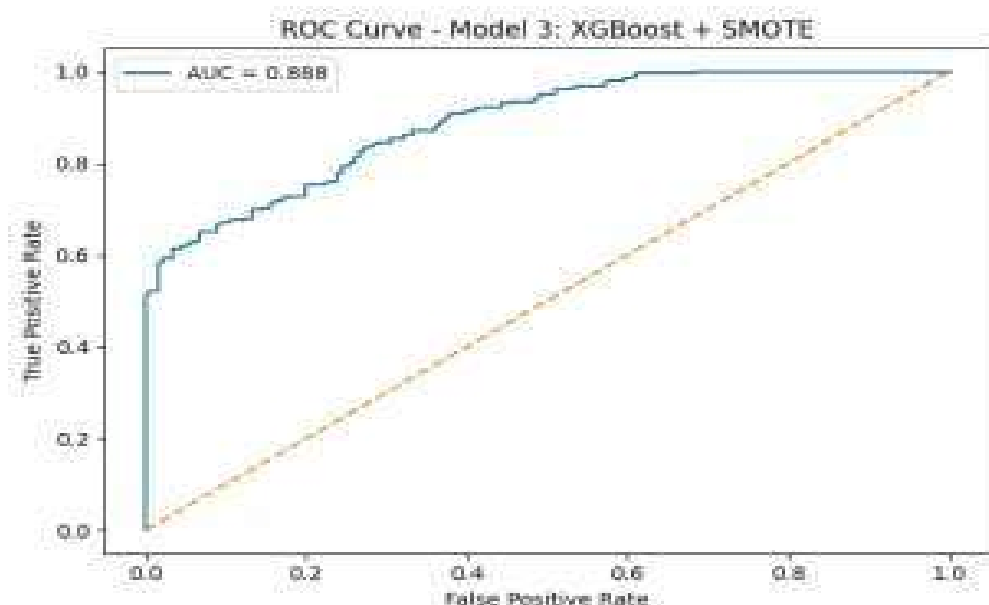
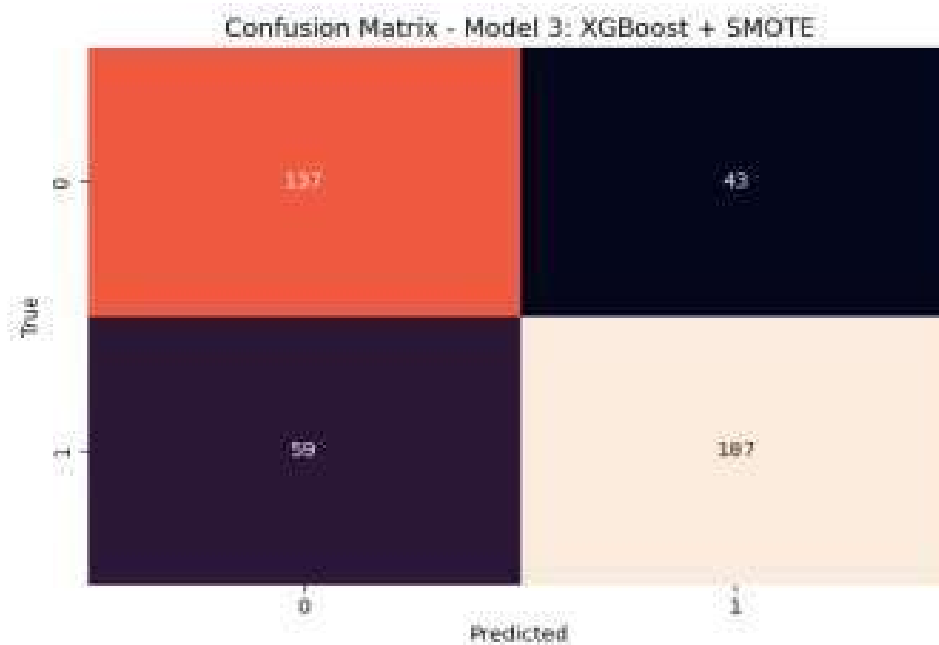


Figure. 2- Error interpretation of dataset through confusion matrix showing 206 as TP, 138 as TN, 42 as FP and 40 as FN.

XGBoost+SMOTE showed strong classification performance and are effective for high dimensional microbiome data.



18 Figure.3- ROC curve for the XGBoost +SMOTE model demonstrating classification performance on CLR- transformed OTU profiles



11 Figure 4- Confusion matrix for XGBoost +SMOTE showing True positive, false positive, True negative and false negative.

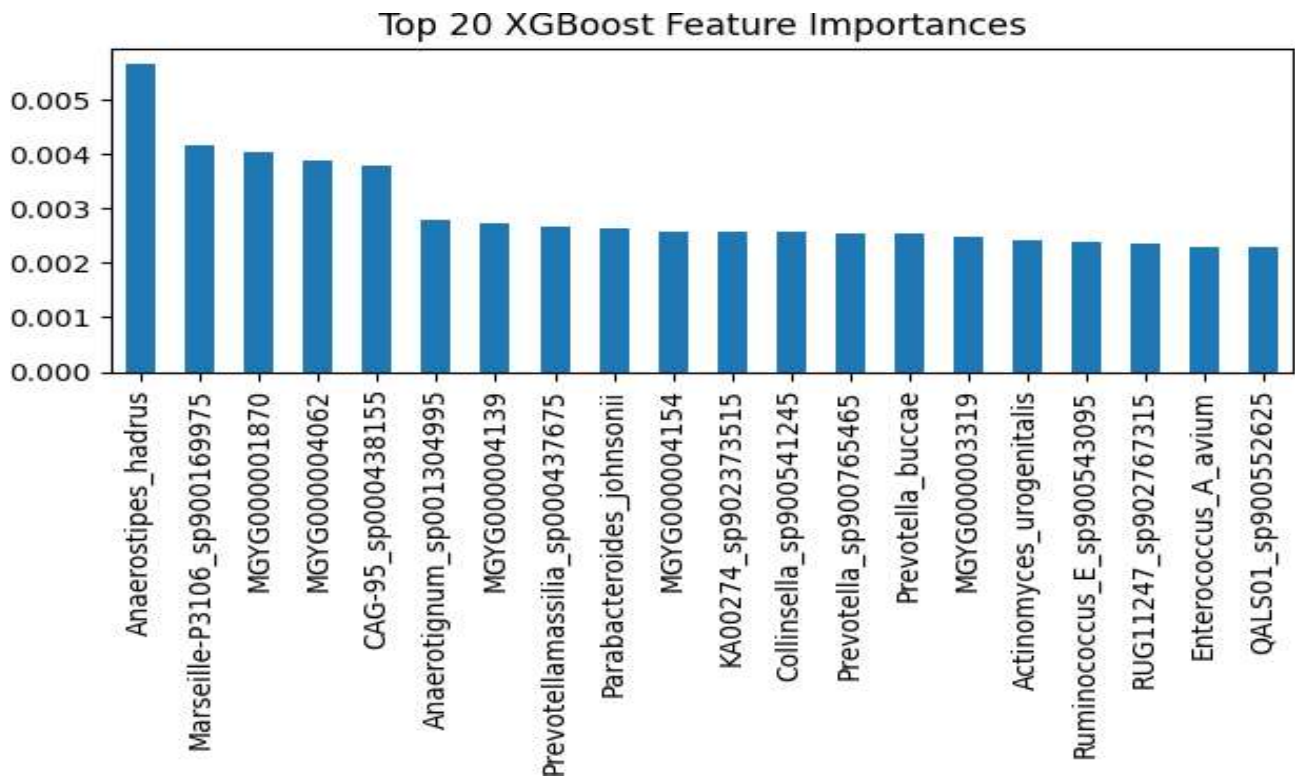


Figure 5- Top 20 XGBoost Feature importances

LightGBM+SMOTE are good for classification stability and handle sparse OTU features efficiently.

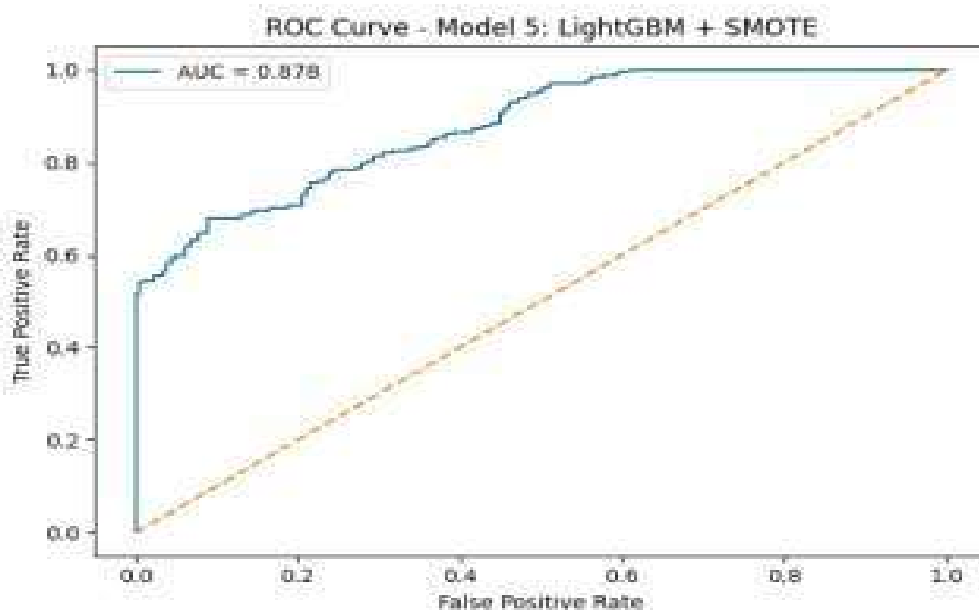


Figure. 6- ROC curve for LightGBM+SMOTE model

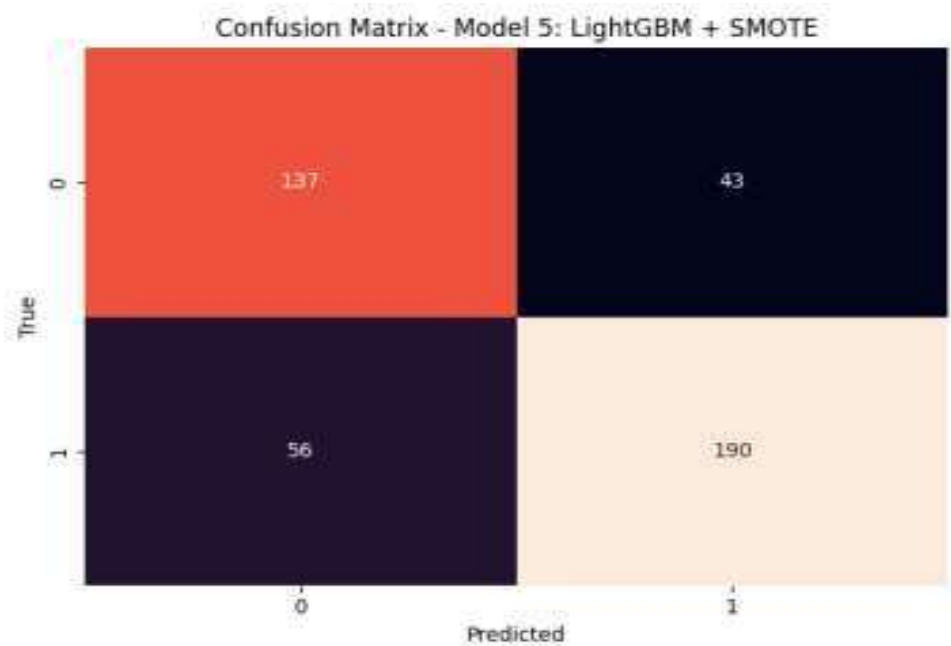


Figure 7- Confusion matrix for error interpretation

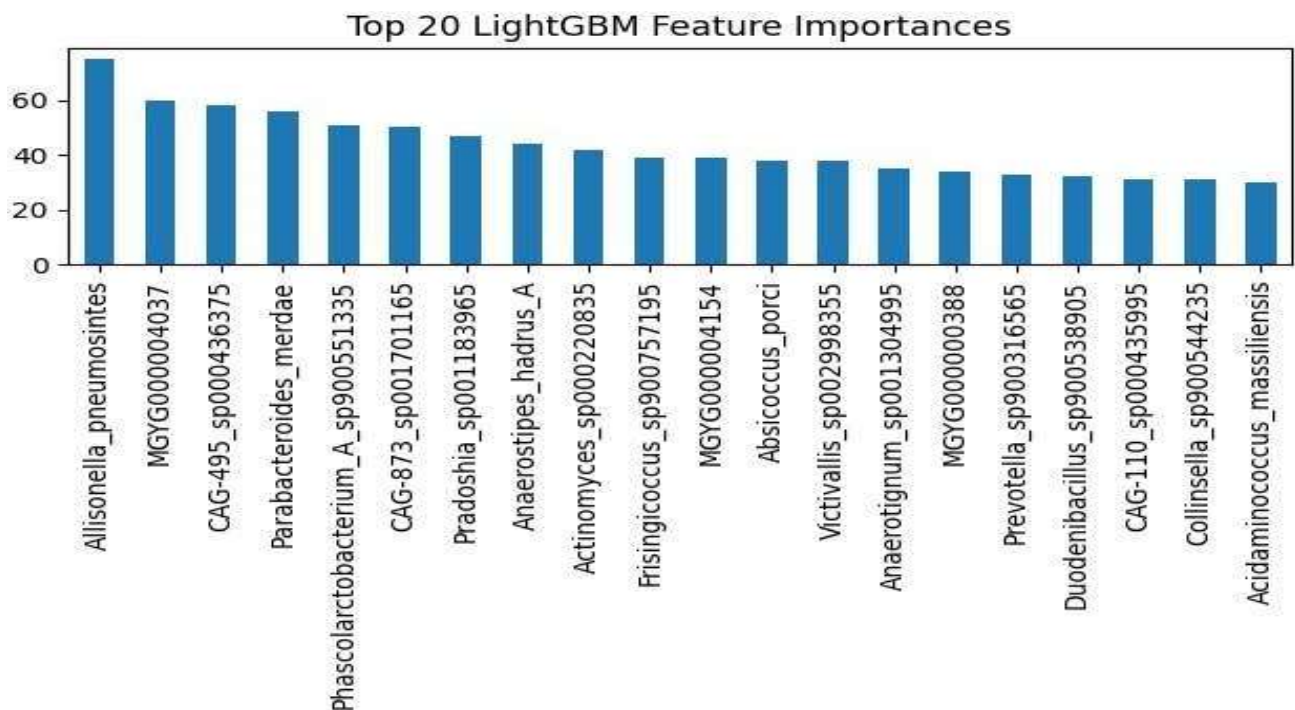


Figure 8- Top 20 LightGBM feature importances

Tuned SVM showed competitive performance and good margin optimization.

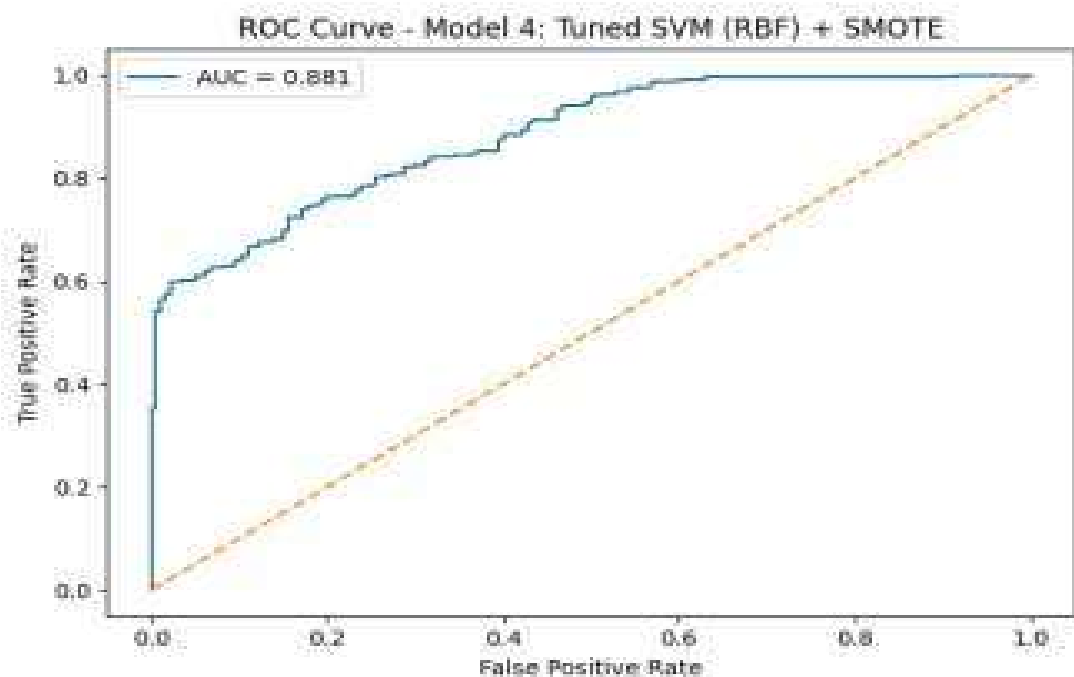


Figure 9- ROC curve for the Tuned SVM(RBF)+SMOTE model

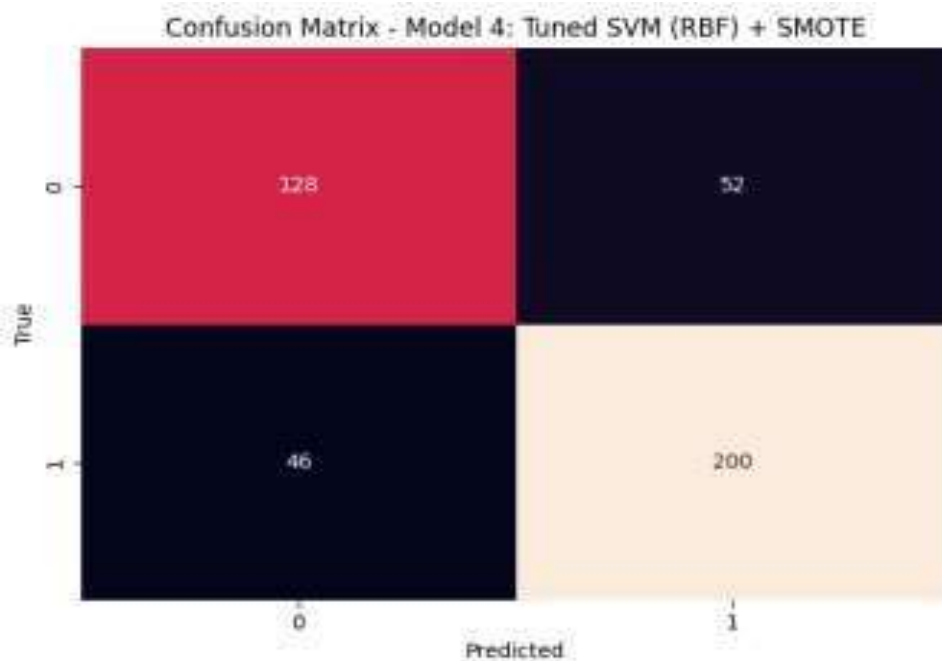


Figure 10- Confusion matrix showing TP, FP, TN, and FN samples.

Random forest is useful for feature important analysis and showed moderate classification performance.

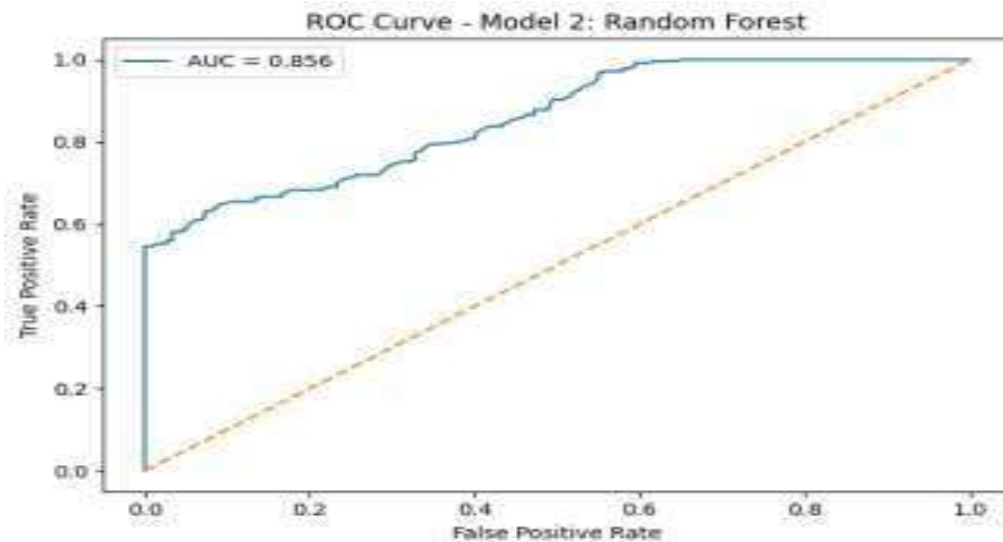


Figure.11- ROC curve for the Random Forest model showing AUC of 0.856. Evaluated on Microbiome OTU features for healthy and non-healthy samples.

The KNN model showed lowest accuracy among all other models and distance-based algorithm is less effective for sparse microbiome datasets.

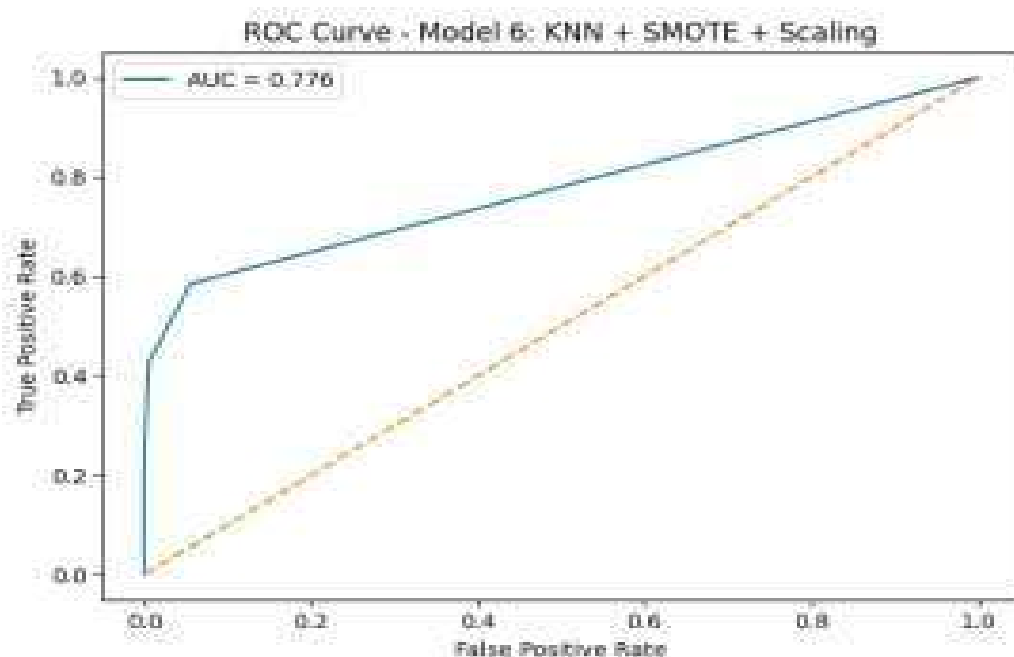


Figure 12- ROC curve for KNN+SMOTE+Scaling model

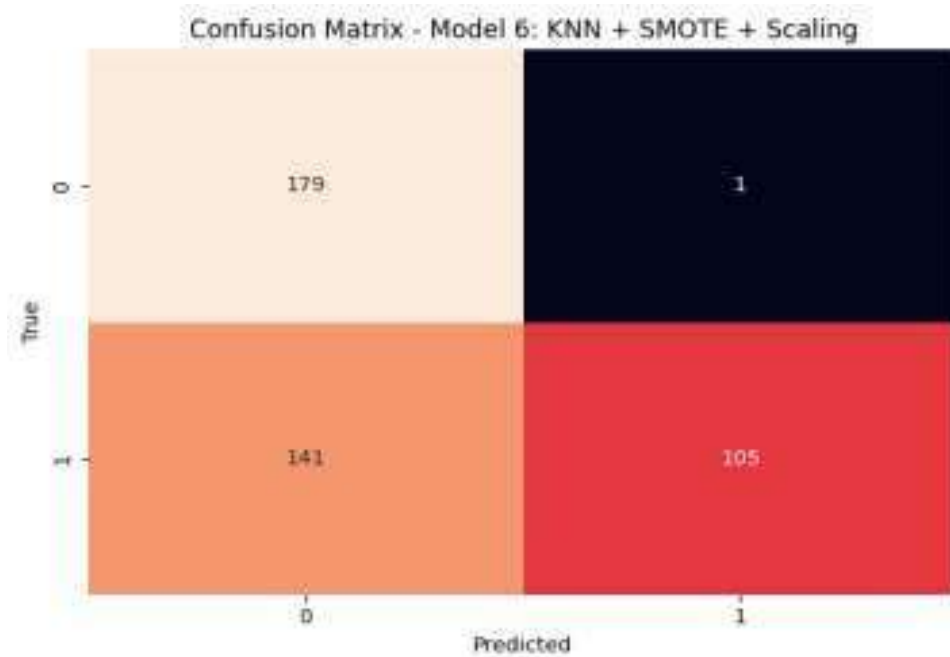


Figure 15- Confusion matrix representing true positive, true negative, false positive and false negative for KNN+SMOTE+Scaling model

TABLE 2: COMPARATIVE ANALYSIS OF ALL SIX MACHINE LEARNING MODELS

<i>ML Models</i>	<i>Accuracy (%)</i>	<i>Precision (Weighted)</i>	<i>Recall (Weighted)</i>	<i>F1-Score (Weighted)</i>
Logistic Regression +SVM(RBF)	80.7	0.81	0.81	0.81
Random forest	72.3	0.72	0.72	0.72

XGBoost+ SMOTE	76.05	0.76	0.72	0.72
Tuned SVM(RBF)+ SMOTE	76.9	0.77	0.77	0.77
LightGBM+ SMOTE	76.76	0.77	0.77	0.77
KNN+SMOTE+ Scaling	66.66	0.81	0.67	0.65

There is important biological significance from these results that support gut microbiome as potential biomarker for disease classification as diseased individuals showed microbial dysbiosis. The possible biological associations may include inflammation, immune suppression and tumor promoting microbial activity.

The result of this study signifies a potential approach for non-invasive diagnostic systems and microbiome-based screening tools. It may reduce dependency on invasive methods like colonoscopy.

The transcriptomics drug repurposing study focused on therapeutic research on ovarian cancer. The target protein used is EPCAM. The transcriptome dataset on ovarian cancer analysed using GEO2R. The analysis identified significantly upregulated genes in ovarian cancer and upregulated genes selected for further study.

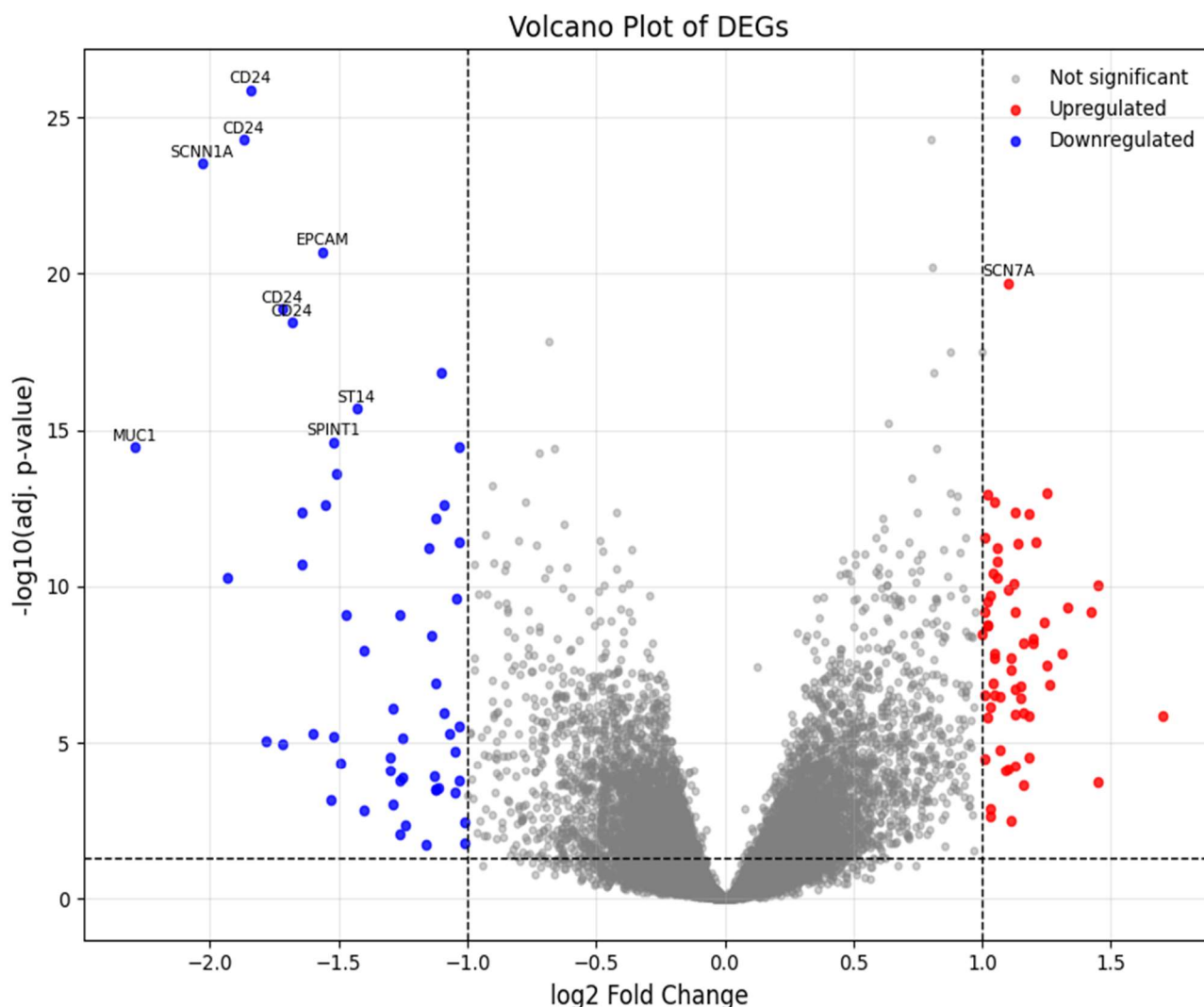


Fig. 1. Volcano plot of differentially expressed genes (DEGs) representing comparing between ovarian cancer and normal ovarian tissue to identify biological relevant and statistically significant

The downstream computational analysis includes Enrichr analysis, LINCS L1000 analysis and drug screening. The shortlisted compounds analysed using SwissADME and Molecular docking.

TABLE 1. PHARMACOKINETIC AND DRUG-LIKENESS PROPERTIES OF SHORTLISTED COMPOUNDS PREDICTED USING SWISSADME

Compound Name	BBB permeability	Lipinski rule	PAINS alert	Brenk
Regadenoson	No	Yes	No	0
Cicloprofen	Yes	Yes	No	0
AT-7519	No	Yes	No	0
Acronin	Yes	Yes	No	1
SYK inhibitor	No	Yes	No	0

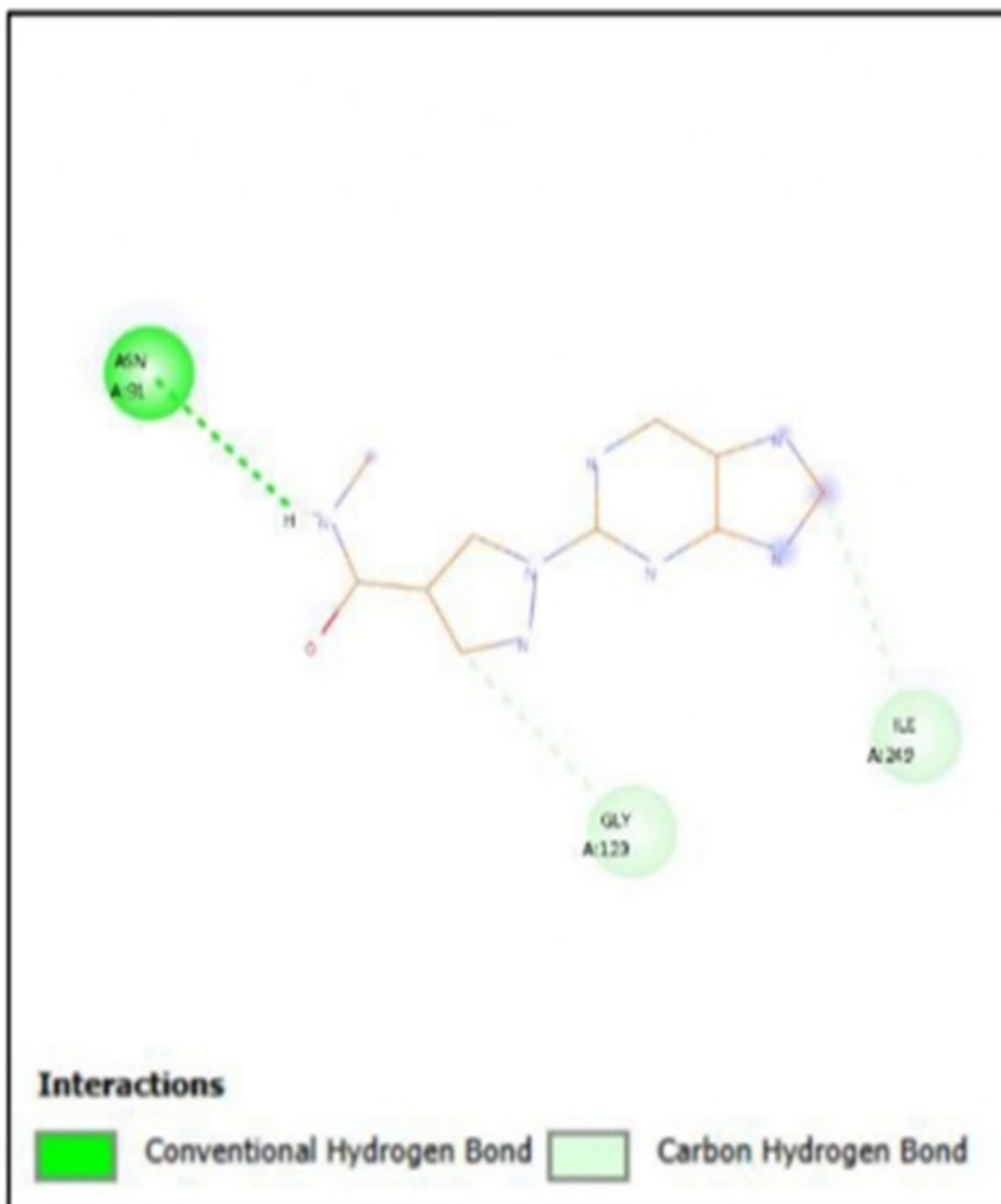
AT-7519, found to be the best drug candidate for EPCAM inhibition, showed strongest binding affinity with a docking score of -8.288 kcal/mol.

TABLE 2. MOLECULAR DOCKING RESULT SHOWING BINDING AFFINITY AND KEY INTERACTING RESIDUES OF SHORTLISTED COMPOUNDS AGAINST EPCAM

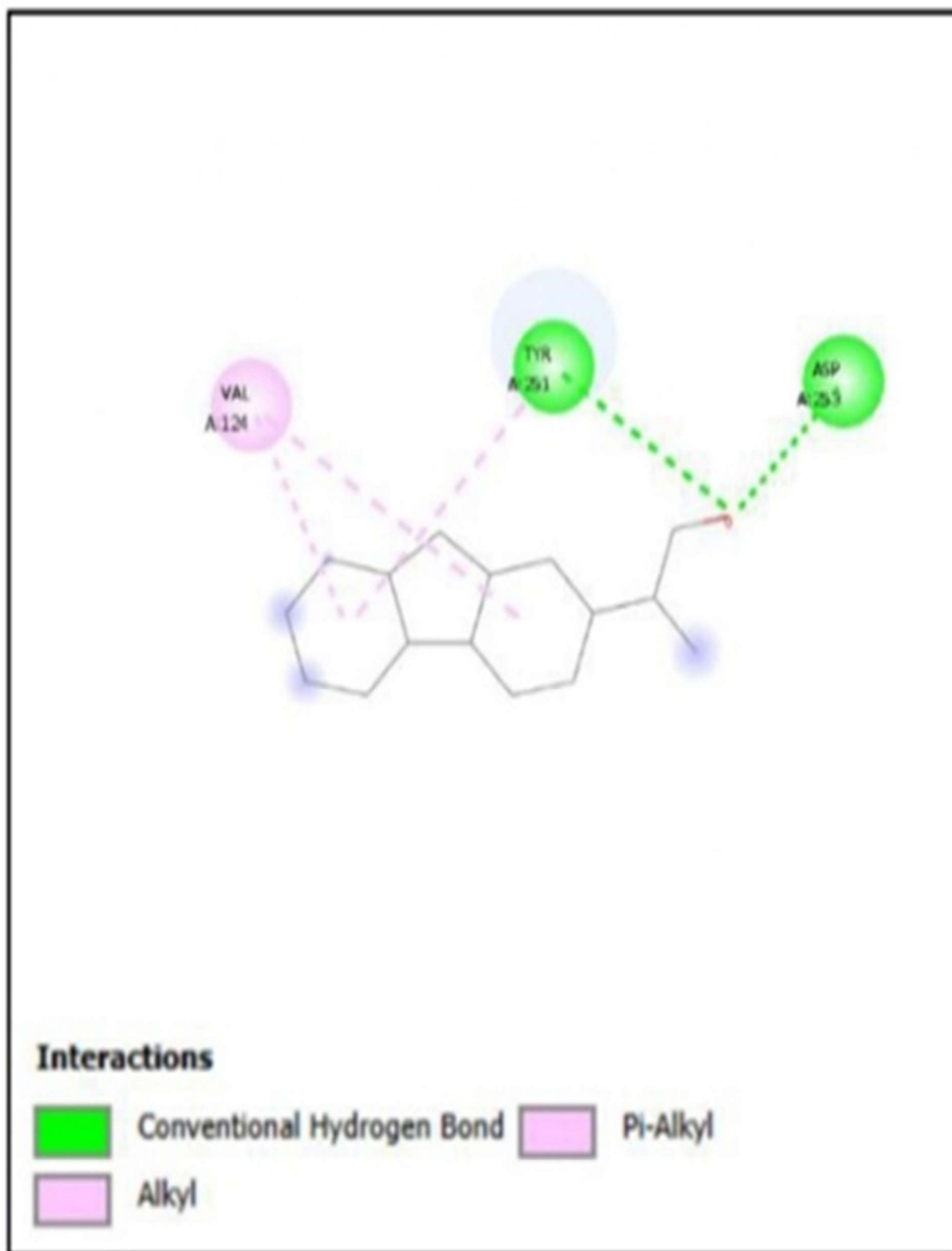
Compound Name	Binding affinity (kcal/mol)	Key interacting residues
Costunolide (control)	-7.394	Tyr 251 Pro 257 Lys 265
Regadenoson	-7.883	Asn 91 Gly 123 Ile 249
Cicloprofen	-7.628	Val 124 Tyr 251 Asp 253
AT-7519	-8.288	Tyr 251 Val 124 Leu 264
SYK inhibitor	-7.302	Tyr 251 Val 124 Gly 123 Gly 93
Acronin	-7.406	Lys 221 Glu 217 Ile 249

This shows better binding affinity than reference inhibitor Costunolide. The stable interactions observed between AT-7519 and EPCAM active site.

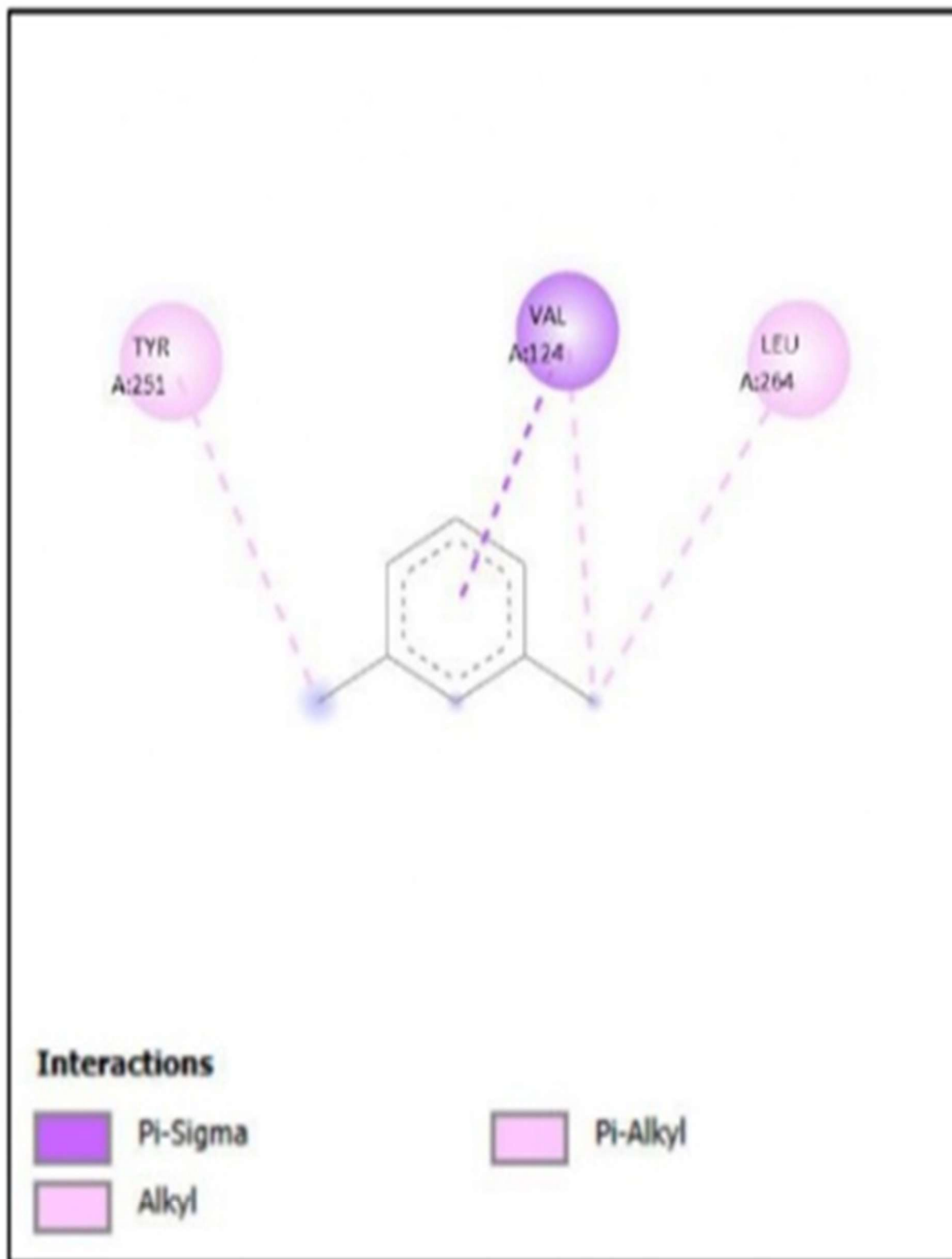
The important interacting residues are TYR251, VAL124, ASN91 and GLY123. AT-7519 could function as an EPCAM inhibitor and be repurposed for potential suppression of oncogenic signalling pathways and tumor progression mechanisms.



(a) Interaction of AT-7519



(b) Interaction of Regadenoson



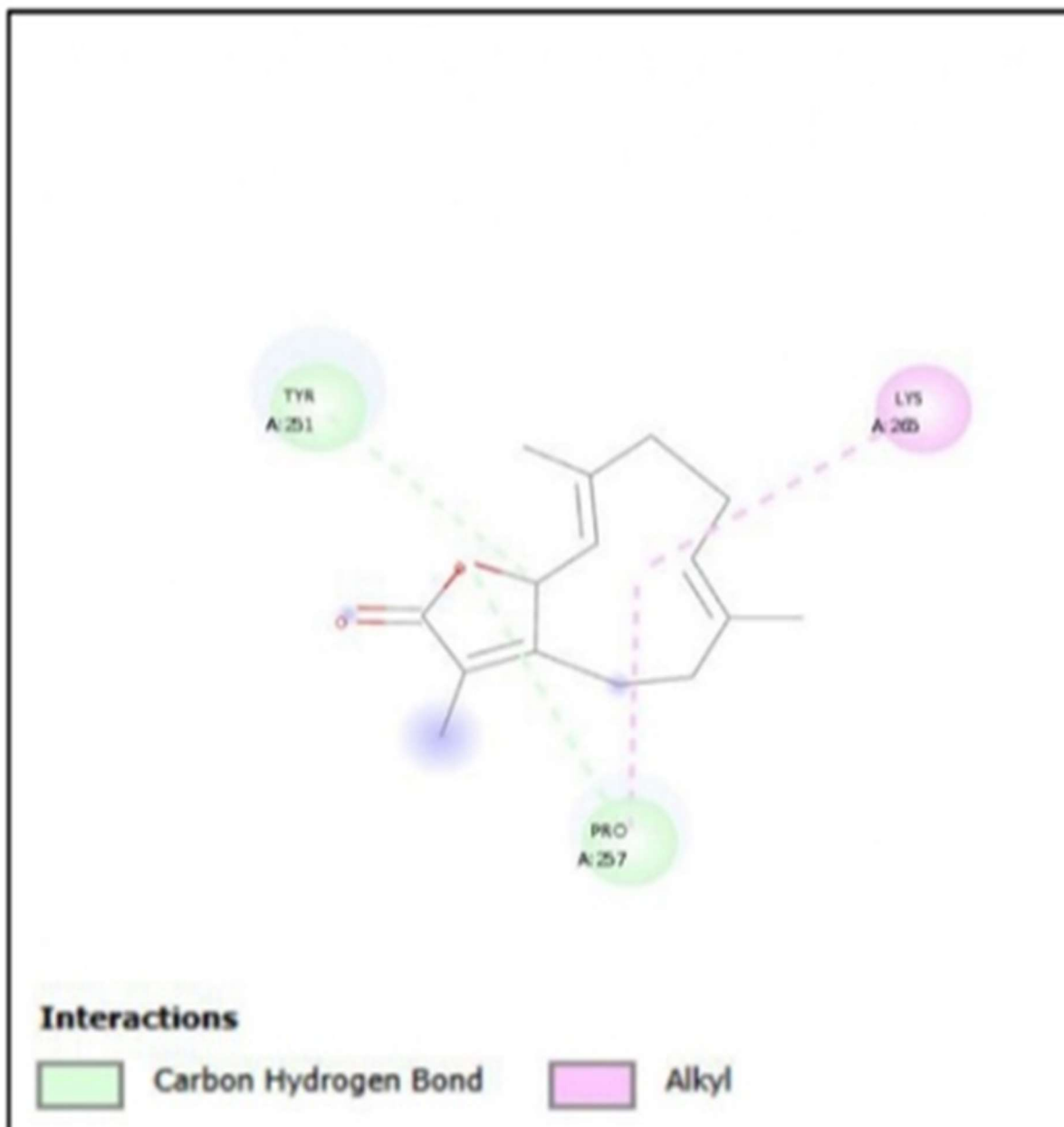
(c) Interaction of Cicloprofen



(d) Interaction of SYK inhibitor



(e) Interaction of Acronin



(f) Interaction of Costunolide (reference control).

Fig.3: 2D protein ligand interaction diagrams illustrating binding interaction between EPCAM and shortlisted compound (a) AT-7519 (b) Regadonosan (c) Cicloprofan (d) SYK inhibitor (e) Acronin and (f) Costunolide (reference control). Green dashed line indicates conventional hydrogen bonds, while pink and purple dashed line represents hydrophobic (alkyl, pi-alkyl, pi-sigma) interactions. Key EPCAM residues involved in ligand stabilization are highlighted



Fig. 2. Three-dimensional ribbon representation of the EPCAM protein structure showing α -helices (red), β -sheets (cyan), and loop regions (green). The structural architecture highlights the binding cavity utilized for molecular docking analysis of shortlisted compounds

Drug repurposing approach is faster than conventional drug discovery and reduces research cost, trial burden and time. Both studies demonstrate integration of bioinformatics, omics technologies, computational analysis and machine learning[3], [4]. Both studies impact the advancement of AI-driven biomedical research, personalized healthcare, early detection and precision medicine.

CHAPTER-5

FUTURE SCOPE

Both the studies demonstrate potential of computational biotechnology, machine learning, metagenomics, transcriptomics and Bioinformatics. These results have potential but further improvements are possible. The future work could possibly improve biological relevance, predictive accuracy and clinical applicability.

The future scope in gut microbiome- based colorectal disease could include dataset expansion. The increased microbiome dataset size could include samples from different geographical regions, dietary habits and different age groups. The larger dataset may improve model generalization and diagnostic reliability. Combining different microbiome data that are metabolomics, proteomics and meta transcriptomics. It will help in understanding host-microbe interactions. Disease-associated microbial functions and molecular mechanisms could also help in understanding by integrating multi-omics.

The model development uses only six models that could be modified using advanced AI and deep learning models. The advanced model could be CNN used mainly for pattern recognition, pattern recognition and used to find hidden local relationships in large datasets.

It was originally developed from image analysis but now helpful in genomics, metagenomics and biological sequence analysis. CNN could be helpful in improving the accuracy and handling the complexity of gut microbiome dataset. It could reduce dependency on manual handling of datasets. It can automatically detect microbial features and patterns that are important. Artificial neural networks copy the structural behaviour of the human brain using input, hidden and output layers. This could be helpful in analysing the non-linear interactions between OTUs. It will continuously improve the prediction using the training dataset. It could be helpful in biomarker identification and handling in noisy biological datasets. Transformer based architectures are advanced deep learning models that could be helpful in understanding sequential data analysis, attention-based learning and relationships among larger datasets. As the large number of factors like OTUs influence disease and long-range relationships are difficult in traditional ml models.

The diagnostic applications could include real time microbiome diagnostic systems and non-invasive colorectal disease scanning systems. This would be helpful in reducing the dependency on invasive procedures this would be helpful in many ways of cost and time. This study has border application in

identifying disease specific biomarkers for early detection. Future applications are personalised medicine, targeted therapies and probiotic based intervention.

The study of transcriptomic drug repurposing requires further experimental validation. The best drug found to be tested in clinical trials. Both in vitro and in vivo studies are the scope for experimental validation that could include the side effects, working effect and microenvironment of EPCAM in inhibition. Before the experimental validation, molecular dynamics simulation study analysis could be the strong evidence for stability of ligand-protein interaction. The study only includes the upregulated genes and shows good results using both downregulated and upregulated genes could give better finding and large number of drug candidates for repurposing and could be helpful in targeting multi target drug discovery. This approach should also be included that could help in exploring multi-oncogenic pathways and multicancer- related proteins. The use of found drug AT-7519 with existing chemotherapy drugs analysis could give better results in ovarian cancer management.

AI and system biology integration could future show a pathway for directional biomedical research and open path for multi omics analysis. The computational platform for the drug repurposing may help in reducing the drug development time by reducing time for clinical trials as earlier it was done. It will also support lower research cost and precision medicine.

Both thesis methodology can support translational life sciences research, personalized healthcare and AI-driven biomedical innovations.

CHAPTER-6

CONCLUSION AND SOCIAL IMPACT

Computational and bioinformatics approaches such as metagenomics, transcriptomics, machine learning are used for diagnosing of disease and drug discovery. There are two challenges studied related to healthcare: Gut microbiome data is used for classification of colorectal cancer while drug repurposing in ovarian cancer is studied on the basis of transcriptomics analysis. Computational methods are extensively used for drug discovery and precision medicine in cancer research.

Publicly available datasets are used for classification of colorectal diseases in gut microbiome study with the help of machine learning. High quality OTU profiles are obtained by the use of various computational methods and various preprocessing such as SRA toolkit, Kraken2, Bracken2, Bracken, prevalence filtering and CLR transformation. Disease prediction is evaluated by various machine learning models such as Logistic Regression, SVM, Random Forest, XGBoost, LightGBM, and KNN. Best classification performance with high accuracy and ROC-AUC curve is demonstrated in Logistic Regression combined with SVM(RBF). Differentiation of healthy and diseased individuals are possible non-invasively for colorectal disease.

Molecular docking is performed for prediction of potential EPCAM inhibitor for drug repurposing in ovarian cancer[4], [14], [28]. GEO2R and Enrichr were used for performing differential gene expression and pathway enrichment analyses. It was followed by prediction of drug through LINCS L1000 and screening is performed by SwissADME[27]. AutoDock vina identifies AT-7519 as the promising compound for EPCAM protein due to its high binding affinity. Identification of potential drugs can be accelerated with the help of computational methods which is indicated by the results.

Integration of machine learning with gut microbiome offers opportunities for non-invasive treatment while drug repurposing is powerful approach for identification of drug candidates in ovarian cancer.

Contribution of this thesis is in the field of healthcare, disease diagnosis, and computational medicine. Early detection of colorectal cancer and identification of drug candidates for ovarian cancer is the contribution of this study to the society. Colorectal disease classification is non-invasive if machine learning is performed on gut-microbiome. Early detection improves outcome of treatment easily and it reduces anxiety among patients and it is cost effective too. AT-7519 acts as inhibitor EPCAM thus,

computational method reduces the cost and time to test drugs which are already for drug repurposing and it improves diagnosis of disease.

REFERENCES

- [1] D. Hanahan, "Hallmarks of Cancer: New Dimensions," Jan. 01, 2022, *American Association for Cancer Research Inc.* doi: 10.1158/2159-8290.CD-21-1059.
- [2] K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2019," *CA Cancer J. Clin.*, vol. 69, no. 5, pp. 363–385, Sep. 2019, doi: 10.3322/caac.21565.
- [3] E. Hernández-Lemus and S. Ochoa, "Methods for multi-omic data integration in cancer research," 2024, *Frontiers Media SA.* doi: 10.3389/fgene.2024.1425456.
- [4] S. Pushpakom *et al.*, "Drug repurposing: Progress, challenges and recommendations," Dec. 28, 2018, *Nature Publishing Group.* doi: 10.1038/nrd.2018.168.
- [5] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," *PLoS Comput. Biol.*, vol. 12, no. 7, Jul. 2016, doi: 10.1371/journal.pcbi.1004977.
- [6] D. Incognito *et al.*, "Fusobacterium Nucleatum in Colorectal Cancer: Relationship Among Immune Modulation, Potential Biomarkers and Therapeutic Implications," Oct. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI).* doi: 10.3390/ijms26199710.
- [7] A. Tang, S. Woldemariam, J. Roger, and M. Sirota, "Translational Bioinformatics to Enable Precision Medicine for All: Elevating Equity across Molecular, Clinical, and Digital Realms," *Yearb. Med. Inform.*, vol. 31, no. 1, pp. 106–115, 2022, doi: 10.1055/s-0042-1742513.
- [8] N. Imanbayev *et al.*, "Colorectal cancer and microbiota: systematic review," 2024, *Termedia Publishing House Ltd.* doi: 10.5114/pg.2024.136228.
- [9] A. Kasagga *et al.*, "Gut Microbiota and Colorectal Cancer: An Umbrella Review of Methodological Trends and Clinical Correlations," *Cureus*, Feb. 2024, doi: 10.7759/cureus.54210.
- [10] Y. Cheng, Z. Ling, and L. Li, "The Intestinal Microbiota and Colorectal Cancer," Nov. 30, 2020, *Frontiers Media S.A.* doi: 10.3389/fimmu.2020.615056.
- [11] J. D. Dahmus, D. L. Kotler, D. M. Kastenber, and C. A. Kistler, "The gut microbiome and colorectal cancer: A review of bacterial pathogenesis," Aug. 01, 2018, *AME Publishing Company.* doi: 10.21037/jgo.2018.04.07.
- [12] A. M. Thomas *et al.*, "Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation," *Nat. Med.*, vol. 25, no. 4, pp. 667–678, Apr. 2019, doi: 10.1038/s41591-019-0405-7.
- [13] G. Roy, E. Prifti, E. Belda, and J. D. Zucker, "Deep learning methods in metagenomics: a review," 2024, *Microbiology Society.* doi: 10.1099/mgen.0.001231.
- [14] G. Spizzo *et al.*, "Overexpression of epithelial cell adhesion molecule (Ep-CAM) is an independent prognostic marker for reduced survival of patients with epithelial ovarian cancer," *Gynecol. Oncol.*, vol. 103, no. 2, pp. 483–488, Nov. 2006, doi: 10.1016/j.ygyno.2006.03.035.
- [15] M. E. Ritchie *et al.*, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Jan. 2015, doi: 10.1093/nar/gkv007.
- [16] J. Guarro *et al.*, "Case of Keratitis Caused by An Uncommon Fusarium Species," *J. Clin. Microbiol.*, vol. 41, no. 12, pp. 5823–5826, Dec. 2003, doi: 10.1128/JCM.41.12.5823-5826.2003.

- [17] M. Guelhan Inci, R. Richter, K. Heise, R. Dukatz, H. Woopen, and J. Sehouli, "Patient-Reported Outcomes (PROs) and Health-Related Quality of Life (HR-QoL) in Patients with Ovarian Cancer: What Is Different Compared to Healthy Women?," *Cancers (Basel)*, vol. 13, p. 631, 2021, doi: 10.3390/cancers.
- [18] Z. Wang *et al.*, "Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd," *Nature Communications*, vol. 7, Sep. 2016, doi: 10.1038/ncomms12846.
- [19] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, Jan. 2010, doi: 10.1002/jcc.21334.
- [20] P. Przymus *et al.*, "Deep learning in microbiome analysis: a comprehensive review of neural network models," 2024, *Frontiers Media SA*. doi: 10.3389/fmicb.2024.1516667.
- [21] M. V. Kuleshov *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Res.*, vol. 44, no. 1, pp. W90–W97, Jul. 2016, doi: 10.1093/nar/gkw377.
- [22] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "Fastp: An ultra-fast all-in-one FASTQ preprocessor," in *Bioinformatics*, Oxford University Press, Sep. 2018, pp. i884–i890. doi: 10.1093/bioinformatics/bty560.
- [23] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, 2012, doi: 10.1038/nmeth.1923.
- [24] J. Wirbel *et al.*, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nat. Med.*, vol. 25, no. 4, pp. 679–689, Apr. 2019, doi: 10.1038/s41591-019-0406-6.
- [25] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," 2014. [Online]. Available: <http://ccb.jhu.edu/software/kraken/>.
- [26] H. Ma, T. W. Tan, and K. H. K. Ban, "A multi-task CNN learning model for taxonomic assignment of human viruses," *BMC Bioinformatics*, vol. 22, Jun. 2021, doi: 10.1186/s12859-021-04084-w.
- [27] A. Daina, O. Michielin, and V. Zoete, "SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules," *Sci. Rep.*, vol. 7, Mar. 2017, doi: 10.1038/srep42717.
- [28] S. Tang, J. Ding, X. Zhu, Z. Wang, H. Zhao, and J. Wu, "Vina-GPU 2.1: Towards Further Optimizing Docking Speed and Precision of AutoDock Vina and Its Derivatives," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2024, doi: 10.1109/TCBB.2024.3467127.
- [29] G. Bottegoni, "Protein-ligand docking," 2011.
- [30] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings," *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3891–3898, Aug. 2021, doi: 10.1021/acs.jcim.1c00203.