DESIGN AND DEVELOPMENT OF ML MODEL BASED ON FEATURE EXTRACTION FOR CLASSIFICATION AND SURVIVAL PREDICTION IN LUNG CANCER PATIENTS

A Thesis Submitted

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

By

Ms. SONIA

Regd No. – 17SCSE301033

Supervisor

Dr. MUNISH SABHARWAL
Professor

Co- Supervisor

Dr. ALOK KATIYAR
Professor

SCHOOL OF COMPUTING SCIENCE & ENGINEERING
GALGOTIAS UNIVERSITY
UTTAR PRADESH
FEB 2024

“श्री राधे कृष्णा अर्पण”

APPROVAL SHEET

This thesis entitled "**DESIGN AND DEVELOPMENT OF ML MODEL BASED ON FEATURE EXTRACTION FOR CLASSIFICATION AND SURVIVAL PREDICTION IN LUNG CANCER PATIENTS**" by **SONIA** is approved for the degree of **Doctor of Philosophy in Computer Science and Engineering**.

Examiners

____

____

Supervisor

Dr. Munish Sabharwal

Dr. Alok Katiyar

Chairman

____

Date:_________

Place:_________

CANDIDATE'S DECLARATION

I "SONIA" hereby certify that the work which is being presented in the thesis, entitled "**DESIGN AND DEVELOPMENT OF ML MODEL BASED ON FEATURE EXTRACTION FOR CLASSIFICATION AND SURVIVAL PREDICTION IN LUNG CANCER PATIENTS**" in fulfillment of the requirements for the award of the degree of **Doctor of Philosophy in Computer Science and Engineering** and submitted in School of Computing Science and Engineering, Galgotias University, Greater Noida is an authentic record of my own work carried out under the supervision of Dr. Munish Sabharwal and Co-supervision of Dr. Alok Katiyar.

The matter embodied in this thesis has not been submitted by me for the award of any other degree of this or any other University/Institute.

Ms. Sonia

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

Dr. Munish Sabharwal
Supervisor

Dr. Alok Katiyar
Co-Supervisor

The Ph.D. Viva-Voice examination of ___________________Research Scholar has been held on________________.

Sign. of Supervisor Sign. of Co-Supervisor Sign. of External Examiner

ABSTRACT

Cancer is a multifactorial complicated illness that is caused by numerous gene mutations or dysregulation of gene interactions. This study, on the other hand, proposes a unique method to cancer categorization. A large dataset of tumors has been taken and use the Naive Bayes classifier to categorize them. The Class Topper Optimization method is used to extract the features. The CTO algorithm is a novel artificial intelligence technology that is rapidly converging. The suggested method is straightforward, less complicated, accurate, and has a low error rate, which is critical in cancer classification. The accuracy, precision, error rate, and classification efficacy of the suggested method are shown. The results are also compared to the KNN classifier, which has been used by numerous researchers in the past to classify cancer. Experiments on a variety of datasets revealed that our innovative technique was both reliable and effective. The results demonstrate that the proposed method is both quick and accurate, making it an excellent choice for real-world cancer diagnosis. Cancer has long been a major danger to human health and well-being, having posed the greatest challenge in the history of human illness. Cancer's high death rate is primarily owing to the disease's complexity and the wide range of clinical outcomes. Because cancer survival prediction has become a major focus of cancer research, it will be important to increase the accuracy of this prediction. Many models have been suggested at the moment, however most of them simply use single genetic data or clinical data to construct prediction models for cancer survival. There is a lot of emphasis in present survival studies on determining whether or not a patient will survive five years. The personal issue of how long a lung cancer patient will survive remains unanswered. The goal of this research is to estimate the overall survival time with

lung cancer. Two machine learning challenges are derived from a single customized query. To begin with, determining whether a patient will survive for more than five years is a simple binary question. The second step is to develop a five-year survival model using regression analysis. When asked to forecast how long a lung cancer patient would survive within five years, our models are accurate to within a month of the mean absolute error (MAE). Lung tumors have been linked to a number of biomarker genes. As a part of our investigation into lung cancer prognosis, we integrated a feature selection method with a classification system. Using feature selection approaches to minimize the number of features, we believe that most classification systems may be improved. There are certain factors that have a greater impact on the categorization algorithms than others. The findings of our tests using a well-known classification technique, namely naive Bayes, SSA, have been provided. As a result, naïve Bayes provided superior output without SSA, but SSA enhanced performance. New algorithms and feature selection strategies will be tested in the future as part of this research. Our experiments will include both cluster and ensemble methods.

Keywords: Lung Cancer, Machine Learning, Cancer prediction, Electronic nose, Accuracy, Cancer classification, Cancer detection, Heuristic Class Topper Optimization (HCTO), Naïve Bayes, precision, Recall, Squirrel search algorithm

ACKNOWLEDGEMENT

I **“SONIA”** Working as an Assistant Professor and doing research for the degree of Ph.D. at Galgotias University was a quite magnificent and challenging experience for me. In all these years, many people directly or indirectly contributed to shaping my career. It was hardly possible for me to complete my doctoral work without the precious and invaluable support of these personalities.

I would like to give my small tribute to all those people. Initially, I would express my sincere gratitude to my Supervisor “**Dr. Munish Sabharwal”,** Professor and Co Supervisor **“Dr. Alok Katiyar”** Professor, School of Computing Science and Engineering for his valuable guidance, enthusiasm, and over-friendly nature that helped me a lot to complete my research work promptly.

I must owe a special debt of gratitude to Hon’ble Chancellor Mr. Suneel Galgotia, Mr. Dhruv Galgotia, CEO, and Hon’ble Vice-Chancellor Dr. K. Mallikharjuna Babu, Galgotias University for their valuable support throughout my research work.

Without family support no one can achieve their dreams in life, same is in my life also, so I am here obliged to my loving husband **Mr. Amit Kukreja**, who has been a source of incessant motivation and encouragement to me and who has always extended their profuse support to me in completing this research work. My son **Manik Kukreja** also deserve special acknowledgement for willingly making the sacrifices necessary for me to enable me to work for realizing one of the major ambitions of my life.

In last, my heartiest gratitude to my Parents LATE **Sh. Harbans Lal** and LATE **Smt. Raj Rani** and grandparents LATE **Sh. Gurbaksh Rai** and LATE **Smt. Bhagwan Devi** and the ALMIGHTY GOD, who blessed and enabled me to complete my work.

SONIA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

S.NO	ABBREVIATION	DESCRIPTION
1	EC	Evolutionary Computing
2	GA	Genetic Algorithms
3	SNPs	Single nucleotide polymorphisms
4	MSA	Multiple Sequence Alignment
5	GA	Genetic algorithms
6	PSO	Particle swarm optimization
7	ANN	Artificial neural network
8	SVMs	Support Vector Machines
9	ILP	Inductive logic programming
10	LR	Logistic Regression
11	EGRBF	Evolutionary Generalized Radial Basis Function
12	PLS	Partial Least Squares
13	FCC	Ferocystic changes
14	FCC	Fibroadenosis cystic
15	MRI	Magnetic resonance imaging
16	PEM	Positron Emission Mammography
17	MLO	Mediolateral oblique
18	CC	Craniocaudal
19	GIS	Geographic Information System

20	CT	Computerized tomography
21	WDELM	Weighted discriminative extreme learning machine
22	SVM	Support vector machine
23	HOG	Haar Wavelet feature
24	TL	Transfer learning
25	GEO	Gene Expression Omnibus
26	HCTO	Heuristic Class Topper Optimization
27	SE	Level of Section
28	ST	Section topper
29	CT	Class topper
30	PI	Performance Index
31	NB	Naïve Bayes
32	MIL	Multiple instance learning
33	OMBRFE	Optimal Mean based Block Robust Feature Extraction method
34	SVD	Singular value decomposition
35	IPcc	iterative Pearson's correlation coefficient
36	SSA	Squirrel search algorithm

LIST OF PUBLICATIONS

1. Sonia Kukreja, Munish Sabharwal, Mohd Asif Shah, D. S. Gill, "A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival", Computational Intelligence and Neuroscience, vol. 2023, Article ID 4506488, 9 pages, 2023. https://doi.org/10.1155/2023/4506488

2. S. Kukreja, M. Sabharwal and D. S. Gill, "A Survey of Machine learning algorithms for Lung cancer detection," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 338-342, doi: 10.1109/ICAC3N56670. 2022.10074272.

3. Sonia Kukreja, Munish Sabharwal, Alok Katiyar, D.S. Gill, "Automated Classification of Cancer using Heuristic Class topper optimization based Naïve Bayes classifier", SN Computer Science 2024-02-10, Journal article, DOI: 10.1007/s42979-023-02586-3.

4. Sonia Kukreja, Munish Sabharwal, D.S. Gill, "An effective Lung Cancer Diagnosis Model using the CNN algorithm", IACIDS23 (Waiting for publication), 2023.

CHAPTER 1
INTRODUCTION

Cancer is the preeminent reason of mortality in the modern world, with lung tumour being among the top 3 largest deadly diseases. Oncologists would be able to confirm changes in tumour size more easily if liver tumours were segmented. This information may then be used to assess the patient's reaction to therapy and, if required, make adjustments to the medical device [1]. Categorization of images in medical field is the utmost considerations in ample scope of applications. Because medical picture data has a lot of variation, it's important to apply the right categorization methods.

The computer tomography modality is used in clinical diagnosis and aids radiologists. Different tissues may be differentiated in CT images based on grey levels, which provide information that aids medical diagnosis. In many industrialised nations, cancer is a largest reason of death from illness. In medical practise, cancer categorization is based on clinical and histological facts, which might lead to inaccurate or misleading findings.

The human liver is located in the upper abdomen and is responsible for digestion and the elimination of waste materials from the circulation [2]. The presence of excess waste cells in the liver might result in a mass of tissue known as a growth or tumour. Cancers are classified as benign or malignant. Doctors remove benign tumours since they are not malignant.

Benign tumours, in the vast majority of cases, do not recur after they have been removed. Haemangiomas are tumours made up of masses of twisted and clogged blood vessels that are benign [3]. Cancer is defined as malignant tumours. The majority of primary liver tumours start in hepatocytes, which are liver cells. Hepatocellular carcinoma, often known as malignant hematoma, is a form of cancer that affects the liver. PAT scan, Ultrasonography (US), magnetic resonance imaging (MRI) and computed tomography (CT), are the most frequent medical picturing procedures for early identification and diagnosis of liver tumours (MRI) [4-7].

Because a CT scan allows a clinician to confirm the presence of a tumour as well as quantify its size, precise position, and amount of tumour interaction with other surrounding tissue, it is the most widely used and recommended tool for detecting many different types of malignancies, including colon cancer. CT scans can be used to guide biopsies and other minimally invasive operations, as well as schedule and deliver radiation treatments for malignancies.

Smoothing a picture is necessary for feature extraction and categorization in image processing [8-10]. As a result, in biomedical image processing, a precise filtering approach is a must. To achieve good performance, a suitable denoising method for liver tumour images is required. Using a suitable segmentation technique, the undesirable portions of the liver pictures may be removed. After preprocessing and segmentation [188], feature extraction comes next, followed by feature selection.

When they've been chosen, they'll go through the categorization procedure. The difficulty of selecting the best de-noising, segmentation, feature selection, and prediction algorithm for the classification of liver tumour pictures is still a major research challenge [11-14].

1.1 ***Dedicated systems can be solved using Soft Computing methods***: Because of its structures, such as nonlinear programming, intelligent control, decision making support, optimization, and soft computing, a researcher can combine genetic algorithms and evolutionary algorithms, fuzzy systems and fuzzy set theory, chaotic systems and chaotic theory, and neural science and neural net systems. Soft computing has grown in popularity as a result of current technological advancements and the diverse backgrounds of researchers. Its use spans a variety of scientific and technical areas [15-16].

1.1.1 Soft computing is a new approach for calculating the human mind's substantial ability to learn environmental unpredictability, limited truth, and imprecision while maintaining tameness, resilience, and cheap output costs. Soft Computing technologies are affected by nature and aid in the conversion of information from natural systems to a computer system. It implies to a mixture of artificial intelligence, computer science, and ML techniques used in engineering areas such as spacecraft, aircraft, heating and cooling, mobile robots, communication 2 networks, converters and inverters, power electronics, electric power systems, and motion control, among others [17-20].

1.1.2 Probabilistic Approach or Probabilistic Reasoning is one of the fundamental soft computing methods (PR). This method proposes several probabilistic models for assessing evidence and inferring solutions based on probability theory. Probabilistic

Models enable the system to assess randomised, probabilistic, and dynamic variables in real-world issues. Fuzzy logic was created to imitate the nature of human thought in order to do real-time data calculations with less complexity.

To increase the efficiency of the Genetic Algorithm (GA), Fuzzy Logic methods employ evolutionary computation techniques to produce Fuzzy Logic Controller (FLC). Neuron Fuzzy Controller (NFC) is a Fuzzy Logic Controller that uses Neural Networks as one of the greatest tuning tools. Experimental findings show that NFC outperforms FLC in terms of performance. One of the most fundamental ways among SC techniques is Neural Networks (NNs) [21-22].

Adaptation, parallel processing, and learning are all capabilities of Neural Networks. Human interaction, adaptability, and knowledge representation are all missing from these NNs. Recurrent Neural Networks and Feed onward Neural Networks are the two types of networks based on their structure [21-22]. One of the most fundamental SC algorithms using a randomised search pattern is Evolutionary Computing (EC). The primary streams of EC are Evolutionary Algorithms (EA), Evolutionary Programming (EP), Genetic Algorithms (GA), and Genetic Programming (GP).

Each EP approach operates in a similar way, employing different processes to identify similarity and random changes in a trial population in order to pick an ideal solution that is more adaptable for future generation tests [23-24]. The techniques which are inspired from biology are known as soft computing methodologies such as evolution, particle swarming, ant behaviour, and human nerve systems, among others.

Soft computing is now the only way to solve issues (i.e., algorithms) when no mathematical modelling is available, discover a solution to an unresolved problem in real time, adapt to diverse conditions, and execute with comparable computing [25].

1.2. Bioinformatics: Bioinformatics is a field of biology concerned with storage, gathering of data, and statistical analysis of nucleic acid and gene sequence data. To deal with the problem, expertise in bioinformatics software and computers are required. It's a link between several fields of knowledge, such as computer science, information engineering, mathematics, statistics, a mix of biology and bioinformatics, disciplines that can analyse data.

In Bioinformatics, tools such as statistical and mathematical analysis are employed to analyse the data. Pipelines are frequently used in the field of genomics to analyse and study computer programming in order to practise biology in bioinformatics. The discovery of single nucleotide polymorphisms (SNPs) and candidate genes are two common bioinformatics applications.

Typically, such identifications are done to better understand the genetic origins of illness, adaptations of uniqueness, desirable features (such as species in agriculture), or demographic differences. Meanwhile, bioinformatics seeks to understand the principles of nucleic acid and protein sequences. Computational biology and bioinformatics are used to analyse RNA, protein, and DNA sequences in biological data. It needs enough sequence data to map the genes of challenging disorders like infertility, Alzheimer's, and breast cancer disease before it can start next-generation sequencing [26].

Thousands of DNA variants have been identified, all of which are linked to similar illnesses and behaviours. The ability to use genes for diagnosis or therapy is one of the most important uses, and prognosis is one of the most important. Many research is looking at utilising genes to prognostic or forecast the existence of illness in a promising method. In genetics, gene expression is important because it observes the collection of traits of an individual that interact with their genotype environment. Exons are the parts of a gene that are supposed to be produced in mature mRNA, while introns are the intervening DNA regions between exons.

Data from nucleotide sequences (protein coding) areas, DNA regions, and RNA coding regions between antigenic regions can be used to study the structure of genes. A gene is a piece of hereditary data that occupies a specific (locus) location on a chromosome (genetic code). In protein synthesis, genes play a crucial role. Alignment of Multiple Sequences Multiple Sequence Alignment (MSA) is a critical and widely used computational tool in bioinformatics, computational biology, and molecular biology for examining biological sequences.

MSA is used to analyse protein secondary, tertiary, and tertiary structure, as well as protein function prediction and phylogeny reconstruction, wherever homologous sequences are found [27]. Natural correct and excellent alignments may deduce important meaning, connections, and homology among different sequences, as well as provide useful information. They are also used to identify members of protein families.

MSA accuracy is one of the most essential aspects of bioinformatics procedures, and it is vital since many bioinformatics techniques and medical techniques rely on MSA findings. Nucleotide (DNA or RNA) or amino acid sequences can be found in gene sequences (Proteins). When optimum alignment is used on gene sequences, the highest number of related or identical residues is extracted through changes in their DNA through time, all living creatures acquire certain common traits from their predecessors.

These inherited traits, including as physical or behavioural features, or genetic (hereditary) illnesses, are discovered using comparable residues created following alignment: Pair-wise Sequence Alignment and Multiple Sequence Alignment are 2 different kinds of equalization methods. In terms of delivering more accurate functional, structural, and evolutionary biology information, MSA continues to outperform PSA [28].

1.3 Soft Computing Techniques for Bioinformatics: Advances in soft computing techniques establish extraordinary ideals of tools, algorithms, and technology in bioinformatics for committed resolutions such as parallel genome sequencing, fast sequence comparison, and reliable, examine the efficient modelling, automated gene identification, storage of heterogeneous data, and databases, and so on.

The primary challenges in bioinformatics include phylogenetic inference, protein structure prediction, multiple alignment, and other NP-hard tasks. Soft computing provides an advantageous technique for achieving efficient and dependable solutions to these issues. On the other hand, the quality of biotechnology is constantly improving, as evidenced by techniques like as microarray, which offer typical patterns for the undeviating classification of tissue activities, and bioinformatics provides favourable study opportunities. So, for a unified consideration of key processes in a systems biology, bioinformatics must transcend the boundary toward a massive integration of the characteristics and abilities at diverse vital concerns like statistics and computer science [29].

Prediction from the structure of Protein, DNA, genomics, and Statistical data analysis from the gene expressions are only some of the activities that are involved in biological data analysis. Statistical approaches for data processing, such as regression and estimate, were formerly utilised in bioinformatics.

It can manage data sets like complicated, ambiguous, and intrinsically huge in biology in a stable and computationally well-organized fashion utilising soft computing approaches in bioinformatics, thus fuzzy sets are utilised as a framework for analysing them. The majority of bioinformatics jobs entail the optimization and search of many criteria (i.e equalization score, energy, and overlay strength), yet robust, near, and quick approximation solutions are available.

PSO, SA, ACO, TS, and other soft computing searching algorithms were evaluated as important searching strategies for discovering vast and multi-layer output spaces. In the research of molecular biology, new data and models are produced every time, and old data and methodologies are replaced with new ones. Soft computing may easily be adapted to many situations. The designers have benefitted since the system does not need to be redesigned every time the environment changes. Meanwhile, because many problems have many competing objectives, soft computing multi-objective optimization techniques that are similar to more than one objective GA (genetic algorithms) appear to be common and perfect. Soft computing technologies may be reused to analyse biological data in order to excerpt richer and more expressive information and visions, either separately or in a hybridised fashion [30].

Evolutionary Computing Techniques Among all SC methods, evolutionary algorithms play a major role in genetic-based investigations. To alert any 10 hereditary or connected characteristics in gene sequences and to observe any severe changes, Evolutionary Computing (EC) algorithms take into consideration all people from each generation. After applying Evolutionary Computational methods to execute advance experiments, subsequent generations of the population will be sorted out depending on optimal values created. In the category of Evolutionary Computations, several algorithms are suggested. GA (Genetic Algorithms), particle swarm optimization (PSO), and differential evaluation (DE) are examples of population-based stochastic search approaches that are appropriate for genetic studies. GA is mostly used to quantify various limitations in biological sequence evolution. Applying GA to gene data sequences to predict certain features will yield the optimal chromosomal sequences.

The importance of PSO and DE algorithms in gene sequence analysis to predict some traits has been demonstrated. On a few data sets of gene expression relevant to breast cancer, the efficacy of PSO and DE methods were proven. Some excellent test routines to validate the obtained optimal results, the sphere, Ackley's, Beale's, and suggested hybrid functions were employed. Some values to categorise some set of gene sequences as diseased (breast cancer) and normal (non-diseased) sequences using this population-based stochastic search methods and certain optimum test functions. In addition, to empirically confirm the results, optimal values of the proposed hybrid approach to those of other existing methods has been compared.

1.4 Motivation: Tumour is one of the biggest life-threatening diseases among hereditary disorders. Cancer comes in various forms, the most common of which is breast cancer, which affects more women than men and can result in mortality as well as emotional and socioeconomic consequences. Cancer has an impact on both the body and the psyche. Because of early identification and government-supported research, the survival rate for breast cancer has improved dramatically, and the death rate has continuously decreased. Illness prediction is also a significant motivator that aids in disease prevention rather than cure. Another case study it may use as an example is the preservation of newly born stem cells, which are utilised to create immune cells in preparation for future health concerns. This research focuses to provide a framework for categorising cancer and non-cancer sequences by identifying structural similarities between gene sequences and optimising outcomes using sophisticated soft computing techniques.

This research may categorise a sequence based on the created ideal values or outcomes after using each approach, whether the sequence structure resembles or is close to cancer or healthy sequences. Using this as a fundamental method, there is the potential to apply a plethora of advances in the future to enhance the precision of the generated ideal outcome. This method can be used not just for cancer but also for other hereditary disorders.

1.5 Objective: The objectives of my research are:

- To gain knowledge about various machine learning algorithms for Cancer Detection.
- To Design a Classification model for cancer which minimize human error and making the process more accurate and hassle - free.
- Design an efficient model which predicts Overall survival time in Lung cancer patients

1.6 Organization of Thesis: This research is categorized into six sections. The first chapter covers the introduction and fundamental ideas of machine learning Techniques, Bioinformatics, Problem Domain, Motivation, and the present efforts' objectives. It basically explains about what is the cancer, how it is starting and how it is growing. What all different types of cancers are and the patterns which effects the cancer. The current work's Literature Review is described in Chapter 2.

By studying various papers, knowledge of the various machine learning algorithm has been obtained. The parameters used to achieve the accuracy has been studied and computational paradigm achieved through different machine learning algorithms has been considered carefully. The 3rd chapter discusses about various machine learning or artificial techniques being used for the detection of tumour. The 4th chapter discusses the use of ML for cancer classification.

In this, Research has been carried out on 4 types of cancer i.e breast cancer, liver cancer, bladder cancer and Kidney cancer. After collecting data-set it has been pre-processed and features were extracted through hybrid Class topper optimization algorithm. During this process the dataset has been divided in the ratio of 80:20 for the training set and testing set. In the last step, cancer has been classified using Naïve bayes algorithm. Computational paradigm like accuracy and precision were considered. Chapter 5 discusses cancer prediction model for the lung cancer patients. In this paper, Features have been extracted through squirrel search algorithm. This work basically gives the idea whether the lung cancer patient would survive for 5 years or not. This model also divided the dataset in the ratio of 80:20 for testing and training and gave the best accuracy. In the last chapter 6, conclusion and future scope has been defined.

CHAPTER 2
REVIEW OF LITERATURE

Data mining methods are used to excerpt biological information. Computer science, applied mathematics, artificial intelligence, informatics, statistics, biochemistry, and chemistry are all approaches used in computational biology and bioinformatics to tackle biological issues at the molecular level.

The fundamental idea of many approaches is to use computational resources to address issues on scales far beyond human perception. Research in bioinformatics and computational biology frequently overlaps with systems biology research. Chromosome assembly, genetic exploration, genome position, structure of proteins alignment, protein structure prediction, protein-protein interactions, gene expression prediction, and evolutionary modelling are just a few examples of major research projects and endeavours in this field. Because of the large quantity of data involved in various study disciplines, data mining techniques appear to be a viable option.

These approaches seek to reveal previously undiscovered information and correlations using data from a variety of sources, such as the findings of high throughput studies or health records. In the last few years, a variety of data sources have been available. DNA microarray studies, for example, create thousands of gene expression measurements and give a quick and easy approach to acquire large volumes of data. They are used to acquire information on gene expression changes from tissue and cell samples. Techniques based on gene expression profiles are more objective, accurate, and trustworthy than traditional tumour diagnostic methods, which are mostly dependent on the physical appearance of the tumour [30].

To model gene and protein interactions and construct a knowledge base of biological processes, multiple data sources must be analysed. Microarray data analysis enables the identification of the most important genes for a specific illness, as well as groups of genes with similar patterns, under various experimental circumstances. The most frequently utilised techniques to address these challenges include feature selection, classification algorithms, and clustering algorithms.

2.1 Overview of Data Mining Clustering and classification are two popular data mining techniques for uncovering hidden patterns. Classification is a crucial data mining (machine learning) approach that may be used to predict data instance group membership. Classification is a type of supervised learning in which a technique is learned for predicting class instances from pre-labelled (classified) examples.

Neural networks and decision trees are the most common classification approaches. Clustering is a data mining technique for classifying abstract or physical items into groups. Clustering is an unsupervised learning technique that uses unlabelled data to discover natural groupings of occurrences. A cluster is a collection of related things.

The distance between any two things within the cluster is smaller than the distance between any two objects outside the cluster. A linked region of a multidimensional space with a relatively high item density. Clustering is a technique for separating a set of data (or objects) into a number of distinct subclasses known as clusters.

2.2 Data Mining with Bioinformatics: Data mining, as defined by Hand, Mannila, and Smyth [31], is the study of large qualitative records to find earlier unknown, subconscious, complicated, and possibly relevant connections and summarize the data in fresh manners that are useful to the data owner, whereas Han [32] defines it as the extraction of interesting (previously unknown, implicit, non-trivial, and potentially useful) information or different patterns from the data. Bioinformatics is the use of computer science, molecular biology, mathematics, and artificial intelligence to find relevant information in big molecular biology databases.

Because classification is such an important element of biology, classification algorithms play an important role in bioinformatics, where they typically use structural similarities to infer functional similarity. For both probabilistic and deterministic problems, a wide number of such approaches are utilised. New discoveries will rely heavily on data mining. Data mining aids pattern identification in large datasets with several persona characteristics and disorders, and it may also build disease onset prediction models based on a person's genomic profile.

The feature selection from data mining approach has been proven to be effective in studies of the genome, where there are over 20,000 gene characteristics but only a limited number of data points. Even when there are one million persons in the data, tiny data points may still be a problem when just a few groups of people have comparable conditions. The potential of genomic research is inextricably linked to data mining.

2.3 Machine Learning: The field of machine learning arose from the effort to create artificial intelligence. Its fundamental goal is to make a machine that can learn and adapt to new knowledge. The perceptron model was created in 1957, therefore this field may be traced back to that year. The neurons in the human brain are used to simulate machine learning.

This is what led to the creation of the neural network model in the late 1980s. Because of C4's capable package, the decision tree approach became highly popular in the 1980s and 1990s.SVMs (support vector machines) were first found in the mid-1990s and are now widely utilised in industry. After 2001, there has been an increase in the use of logistic regression, an older statistical approach. Machine learning is the art and science of making computers work without being explicitly programmed. Machine learning has provided us with realistic voice recognition, self-driving vehicles, quick online searches, and a vastly enhanced understanding of the human genome in the previous decade. This is so pervasive now that you may unknowingly use it many times each day. The majority of the researchers participating believe it is a great method for advancing Artificial Intelligence to the human level.

2.4. Bioinformatics The algorithmic element of the data mining process is machine learning, which is a computer programme that improves its performance at a task via knowledge or experience.

- A learning system uses preliminary information to update the foundation for better performance on future data from the same source, and it conveys the new origin in an understandable symbolic form.
- Learning denotes system alterations that are adaptive in the sense that they enable the system to do the same task or tasks selected from the same population more successfully the following time around Learning association rules. The most common and well-studied approach for identifying interesting connections between variables in big databases is association rule learning. It is defined as the process of finding strong rules found in databases.

Learning using decision trees in statistics, information mining, and the field of machine learning, decision tree modelling employs a decision tree as a prediction model to convert an item's observations into inferences about the item's desired value. Tree models are known by various names, including regression trees and classification trees. In these tree structures, leaves represent class labels, and branches represent feature combinations that result in class labels. The decision tree can be used to express decisions instead, the classification tree that results can be used as an input.

Artificial neural networks: The neural network learning algorithm is an artificial neural network (ANN) learning method. It's a learning algorithm that's influenced by the form and functionality of biological neural networks. With a connectionist approach to computing, computations are regulated in terms of an interconnected network of artificial neurons processing input. Neural networks are nonlinear statistical data modelling techniques that have recently become popular.

They are commonly used to describe complex relationships between intakes and outcomes, to identify trends in facts, or to record the mathematical framework of an unknown combined likelihood distribution between observable variables. They usually appear as networks of neurons that are interconnected that can compute values from inputs by passing data across the network.

Support Vector Machines (SVMs). Support vector machines, also referred to as SVMs, are a set of supervised learning algorithms for classification and regression. An SVM training method creates a model to predict whether a new example belongs to one of two types of categories given a collection of training instances, each of which is tagged as belonging to one of two kinds of categories. Support vector machines (also known as support vector networks) are supervised learning models with associated learning algorithms that examine data and recognise patterns, and are used for classification and regression analysis in machine learning.

The basic SVM takes a set of input data and predicts which of two possible classes will form the output for each input. Support vector machines, in addition to doing linear classification, may also effectively conduct non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces. Group evaluation is the process of dividing a collection of data into segments (known as clusters) so that findings within the same group are analogous based on a predetermined criterion, whereas findings collected from other groups differ.

Different methods for clustering produce various types of presumptions about the structure of the data, which are expressed by some type of similarity metric and then assessed and divided across groups. Expected density and graph connectedness are used in several additional approaches. Clustering is an unsupervised learning process and a generic statistical data analysis tool.

Cluster analysis is the process for arranging an assortment of items so that objects in the same group (designated as a cluster) are extremely similar to those in other groups (clusters). It is a key component of exploratory data mining, as well as a general statistical data analysis approach used in a variety of fields such as bioinformatics, algorithmic learning, recognition of patterns, information retrieval, and statistical analysis of images.

A directed acyclic graphical model known as a Bayesian network. It is also a graphical model of probability that relies on an acyclic graph with direction for displaying an ensemble of random factors and their dependent relationships. The probabilistic connections between symptoms and illnesses, for example, may be represented using a Bayesian network. This network might be used to calculate the chances of many illnesses arising from each symptom.

There are efficient inference and learning methods available. Inductive logic programming (ILP) Inductive logic programming is a branch of machine learning that employs logic programming to represent hypotheses, prior information, and examples in a consistent manner. A logical database of facts is represented as an encoding of known prior information and a set of instances. The ILP system runs a fictitious logic programme that includes all positive instances but none of the negative ones. Bioinformatics and natural language processing are two areas where inductive logic programming comes in handy.

Algorithm Types in Machine Learning 2.5 The desired output of a machine learning methods or the kind of intakes available at the time may be used to classify it into a taxonomy. Supervised learning produces a function capable of mapping inputs to desired outputs. In a categorization problem, for example, the learner appears at the function's

input and output instances to estimate a function that translates a vector into classes. Clustering is an example of unsupervised learning.

During training, the labels in this model are not known. Semi-supervised learning gathers unlabelled and labelled instances to produce a classifier. Transduction, also known as transudative inference, is a method of predicting new outputs on exact and fixed (test) instances using observable and precise (training) data. Reinforcement learning teaches a person how to make a certain observation of the environment.

Each action has a minimal influence on the environment, and the environment provides feedback or responses in the form of incentives, which guides the learning process. Based on prior experiences, learning to learn generates its own inductive bias. By using autonomous self-exploration and supervisory processes including active learning, imitation, maturation, and motor synergies, developmental learning, also known as Robot learning, builds its own sequences of learning scenarios in order to collect repertoires of new abilities. One of the research's aims is to convert biological challenges into machine learning difficulties.

Machine learning is an artificial intelligence area concerned with the development and study of data-learning systems. Machine learning is all about generalisation and visualisation. After being trained on a learning dataset, an algorithm's ability to work correctly on disguised and novel situations is referred to as generalisation. The visualisation of data instances and the evaluation of functions on these data instances are part of all machine learning procedures. The explosion of data in recent years, there has been an explosion of data in the field of biology. Thanks to modern techniques, whole genome sequencing is now achievable.

The whole genetic code can now be mapped, sequenced, and analysed, from bacteria and viruses to humans. With the explosion in genomic data has come a massive increase in other biological data from the proteome, metabolome, transcriptome, combinatorial chemistry, and other sources. The magnitude of the resources necessary to produce data matches the level of the enthusiasm for its potential: hundreds of computers creating terabytes of data.

The most serious issue right now is data analysis, and computers are an important part for that. The technique of extracting information from big databases is known as data mining, or database knowledge discovery. Categorization, connections, and sequencing are the three forms of knowledge discovery outlined by Agrawal et al.

Attempts to categorise the data and divide it into classes, a categorization of the classes can then be used to anticipate the most recent unclassified data. Classes can be simple binary divisions (like "is an enzyme" or "is not an enzyme") or composite and highly valued hierarchies (like gene functional hierarchies). Associations are different patterns in data, frequently recurrent groupings of items that belong together.

When extracting patterns from DNA sequence motifs, sequences provide information about data in which time or some other type of ordering is essential. Due to the amount of data in today's databases, data mining by hand is impractical, thus machine learning techniques are commonly used. Machine learning is a method for creating or changing a system automatically using "training" data. The goal of machine learning is to make predictions based on the qualities acquired from the training data. Data mining (the analytical step of Knowledge Discovery in Databases) aims to identify previously unknown characteristics in data.

Depending on the algorithm's desired output or the type of input available during the machine's training, machine learning algorithms are categorised into taxonomies. When a function is trained under supervision, it may map inputs to desired outputs (also called labels, because they are frequently provided by the human experts labelling the training examples). Clustering is a good example of unsupervised learning from a set of inputs. Because labels aren't essential, they aren't utilised.

2.5 Neural Network-Based Cancer Classification: Despite the fact that many conventional cancer classification approaches are imprecise or incorrect in clinical practise. Microarray gene expression patterns can be used in molecular diagnostics to provide an objective, efficient, and accurate method to cancer classification.

Hong Hee Won et al. [33] proposed using a set of network classifiers trained on negatively correlated characteristics to correctly diagnose cancer sickness and comprehensively test the proposed approach's performance on datasets. According to the findings, the assemble classifier with negatively correlated features has the best recognition rate on the datasets. Authors [34] proposed and evaluated cancer cell classification performance using unsupervised and supervised learning techniques.

This research designed and implemented a single hidden layer FFNN with back propagation training for supervised learning. Non fuzzy, fuzzy, and c-means clustering techniques are used for unsupervised learning. The classification of various tumour types is critical for cancer diagnosis and drug development. On the other hand, most early cancer classification studies lacked diagnostic capability and were clinically based. Cancer classification based on gene expression data is well-known for holding the key to overcoming some of the most challenging issues in cancer diagnosis and treatment development[35].

Thousands of gene expressions may now be monitored at the same time thanks to a new DNA microarray technique. A thorough examination of the problem and the suggested solution is required to have a better grasp of the cancer categorization difficulty. This thesis provides a comprehensive analysis of several cancer classification approaches and evaluate them based on their classification accuracy, processing time, and ability to link physiologically significant gene information.

Cancer research, which makes use of bioinformatics, is one of the most important study areas in the medical field. In terms of giving better medication and lowering patient toxicity, accurate tumour type prediction is more useful. To acquire a better grasp of the cancer classification problem, effective approaches based on comprehensive gene expression analysis have been planned. The quantity of gene expression has been shown to be a key to solving basic difficulties in biological growth mechanisms, disease prevention and cure, and drug discovery.

Various machine learning and statistical classification techniques have been tried to classify cancer, but there are a number of challenges that make the process challenging. Data is One of the most significant recent developments in experimental molecular biology is microarray technology. This cutting-edge technique allows for the simultaneous monitoring of hundreds of genes' expression levels in cells, and it's increasingly being used in cancer research to better understand the molecular distinctions between tumours and allow for more precise categorization. cDNA microarrays were developed by the Brown and Botstein Laboratory at Stanford [36], and high-density oligonucleotide chips were manufactured by the Affymetrix company [37].

In cDNA microarrays, which are also known as DNA microarrays, the probes are mechanically placed onto modified glass microscope slides by a robotic array [38]. For a more detailed description and comparison of the two systems' biology and technology, see [39]. Despite the fact that both the number of experiments and the number of genes per experiment are rapidly increasing, gene expression data from DNA microarrays is defined by many measured variables (genes) on just a few observations (experiments).

For all talks and research on gene expression profiles in this thesis, the first way of data presentation is employed. Microarray studies face a variety of statistical challenges, including image processing, experimental design, cluster and discriminant analysis, and multiple hypothesis testing [35]. The main aims of most microarray studies are class comparison, discovery, or prediction [38].

A class comparison is used to determine if the expression profiles of distinct classes vary. If so, which genes differ in expression between the groups, i.e. gene identification. For example, which genes can be utilised to distinguish tumour samples from non-tumour ones [40]. Class discovery, for example, entails identifying subclusters or structure among specimens or genes in order to define previously undiscovered tumour subtypes. Class prediction is the process of predicting phenotypes from data from gene expression profiles. This includes classifying malignancies (tumour or non-tumour) or tumour samples into previously identified subtypes, as well as predicting patient outcomes such as which patients are more likely to experience severe drug toxicity [41] versus those who are not, or which breast cancer patients are more likely to relapse within five years of treatment versus those who are not.

The focus of this thesis will be on class comparison and prediction. For these two objectives, the most successful supervised analysis approaches are those that use known class information [42]. In practise, feature selection techniques are employed to identify distinguishing genes, while classification algorithms are utilised to construct models from training data and predict the phenotype of blind test cases.

Expression profiling of genes Expression microarray profiling is a high-throughput method used in biotechnology and molecular biology to examine the gene expression patterns of thousands of genes at the same time [43]. A typical microarray chip is made up of thousands of tiny DNA oligonucleotides patches, each of which carries a little amount of a specific DNA sequence.

Under the appropriate circumstances, this may be a short piece of a gene or another DNA fragment employed as a probe to hybridise a cDNA or cRNA sample. The hybridization is quantified and detected utilising fluorescence-based recognition of fluorophore-labelled targets to assess the relative amount of nucleic acid sequences in a sample.

In the microarray technique, gene-specific probes represent hundreds of distinct genes. After the probes have been collected on an inert substrate, the gene expression levels in a target biologic sample are evaluated.

RNA is extracted from the tissues of interest, labelled with a detectable marker (usually a fluorescent dye), and allowed to hybridise with the arrays. Samples of messenger RNA (mRNA) hybridise with complementary gene-specific probes on an array. The fluorescence intensity of each gene-specific probe is utilised to quantify the gene's expression level, and pictures are generated using a confocal laser scan.

The higher the degree of expression, the stronger the signal. One of two approaches can be used to produce microarray data. In a two-colour array, two samples of RNA, each labelled with a different dye, are hybridised to the array at the same time. The query sample (for example, breast cancer tissue) is labelled with one dye, while the reference sample (for example, normal breast tissue) is labelled with a different dye; the two samples are mixed in an approximate 1:1 ratio depending on dye incorporation. In this scenario, the expression of the paired samples is compared and reported. For single-color arrays like the Gene Chip, each sample is labelled and incubated individually with an array.

The quantity of expression of each gene is reported as a single fluorescence intensity that indicates an expected level of gene expression after any non-hybridized material in the sample is washed away. The data may be utilised in future research as expression values for each gene in each sample, regardless of the method or approach used. Every gene in each sample is examined, and the expression vectors from the many samples are merged into a single expression matrix.

In the matrix, each row represents a gene and its expression levels across all samples, whereas each column represents a single sample and its estimated gene expression levels (the sample expression vector) (a gene expression vector). Coloured matrices are often used to depict expression matrices (usually, red or green, even though other combinations, such as blue or yellow, are also common). In a coloured matrix, the hue and intensity of a variation in gene expression represent the relative direction and magnitude of the change.

Expression Data Pre-Processing Before being studied further, gene expression data, like much of the data provided to machine learning algorithms, must be pre-processed. The characteristics of the experimental data are used to guide scale transformation, data normalisation, missing value management, replicate handling, and other common preparation operations [44].

A gene acts as a feature or attribute, and a sample acts as an instance or data point, in a microarray dataset, which follows the conventional data format for data mining and machine learning. This data type is characterised by a large number of genes (usually tens of thousands) and a small number of samples (less than one hundred). This one-of-a-kindness leads to one-of-a-kind issues in assessing microarray data (e.g., intricate data connections, high levels of noise, and a lack of biological absolute knowledge), which data mining algorithms must deal with [45].

A huge number of microarray datasets have become publicly available in recent years as a result of increased growth in the bioinformatics business. The Array Express archive (http://www.ebi.ac.uk/microarrayas/ae/) has a massive quantity of publicly available microarray data. Although raw data (for a subset of the collection) can be downloaded, the datasets are all pre-processed and stored in MIAME and MINSEQE formats.

One of the archive's best features is the ability to browse the whole collection or do searches based on experiment characteristics, submitter, species, and other parameters. In response to inquiries, the system receives summaries of trials as well as complete data. Other datasets can be downloaded from the author's or tool's websites as well (e.g., LibSVM software by [46], GEMS software by [47]).

Categorization In microarray studies, particularly in cancer research, classification of biological samples using gene expression data is a key challenge. Traditional diagnostic techniques rely on a subjective assessment of the tissue sample's morphological appearance, which requires the presence of a visible phenotype and interpretation by a competent pathologist.

In some cases, cell shape or cell-type distribution can help identify the class, but in many others, seemingly similar illnesses might have vastly divergent clinical outcomes. Diagnostic classifications include cancer vs. non-cancer, tumour subtypes, and treatment response prediction. The technique of determining the diagnostic category of a tissue sample based on its expression array phenotype and the availability of analogous data from tissues in specified categories is known as classification [48]. For the first time, the feasibility of cancer classification based only on gene expression monitoring is demonstrated in [49].

2.6 Overview of Tumour Classification Systems Shifting the focus of tumour classification from morphologic to molecular is becoming increasingly common. Gene expression profiles might provide more information than morphology and could be used to replace morphology-based tumour classification methods. Gene selection [50] includes searching for gene subsets that may differentiate tumour tissue from normal tissue and may have either a clear biological relevance or a function in the tumorigenesis molecular mechanism.

Gene selection is crucial in cancer classification based on gene expression. In compared to the amount of tissue samples, the number of genes involved in the creation of a discriminant rule is huge. Many genes can reduce the efficiency of the tumour categorization system while also raising the cost. Look at criteria and techniques for reducing the number of genes and choosing an optimal (or nearly optimal) subset of genes from a larger collection.

Gene selection has several practical benefits over other techniques to reducing dimensionality, including its simplicity, predicted cost savings, and improved probability of being used in a clinical setting (e.g., principal components). According to the findings, a limited number of selected genes might be used as biomarkers for tumours after a preliminary evaluation of tumour categorization using microarrays.

2.6.1 Tumour classification and definition: A tumour is an abnormal mass of tissue that has the ability to develop. It's a term for unregulated cell division and proliferation induced by alterations in cell DNA [50]. As a result of these changed extra cells using oxygen, nutrients, and taking up space from healthy cells, tumour tissue would grow. There are 2 kinds of tumours: benign and malignant [51]. Malignant (cancerous) tumours include: Cells in malignant tumours that break free and travel via the circulation or lymphatic system can infiltrate neighbouring tissues and cause new malignancies in other parts of the body. The spread of dangerous cells from one part of the body to another is known as metastasis, and it is a primary cause of cancer death.

The primary tumour is where the cells originally spread, whereas the secondary or metastatic tumour is the affected place nearby. Physical excision of malignant tumours is followed by radiation and chemotherapy to destroy any non-spotted remaining malignant cells, or chemotherapy alone if surgical removal is difficult, especially in late stages. Benign (non-cancerous) tumours include: Benign tumour cells do not continue to grow forever, infiltrate surrounding tissue, or spread to other parts of the body (i.e. metastasise). These can frequently be reduced in size or completely removed without recurrence, and there is usually little risk of death if they are not removed. In essence, there are as many different types of tumours as there are different types of human cells, with somewhat more than 200 variations, some of which are quite common and others of which are extremely uncommon [52].

Almost all tumours are named after the organ or cell type that gave birth to them. Lung tumours, for example, are recognised as lung tumours, but melanoma is characterised as a cancer that starts in skin cells called melanocytes. Authors [53] reported a light reflection analysis of genomic data-based microarray gene expression classification method.

In the first phase of this proposed technique, which encompassed two stages, two filter algorithms selected significant expression genes from hundreds of genes. During the second stage, the recommended technique was applied to selected gene subsets as extra input variables. The method combined Logistic Regression (LR) with Evolutionary Generalized Radial Basis Function (EGRBF) [54] neural networks, both of which have previously been shown to be very accurate in predicting high-dimensional patterns. Finally, nonparametric statistical tests were employed to compare the results, indicating a high synergy between the EGRBF and LR models [55].

The authors [56] described a method for efficiently analysing huge volumes of data from DNA microarray gene expression patterns. This development method was identified based on variations in DNA microarray gene expression patterns inside the same organism by simultaneously measuring the expression of hundreds of genes. Finally, to focus on cancer gene classification, the distribution probability of codes was used. According to the authors [57], microarray analysis, often known as gene expression profiling, is a technique for analysing thousands of genes in a single sample.

Micro array [58] analysis created difficulties in a variety of fields by giving massive volumes of data that could be analysed to obtain useful information. In this study, gene samples were taken from biopsy samples taken from colon cancer patients. They devised a method for identifying artefact states and differentiating malignant from non-malignant genes based on learning vector quantization. Finally, utilising DNA microarray [59] gene expression patterns from the same organism, an organism was found.

A wide range of approaches for functional genomics research have emerged as a result of biotechnological developments in the twentieth century. Microarray technology is a relatively new development that has offered snapshots of which genes are expressed in cells from various tissues and diseases [60]. Technology for getting accurate microarray data is continually improving and developing in order to meet the demands of biological researchers. "Stable feature selection and classification techniques for multiclass microarray data," by Taha et. al, was submitted. Scientists discovered in a recent study that gene expression patterns may be utilised as a feasible alternative for predicting outcomes. When utilising DNA microarrays for classification, the size of the produced data sets is a major concern [61]. The researchers created a multiclass gene selection approach based on Partial Least Squares to identify the genes for classification (PLS). The novel notion is to solve the multiclass selection problem using partial least squares and divide it into two subproblems: one versus one (OvO) and one versus rest (OvR) [62].

The objective of this research was to figure out how to properly identify informative genes. As a consequence, a novel approach for identifying a small number of key genes has been developed. The method allows for the development of a more reliable classifier with reduced classifier error. At the same time, this strategy produces more stable sorted feature lists than previous approaches [63].

Alok Sharma and Kuldip K. Paliwal proposed a Gene Selection Algorithm Using Bayesian Classification Approach. In this paper, they introduced a gene (or feature) selection method based on the Bayes classification methodology [64]. This approach can discover important gene subsets for the cancer classification 80 problem. The recommended technique begins with an empty feature subset and then adds a feature that provides the greatest information to the current subset.

The process of integrating features are completed when no feature may add information to the existing subset. The Bayes classifier is used to determine whether or not a set of attributes is beneficial. It is said to be the most accurate classifier [65-68]. The proposed method has been put to the test on a variety of publicly accessible microarray datasets, with promising results.

To obtain the gene subset, forward selection is utilised. The proposed technique has proved to be highly promising in terms of classification performance on three DNA microarray gene expression datasets. The authors proposed a unique concept dubbed " Microarray Gene Expression for Cancer Classification by Using Fast Extreme Learning Machine with ANP"[69-72]. DNA microarrays have shown to be an effective tool for cancer diagnosis and study in molecular biology. The Extreme

Learning Machine (ELM) is a unique learning algorithm that is used to increase the accuracy of the system's outputs in this current study.

This method avoids the issues that iterative learning systems are known for, such as inadequate learning rate, local minima, and overfitting, while also completing training quickly. The ELM's performance is improved by the Analytic Network Process (ANP). This method was predicted using the Lymphoma data set. The proposed strategy delivers greater classification accuracy while requiring less training time and effort than previous techniques [73-76].

New research named "A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set" has been published. Multilayer Perceptron Networks, Hybrid feature selection techniques [77], and Partial Least Squares (PLS) methods were used to assess the performance of four different supervised classification algorithms in this study (MLP). The experiment's findings show that the PLS regression technique is a good feature selection method, and that combining various feature selection and classification methodologies can result in high-performing microarray data classification models.

2.6.2 Microarray Tumour Classification Challenges: In the past, the statistics, database research, and machine learning groups have all worked on the classification problem. However, due of the particular nature of the problem, gene classification, as a relatively young field of study, poses significant challenges. Here, go through some of the concerns in further depth. The first problem stems from the peculiar nature of the gene expression data collection that is currently accessible.

However, the successful use of DNA microarrays [78-80] and high-density oligonucleotides has enabled the monitoring of thousands of gene expressions possible in a timely and cost-effective manner. The bulk of these data sets, such as the Colon tissue samples and the Leukaemia data collection, have small sample sizes. The attribute space, or number of genes, in the data is enormous.

There are tens of thousands to hundreds of thousands of genes in each tuple. When the samples are mapped to attribute space points, they might appear as exceedingly sparse points in a high-dimensional space. The majority of existing categorization algorithms were not designed to deal with this sort of data. In such a condition of sparseness and high dimensionality, all classification methods face substantial challenges [81-83].

Furthermore, because the tuple contains so many genes, calculation time will be a big issue. As a result, developing an efficient and effective cancer categorization system is a tough endeavour. The second stumbling block is the data set's inherent noise. The two forms of noise are technical and biological noise. Biological noise is the noise introduced by genes that don't have a role in cancer categorization [84-86]. In truth, the great majority of genes have little to do with cancer types.

Technical noise is associated with various stages of data processing, whereas biological noise is associated with biological processes. When noise is present, paired with a small sample size, data classification becomes exceedingly difficult. Dealing with a large number of irrelevant features is the third challenge (genes). Although irrelevant features may be identified in nearly all types of data sets previously investigated, the proportion of irrelevant to relevant attributes in gene expression data is much lower [87-90].

In most gene expression data sets, the number of relevant genes accounts for a tiny portion of the total number of genes. The vast majority of the genes are unrelated to cancer in any way. The presence of these irrelevant genes obstructs those key traits' capacity to distinguish. This not only raises the classification difficulty but also adds more processing time to the classifier's testing and training stages. One method to deal with this is to include a gene selection process that selects a set of relevant genes. Following that, cancer classifiers may be built using these chosen genes. Another method is to incorporate the identification of important genes into the training phase of the classifier [91-94].

It is a difficult task to classify cancer efficiently and effectively using any approach, demanding more research. The cancer classification application sector poses the fourth challenge. Although precision is crucial in cancer classification, it is not the primary goal. Another essential criterion is biological relevance, because any biological data gathered during the operation might aid gene function research and other biological studies.

The geography of diseased cells or tissues, or the identification of genes that function as a cluster in defining malignant cells or tissues, are two instances of important knowledge obtained through the classification process. All of information would aid biologists in learning more about genes and their interactions. As a result, biologists are particularly interested in classifiers that can give high classification accuracy while also revealing crucial biological data [95-97].

83 Microarray study presents a distinct set of data mining challenges. In the web or financial industries, unique data mining applications comprise a large number of records (thousands or even millions), but only a few fields (nearly several hundred). In contrast, a typical microarray data analysis study may only operate with a few hundred records. Because collecting microarray samples is challenging, the number of samples in many fascinating scenarios is likely to remain small [98-100].

However, with so many fields compared to so few data, there's a strong possibility you'll find "false positives" by accident - both in terms of finding differentially expressed genes and constructing predictive models.

***2.7. Breast anatomy**:* If you want to understand why studies on breast cancer are so important, you need to have a basic understanding of how breasts work and some of the diseases that affect them. When cells divide abnormally, a malignant neoplasia is formed. A special kind of radiography called mammography uses radiation doses at certain intervals to obtain breast pictures with the goal of detecting abnormalities that may signal an illness, such as breast cancer [101-103].

When it comes to mammary diseases, it is critical that they be discovered as early as possible. Detection of breast cancer instances has increased as a result of developments in imaging technology. A good prognosis may be determined by early detection of abnormalities using mammography in this case [104]. During the foetal phase, the epidermis creates a depression where the mammary gland is located, resulting in a mammary pit. The top ventral part of the trunk has the mammary glands visible on the left and right sides. Although both men and women have breasts, the mammary glands are more developed in women, save in rare cases involving hormonal imbalances [105].

Nips are little conical protrusions that sit on top of dusky tanned skin contain enormous sebaceous cells, which is generally imperceptible to human eye. From second rib above to the sixth rib below, the female breast's base is nearly spherical. Both medically and laterally, it touches the mid-auxiliary line on the lateral edge of the sternum's body [106-107].

The female breasts, as well as the nipples and areolas, often expand throughout puberty in tandem with glandular growth and fat deposition. Genetic, ethnic, and nutritional variables all have a role in breast size and form. The hue of the areola darkens during pregnancy and remains dark afterwards. While its hue fades once nursing ends, it never completely disappears from the human body. Fatty tissue fills the spaces between lobes of the breast, which are made of calls tissue, fibrous cell, and connective cells [108-109]. The parenchyma of the mammary gland is composed of 15 to 20 lobes of glandular tissue.

They are made up of lobules joined by areolar tissue, blood arteries, and ducts and give the breast its distinctive form thanks to a substantial quantity of fat inside them. Lactiferous ducts open on the nipple of each lobule, which drains each lobule individually. While breastfeeding, milk collects in the lactiferous sinus of each duct just below the areola. A minor portion of the lactiferous ducts is made up of alveoli that enter into the tiniest branches [110-112]. The hormones progesterone and estrogens are responsible for a wide range of breast tissue changes throughout the menstrual cycle and pregnancy. When a lady is not expecting or breastfeeding, alveoli are quite tiny and firm, but at the time of pregnancy they expand and the cells multiply rapidly [113].

However, mammary glands are ready to secrete milk since mid-pregnancy, the milk is only produced after birth. To rid the alveolus of fatty degeneration, colostrums, the first milk, is given to the infant. Breasts develop larger and more pendulous in women who have given birth more than twice, whereas in older women, they tend to shrink due to the loss of fat. However, women who are in young age, the cooper's ligaments generally support and maintain the breasts' original position [114-117].

The gland's lobes are held together by ligaments found in the top section of the gland. People across the globe are affected by cancer. In 1900, cancer was believed a condition that could not be cured, and research in this field started. When cells expand and replicate uncontrolled, they form a tumour or neoplasm, much like other types of malignancies [118-121]. Tumours may be benign if the cancerous cells don't move to other areas of the body, or they can be malignant if the cancerous cells spread to other parts of the body through the bloodstream or lymphatic system, a process known as metastasis.

Although their fat deposition and mammary gland growth are different, both the lobules and alveoli that form the breasts of infants and toddlers are comparable. When a woman reaches adulthood, her breasts take on their mature look thanks to the hormones progesterone and prolactin, which are responsible for driving the latter stages of her development. Estrogen levels rise throughout pregnancy. The breast gland ducts expand and branch as a result of this process, resulting in the accumulation of extra adipose tissue.

2.8. Pathologies of the breast: Pubertal girls are more likely than any other age group to develop fibroadenomas, which fall into three categories: common, enormous, and juvenile. A proliferation of glandular and stromal materials, well-defined boundaries, and hard, rubbery, freely moving, solid breast masses are all characteristics of these tumours. Fibroadenomas do not cause pain or soreness, and their size does not fluctuate with the menstrual cycle [122-125].

Adolescent girls and women in their early twenties are the most often impacted demographics. Rapid development does occur on occasion, but it is much more common for it to be exceedingly sluggish [126-127]. It is common for large fibroangiomas to have a diameter of exceeding 5 centimetres, although the average is just 2.5 centimetres. Women should be aware that these cancers may recur (about 20% recur), thus they should undergo regular checkups [128-130].

2.8.1. Dysplastic breast tissue: Ferocystic changes (FCC), fibrous mastopathy, or fibroadenosis cystic (FCC) are other names for mammary dysplasia. In actuality, these changes are not indicative of an illness and do not need further investigation. Pathologists classify it as a benign breast change characterised by cystic dilatation of intralobular glands and either stromal fibrosis or neither 131].

This lesion's age ranges from 20 to 50 years. Breast fluid retention and development into nodules or cysts, which resemble lumps when felt, are often linked to cyclic fluctuations in ovarian hormone levels in women going through ovulation and the weeks preceding menstruation. When the breasts are at this stage, they have a texture comparable to that of the premenstrual stage. A rise in the volume and density of the breasts, excessive modularity, quick change and fluctuation in cystic area size, increased discomfort, as well as spontaneous nipple discharge are all symptoms of fibrocystic alterations in the breast tissue. It might be one-sided, bilateral, or just affect a portion of the breast [132-133].

2.8.2. Mastitis and breast abscess Acute mastitis and breast abscess, both of which are inflammations of the breast, are very uncommon. Often, these infections occur after childbirth or after a wound. Acute and chronic mastitis are the two forms of mastitis. Neutrophil-rich granulocytes predominate in acute mastitis, which mostly affects nursing mothers.

Reinfection or a relapsed infection may cause chronic mastitis; the first occurs irregularly and is often transferred by the infant, while the second indicates that the pathogen has not been eradicated. When mastitis is left untreated and there is milk retention, it may lead to a breast abscess. Ultrasonography of the breast and needle aspiration under local anaesthetic are two of the most often utilised diagnostic procedures for therapy [134-137].

2.8.3. Cancer and Breast Cancer: Cancer kills one out of every eight people on the planet. In affluent nations, Tumour is the 2nd largest cause of mortality, while in underdeveloped nations, it is the third largest cause. In 2009, the number of new cases of breast cancer in India has gradually climbed, reaching up to 100,000 per year. One in four fatalities in the United States is attributed to cancer; this is the second most common cause of mortality.

The normal features of cells are fundamentally altered by a succession of molecular changes that lead to cancer. Control mechanisms that normally prevent cancer cells from growing and spreading to other tissues have been rendered ineffective. In the presence of signals that ordinarily limit cell development, these changed cells divide and proliferate, and so no particular signals are required to stimulate cell growth and division any longer. Changes in cell shape, a reduction in cell adhesion, and the development of novel enzymes are all features that these cells acquire as they get older [138-139].

Present normal cells, which normally hinder the development of surrounding cells, these heritable modifications enable the cell to divide and thrive. Cells that have undergone these alterations are more likely to propagate and infect nearby tissues. Mutations in the genes that code for proteins that govern cell division are the most common cause of abnormalities in cancer cells. More and more genes get mutated throughout time.

In many cases, this is due to a lack of DNA repair proteins since the genes that produce them are also altered, causing them to malfunction. Because of this, the number of mutant cells and their offspring begins to rise, leading to a cascade of more aberrant behaviour in the original cell. Other mutations may provide the aberrant cell a benefit that permits to proliferate significantly more quickly than perfect cells, resulting in the death of some of these altered cells. Most cancer cells are skilled to proliferate at a much higher speed in comparison of normal cells because they have obtained activities that are normally suppressed in healthy cells [139,140]. To be called benign, these cells must maintain their original position, but if they become nosy they are cancerous. Tumour plasm from malignant tumours may transmit to rest structure of the body, where they can grow into new tumours.

To begin with, cancer is a cellular illness. Cells divide and increase in response to the body's needs under normal circumstances. New cells are created and the body does not require it, and old plasms do not vanish as they should, disrupting this normally occurring process. Tumours are growths that are caused by an overabundance of these extra cells. Benign cancer is one form, while malignant cancer is another [140-141].

2.9. Different kinds of cancer

2.9.1. Benign Tumours that are benign, or noncancerous, are called benign. They're normally easy to remove and don't grow back once they're removed. To put it another way, there is a very little risk of a benign tumour becoming life-threatening.

2.9.2. Malignant Unlike benign tumours, malignant tumours may be cancerous. Unhealthy and erratic cell division is the norm. The cells assault the surrounding tissue vigorously. To create new tumours elsewhere in the body, they may also leave the malignant tumour and travel via the circulation. It is believed that only a small number of the viruses that infect people may cause cancer. A subclass of RNA viruses, DNA viruses and retroviruses are examples of this.

2.9.3. Cancer stages: Tumours are categorised according to their Stage. Tumour stage refers to the microscopic appearance of the cells. There are several stages of cancer in our bodies [142-143].

2.10. Finding out what stage of cancer the patient has*:* The "stage" of cancer will be determined by your doctors after they have identified the sort of cancer you have. A measure of how far along it is in its development. The TNM is an example of a staging system, although there are many more. There are three components to a TNM score: tumour size, number of lymph nodes affected, and metastasis (the advancement of disease to other parts through lymphatic and/or circulatory systems). The likelihood of a successful therapy is higher when the cancer is at an earlier stage.

On the basis of how cancer cells appear under a microscope, several subtypes of breast cancer may be identified. Adenocarcinoma, a kind of carcinoma that originates in glandular tissue, is the most common form of breast cancer. Breast cancer can be prevented, but there is no proven method of doing so. When it comes to battling breast cancer, early diagnosis is essential. It is critical to the detection and treatment of breast cancer.

Breast cancer is the major cause of cancer mortality among females, accounting for 23% of all cancer diagnoses and 14% of all cancer deaths worldwide. Economically developing nations are currently the major cause of cancer mortality for women. Approximately 700 women are diagnosed with breast cancer each year. For women aged 40 to 55, it is the second largest cause of mortality, according to American data. Breast cancer prognosis may be greatly improved if it is detected early enough. As a result, a number of jurisdictions have instituted screening procedures. Mammograms are generated in vast numbers by these programmes.

Breast cancer is a kind of cancer that starts in the breast tissue [144-145]. Intelligently designed computer processing algorithms, programmes, and software may improve diagnostic information from medical imaging. There are two types of breast cancer lesions. Microcalcification in breast cells may appear in knots or patterns and be linked to extracellular activity. A cluster of microcalcifications is often more noticeable than a single microcalcification and may help diagnose breast cancer in its earliest stages. Three or more microcalcifications may be detected in a mammography region of 1 cm or more in these clusters. After cancer has been suspected, the ability to discern between benign and malignant microcalcification is essential [146-147].

Because they resemble the normal breast parenchyma, masses are difficult to diagnose on mammograms in comparison of microcalcifications. Masses may have a variety of shapes and borders, ranging from tightly delimited to loosely spiculated. Detecting a tumour might be challenging, but the form and texture of the lump can indicate whether it is benign or cancerous. Consistently smooth and defined, benign lumps have round-like forms.

In contrast, malignant tumours have uneven shapes and ill-defined borders. An irregular-shaped tumour has more chances to be malignant than a knot with a regular-shaped mass, however this is not always the case. However, despite the wide variety of breast abnormalities, their border forms with the surrounding breast tissue may help distinguish benign from malignant breast cancers. Using mammography or ultrasound, spiculations on the malignant tumour may be examined to differentiate between benign and malignant tumours.

When breast cancer has spread into the surrounding tissue, it causes a stellate distortion called spiculation, and the presence of this feature is critical to identifying the tumour as malignant. Numerous effective approaches based on mammography or ultrasound are available for determining the difference between benign and malignant breast cancers based on spiculation degree. Breast cancer may be categorized into two: non-invasive/in situ and invasive/infiltrating, depending on where in the breast the cancer first appeared (glands, ducts, fat tissue, or connective tissue).

Tumours that are still contained inside the epithelial layer are known as carcinoma, and they are the first stage of carcinoma (an invasive malignant tumour caused by damaged epithelial cells). Infiltrating cancer occurs when cells that originate in glands or ducts invade healthy tissue nearby. There are a multitude of ways this cancer may look. Breast cancers may be both ductal and lobular, depending on the position of the tumour in relation to the breast.

When abnormal cells are detected on the surface of the breast milk duct, they are known as DCIS (ductal carcinoma in situ). In the surrounding breast tissue, the abnormal cells haven't made their way outside of the ducts. Treatment options for early-stage cancers such as ductal carcinoma in situ is excellent; nevertheless, if left untreated or missed, it may spread to the surrounding tissues and cause serious health problems. in situ refers to "in the original location" or "cancer" in the context of "carcinoma in 23 situ" (National breast cancer foundation).

Nearly 80 percent of breast cancer cases are caused by infiltrating ductal carcinoma, which is the most common kind. On mammography, this form of malignant tumour often appears as an irregular lump. The milk glands and the terminal lobules are the places where lobular carcinoma originates. As the lobules grow, a consistent number of tiny but unusual cells populates the area.

The lumen of the acini is usually obliterated by this procedure [148]. There is no way for these abnormal cells to get through the lobules' walls. Mammograms seldom reveal abnormalities associated with LCIS. To remove benign lesions, it is common to find it in biopsies that have been performed. Breast cancer is more likely to occur if you have LCIS. Careful follow-ups help the vast majority of patients.

Lobular carcinoma accounts for around 10% of breast cancer cases. Between DCIS and IDC, the differences are clear DCIS indicates that the cancer has not spread beyond the milk duct, while IDC indicates that the cancer has spread beyond the milk duct and into the surrounding tissue. To determine the most effective course of therapy, Doctors use cancer staging in conjunction with physical examinations and medical histories [149]. Metastization is the process by which cancer spreads to other regions of the body through the blood and lymphatic system.

Approximately 1% to 4% of breast cancer is characterised by an aggressive tumour that has penetrated the dermal lymphatic. Breast inflammation is a common symptom of this kind of cancer. It indicates that the HER-2/neu gene and oestrogen, progesterone, and the three most frequent kinds of receptors known to drive most breast cancer development are absent from the malignancy.

Breast cancer cells have tested negative for HER-2 (hormone epidermal growth factor receptor 2), oestrogen receptors (ER) and progesterone receptors (PR) (PR). Hormone therapy and medicines that target oestrogen, progesterone, and HER-2 are useless because the tumour cells lack the essential receptors. Chemotherapy is still an option for treating triple negative breast cancer. In fact, chemotherapeutic treatment for early-stage triple negative breast cancer may be even more effective than for other types of cancer [150-152].

2.11. Cancer risk: Cancer may strike anyone, but not all women are at the same level of risk. Women having genetic histories, genetic testing, exposures, and other variables are used to determine risk and propose breast screening and risk management measures. When a woman notices a change in the appearance or feel of her breast, she should immediately notify her doctor.

Mammograms are the primary screening tool for breast cancer in women. If you're concerned about your risk of breast cancer, you may utilise breast MRI to screen your patients. Breast ultrasound may be used to determine whether abnormalities are solid masses, or to check pregnant women at high risk for breast cancer who may not be able to undergo MRI or mammography. Biopsies and similar technologies are also used by experts to monitor changes in the breast or tumours found there. Do I have an increased chance of developing breast cancer? So, how do one know about finding out whether one have cancer of the breast?

The lifetime risk of breast cancer for those at very high risk is 30 percent or higher. There is an increased chance of breast cancer for intermediate-risk women, although they have a lifetime risk lower than 30%. Average-risk women with no family history, or specific genetic markers have a lifetime risk of breast cancer of 10-13 percent (20-29 percent) dependent on these circumstances [153-156].

2.12. Other breast diseases: Even while general breast cancer screening tests are useful for research, most women will not benefit from them when it comes to early detection of the disease.

2.12.1 Exam for discharge from the nipples: The fluid from a woman's nipple discharge may be examined under a microscope to discover whether there are any cancer cells present. Secretions are the most common cause of nipple discharge. Cancer is very uncommon if the discharge is milky or clear green [157-160].

It's possible that cancer is to blame if the discharge is scarlet or red or brown in colour, indicating that it includes blood, but other possibilities include injury, infection, or benign tumour. A breast cancer cannot be ruled out even though no cancer cells are discovered in the nipple discharge. Even if a patient's nipple discharge does not contain cancer cells, a biopsy of the suspicious lump is required [161-162].

2.12.2. Aspiration of the nipple and ductal lavage: For women at high risk of breast cancer, ductal lavage has been developed as an experimental test. An accurate picture of the risk of breast cancer may be gained from this test, which does not screen for or diagnose the disease. A Doctor's office or an outpatient facility may perform a ductal lavage. The nipple region is numbed with an anaesthetic cream. The milk ducts' natural apertures may then be found by using gentle suction to bring a little quantity of fluid up to the nipple surface [163-166].

A catheter is then placed into a duct's entrance with the use of a needle. To gently cleanse the duct and collect cells, saline (salt water) is carefully injected into the catheter. The cells in the ductal fluid are examined under a microscope in a laboratory after it is removed via a catheter. The procedure of ductal lavage is not recommended for women who are not at high risk for breast cancer.' No one knows whether this will ever be of any use. No studies have proven that the test can identify cancer at an early stage. It's more useful as a way to assess a person's cancer risk than it is as a way to find out whether they have cancer [167-169].

The usefulness of this test needs more investigation. In addition to looking for abnormal cells in the ducts, nipple aspiration is a simpler procedure since no instruments need to be put into the breast. Aspiration of the female nipple is carried out using tiny cups put on the breasts. Breast nipple fluid is brought to the surface of the breast by heating, gentle compression, and mild suction provided by the device. Analytical testing is then carried out on the nipple fluid. It is possible to utilise the technique as a test of cancer risk, however it should not be used as a screening test. No studies have proven that the test can identify cancer at an early stage [169-171].

2.12.3. Tests: Biopsy Mammograms, other imaging tests, and physical exams may all detect changes (or abnormalities) in the breast tissue that can indicate the presence of cancer. A biopsy is the only means to confirm or deny the presence of cancer in the body. There are many different kinds of pathologists, but the most common is named a pathologist because of their extensive training. The Doctor receives a report from the pathologist that includes a diagnosis for each sample [172-174]. This test's findings will be utilised to assist in the management of patient care. Fine needle aspiration biopsy, core (big needle) biopsy and surgical biopsy are all forms of biopsies [175-177]. Both have their advantages and disadvantages. There are a number of factors to consider when deciding which to utilise [177-180].

As part of their evaluation, Doctors will look at the lesion's appearance, its size and location in relation to other breast lesions as well as the patient's medical history and personal preferences. A patient may want to explore the advantages and disadvantages of various biopsy methods with their Doctor. In order to identify breast cancer, regular mammograms are required [180-182].

Mammography is the most prevalent screening procedure. A mammogram is an x-ray image of the breast that is used to diagnose breast cancer. Detection and sampling of non-palpable breast abnormalities, as well as the extent of breast tumours, both locally, regionally, and at distant locations are critical to breast cancer screening [182-184]. Imaging may also be used to assess a patient's reaction to treatment. As a result, imaging with various modalities is a critical and ongoing component for patients with breast cancer throughout the course of the illness, from the time of initial diagnosis to treatment. X-rays, also known as radiographs, are often employed in the detection of cancer because they provide a two-dimensional representation of the area being examined [185].

Chest radiographs are used to identify early cancers, or to check whether the disease has moved to the lungs or other parts of the chest, such as the lymph nodes [186]. Mammograms are used to identify breast illness in women with symptoms (such as a lump or nipple discharge) or abnormal results from a screening mammography (such as an increased risk of cancer). There are additional photos of the problematic region in a diagnostic mammography. In certain circumstances, a tiny region of aberrant breast tissue might be more easily evaluated using magnified pictures known as cone or spot views. As the name suggests, ultrasound uses high-frequency sound waves that cannot be heard by the human ear to produce images of the inside of the body [187-189].

Sonograms are images created by the echoes of sound waves. A transducer is smeared with gel and placed on the breast flesh using a portable tool called a transducer. Sound waves are emitted, and the echoes are picked up when they bounce off the tissues of the body. Black and white images of echoing sounds appear on a computer screen. No radiation or discomfort are involved in this procedure [190].

If abnormalities are detected through a screening or diagnostic mammography or a physical exam, breast ultrasonography may be performed to further investigate them. It is not common practise to utilise breast ultrasonography for screening purposes. When screening high-risk women with thick breast tissue, several studies advocate using ultrasonography in conjunction with a mammography. Mammograms, on the other hand, may be replaced with ultrasounds [191-192].

For certain categories of women, further research is required to determine whether ultrasonography should be included to regular screening mammography. Some breast lumps may be examined more closely using ultrasound, and this is the only technique to determine if a mass is a cyst without using an aspirator to remove fluid from it. A biopsy needle may be guided into a suspicious region of the breast with the use of breast ultrasonography. Currently, there is a 3-D automated entire breast ultrasound that can be utilised on the breast [193-194].

It may now be used in conjunction with mammography, thanks to FDA approval. When using a portable transducer, the 3-D ultrasound is possible, but most typically, a bigger transducer is put over the whole breast and scanned automatically. In addition to mammography, ultrasound has become a vital tool since it is widely accessible, non-invasive, and less expensive than other methods of detection. Even if this is less relevant with modern automated ultrasound equipment, an ultrasound's worth is still dependent on the operator's expertise and experience [195]. If you want to screen for breast cancer, don't rely only on ultrasounds since they may miss part of the disease that is detected by mammography.

In terms of tumour detection, ultrasound is less sensitive than magnetic resonance imaging (MRI), but it is less expensive and more readily accessible. Soft tissues and fluid-filled cysts are particularly well suited to ultrasound imaging. Doctors may use it to evaluate how far cancers of the uterus, oesophagus, or rectum have migrated and if cancer has gone into blood vessels, notably the liver and pancreas [196-197].

As a guidance for minimally invasive treatment of tumours including liver, prostate, and others, ultrasound is often used. To examine tumours that are difficult or impossible to see on a mammogram, ultrasonography is an essential tool. Ultrasound is sometimes used in conjunction with other medical imaging technologies.

Mammographically occluded breast tumours may be detected and the extent of the illness can be determined using MRI. In addition to the rising use of MRI for breast cancer staging, MRI-guided needle localization and core needle biopsy procedures have been developed. Women with a high breast cancer risk might benefit from MRI screening as well [198-200].

Instead of using x-rays, MRI scans use radio waves and powerful magnets. Depending on the kind of tissue in the body and the illnesses present, radio waves collect energy and then release it in specific patterns. When a pattern is sent into a computer, it may be rendered into a highly detailed picture. A contrast liquid called gadolinium is injected into a vein before to or during a breast MRI scans in order to enhance the clarity of the images. An MRI scan might take up to an hour to complete. In order to have a breast MRI, you must lay face down on a platform created specifically for the operation within a tiny tube. Without compressing the breasts, the platform provides holes for each of them. To get an MRI picture, the platform must include sensors that can collect the data required [201-203].

During the scan, you must maintain complete stillness. Mammograms and MRIs may be used to screen women at high risk of breast cancer or to carefully investigate problematic regions detected by a mammogram, respectively. For women who have been diagnosed with breast cancer, an MRI may be performed to better identify the disease's size and to screen for additional malignancies in the breast. In planning a breast cancer patient's surgery, it's not yet obvious how useful this information is [204].

It's fairly uncommon for someone suspected of having breast cancer to have their opposite breast examined to make sure there aren't any cancers there, too. Mammograms are a vital part of the breast cancer detection process. Mammograms have the potential to minimise a woman's risk of dying from breast cancer. Mammograms have their own set of risks. Overdiagnosis is the greatest danger. Breast cancer that would not have been dangerous to a woman's health if she hadn't been detected is a common occurrence in this situation [205-207].

A lady can't tell right now whether her cancer is going to spread or not. As a consequence, almost all women who are diagnosed with breast cancer get treatment [208]. An overdose of treatments such as surgery or chemotherapy or radiation might result in serious consequences. Another, more prevalent problem is that a mammography might falsely indicate the presence of breast cancer when none is present. A "false-positive" result is what it's termed when this occurs. There is a risk of unnecessary follow-up testing and treatments due to false positive findings. While some women aren't bothered by false-positive test results, others are. Mammograms become more important as it get older since the chance of acquiring invasive breast cancer rises with age.

An imaging process for evaluation of the breast known as mammography provides information on breast shape, anatomy, and pathology [209-210].

Additionally, it is utilised for the identification and diagnosis of cancerous breast lesions. A successful course of treatment for breast cancer depends heavily on its early identification. An X-ray process comparable to conventional X-rays, but with low dosages, great contrast and resolution, and minimal noise is achieved. Because the breast is very susceptible to ionising radiation, using the lowest radiation dosage possible while maintaining great picture quality is preferable [211].

It is easier to detect fatty breasts using mammography, which is more sensitive and specific than dense breast. Women under the age of 30 had the most difficulty assessing dense breast tissue. To help with needle core biopsies and to pinpoint non-palpable tumours, mammography is also used.

Detecting abnormalities in screening mammograms requires a high level of sensitivity and accuracy, thus the breasts must be compressed uniformly to ensure picture contrast. The prototype of Positron Emission Mammography (PEM) is designed to test the concept of PET technology in the detection of breast cancer and ganglion invasion. PET systems with fine-grain crystal segmentation may achieve greater spatial resolution than whole-body PET systems, and their coverage of the area under study is more precise, resulting in higher sensitivity [212-213].

Calcifications and masses are the two most common breast alterations seen by mammography [214]. Breast tissue calcifications, which appear as white dots on an image, are mineral deposits inside the breast tissue. Cancer may or may not be to blame. On a mammography, another major alteration is noticed in the form of a mass that may or may not have calcifications. In addition to cysts and solid tumours that do not harbour malignancy, masses might represent a variety of other conditions.

Biopsies are normally required for any lump that isn't obviously a fluid-filled cyst (A biopsy is taking out a piece of tissue to see if cancer cells are in it). It's critical that you make prior mammograms accessible to the radiologist. Because they may indicate changes over time, they can assist determine whether or not a biopsy is necessary. Detecting breast and armpit cancer at a size of at least 2 mm with the PEM system will be critical for early diagnosis since it will improve resolution ten times that of present PET systems. When injected into a patient, a radioactive chemical (called a radioactive tag) reacts with malignant cells in the PEM system before being dispersed throughout the body via the patient's blood stream. Tumour cells, whose metabolism is greater than that of normal cells, take in more of this glucose-based liquid.

The liquid's radioactive isotope releases positrons (electron's anti-particle) that swiftly recombine with electrons, resulting in, among other things, two photons in the same direction and in the same line. They may be picked up by crystals that light up when they are struck by these photons (collision).

These photons' pathways overlap, indicating the presence of malignant cells. PET imaging of cancer uses fluoro-2-deoxy-Dglucose (FDG) because cells with elevated metabolic demand absorb FDG more readily than normal tissue [174]. Primary breast cancer has become more accessible to be diagnosed because to the widespread usage and availability of FDG PET imaging (and particularly PET/CT) for whole-body evaluations in the assessment of the majority of cancers (including those in the breast).

Particle emission mammography (PEM) is the preferred method for making the first diagnosis of breast cancer (PEM). However, despite its diagnostic accuracy, whole-body FDG PET has a lower sensitivity than other conventional diagnostic imaging modalities for identifying malignant breast tumours. The gamma-ray detector used for PEM imaging is either a twin head or a ring design. Following the annihilation process, the gamma rays from the two systems will move at an angle of around 180 degrees from one another.

The event location in PEM imaging is determined as a line of response between the locations where each gamma-ray impacts the pair of opposing detectors.

Because PEM does not suffer from the same loss of resolution 36 as BSGI/MBI systems, it has an advantage. One drawback of the dual-head PEM detector design is the lack of resolution in Z-axis owing to the constrained acquisition angle (depth). Although ring detectors don't have this constraint since they give a 360-degree acquisition for reconstruction, ring detector devices don't yet have biopsy capabilities [215].

The dual-head detector system recently received a needle biopsy localization device. For individuals at high risk of developing malignant melanoma, PET with FDG is superior to conventional imaging in terms of sensitivity and precision. Studies demonstrate that the information from PET scans have been utilised by Doctors to influence treatment choices in many cancer patients. For cancer patients, changes in therapy resulted in fewer procedures and biopsies, as well as cost reductions.

The scintillation crystal in APDs generates an electrical signal when it is struck by a photon, and the FE system is responsible for processing this signal. The APD's analogue signal processing is the most difficult part of the FE. In fact, it has been determined that the APD signal should be amplified and processed in order to create a high-resolution reconstruction picture in a short amount of time.

Samples of the signal are taken when the FE has built the right form and recognise the signals with more energy. These signals and the crystals associated with them are delivered to the Data Acquisition (DAQ) system, which is responsible for acquiring and processing digital signals received from the FE. All of the components of a FE module are present in the DAQ system, including APD's, ASIC's, ADCs, and serialisation data transmission circuits that are responsible for transmitting digitised data to the DAQ system.

Due to their equal sensitivity, MRI and PEM play a similar function in clinical practise. Examining primary breast lesions in preoperative surgery planning or pre-chemo assessment remains the principal use of the exam. Radiation exposure is a drawback of PEM. On the basis of the relative risk for cancer induction in women aged 40, a single PEM study with label-recommended radionuclide dosage is linked with a 15-fold greater risk than screening films or digital mammograms. There is also a 25-fold greater risk of cancer-related death in general.

Only the fibro-glandular tissue is subjected to a high dose of ionising radiation during a mammogram. In PEM, on the other hand, radionuclides are administered to every organ in the body. As a result, mammography has just the danger of inducing breast cancer, while PEM has the potential to cause cancer in any number of organs that are radiosensitive. To catch breast cancer in its earliest stages, before any symptoms appear, mammograms are the most often utilized screening method. The prognosis is less favorable if the cancer has spread to other parts of the body by the time symptoms appear [216-218]. The goal of this procedure is to help the radiologist discover breast lesions more quickly and prevent the disease from spreading to a more advanced stage.

This imaging-based technique is the only one that is recommended for routine breast cancer screening because of its high sensitivity in finding clinically occult disease. Breast Cancer Detection Using the Hough Transform in Images (VIP image). Mammography is the breast cancer detection method that has the greatest impact on reducing mortality from the disease when used as part of a comprehensive population screening programme.

As the density of the breast rises, the mammogram's performance suffers. X-rays with a low amplitude and high current are used in mammography to evaluate the human breast during a diagnostic.

X-rays are very powerful electromagnetic rays with wavelengths in the range of things and bodies. There are two primary ways in which X-ray photons interact with tissue: the photoelectric effect and the Compton scattering. Electrons are ejected from an atom nucleus when X-ray photons of short wave length interact with the electric field there. An ionizing particle is formed when a free electron is released. Compton scattering occurs when an X-ray photon collides with an electron from an outside source and is released.

A scattering electron receives energy from the incoming photon, which is then expelled and ionized. It's time for a new direction for the photon to go. To understand the contrast in radiologic images, one must first understand the main process of Compton scattering. Mammography machines now use an X-ray tube to generate X-rays. The beam is narrowed as it passes through a collimator and a metal filter. To reach the image receptor, a part of the radiation travels through the breast and is reflected by an anti-scatter grid.

Photons interact and deposit energy locally, permitting the development of a picture. Automatic exposure control is activated by letting a small percentage of x-rays flow through the receiver uninterrupted to a sensor. When the X-rays penetrate the structures, the picture creation is influenced by the densities of the structures, since absorption is density-dependent.

Micro-calcifications, for example, need a picture with a high spatial resolution to clearly show their edges. There are two conventional image projections: craniocaudal (CC), which is a view from the top, and mediolateral oblique (MLO), which is a lateral view from a central position, providing a better view of the glands.

This gives a sense of three-dimensionality as well as a grasp of how structures overlap. A high-quality mammography with excellent spatial resolution and good contrast separation enables radiologists to see tiny structures clearly. According to research, the death rate for women over 50 may drop by 30% if they all had regular mammograms. A huge number of mammograms have had to be examined by radiologists in the United States due to the widespread use of screening programs. It takes a lot of practice and a lot of skill to read mammograms.

Because of radiologist weariness, breast tissue's complicated image structure, and the cancer's modest appearance, it has been estimated that 20% to 40% of all breast cancers are missed during routine screenings. However, even the most skilled readers of mammograms can only identify between 85 and 91% of cancerous breast tissue. In addition, research indicated that radiologists misread between 2.6 percent and 15.9 percent of negative or benign mammograms.

Researchers found that double reading by two radiologists might increase sensitivity up to 15%. While double reading may be expensive, time-consuming and theoretically difficult to execute, it is an option that might be considered. Mammography has its limitations, Normal breast structures may conceal malignant tumours, especially in thick breasts with a high proportion of fibro glandular tissues. Overlapping tissue can lead to unnecessarily repeated testing following diagnosis.

- There is a considerable degree of inter-observer variability. Biopsy recommendations have a low positive predictive value, and there is a considerable risk of erroneous interpretations resulting to false positives and negatives. This may lead to overdiagnosis and therapy.

The identification of malignancies in a breast Mammogram picture is a common Digital Mammogram application. Breast Mammogram systems may assist radiologists in evaluating pictures and spotting signs of cancer in the breasts. In addition to the human assessment of the diagnosis, several systems are utilized. Using a breast mammography system, not only does it enhance the quality of the cancer images, but it also minimizes the human effort connected with the diagnosis, and it improves the accuracy of detection and diagnosis of breast cancers.

In spite of the billions of dollars and decades spent on breast cancer research, there is still a huge gap in our knowledge of the causes and risk factors for breast cancer. Familial/hereditary and environmental variables are the two major categories of breast cancer risk factors. A wide range of non-hereditary factors, including dietary factors, exposures to chemicals, industrialization, agricultural processes that use ionizing radiation, occupational stress, physical activity, exposure to tobacco and alcohol, are referred to as environmental factors in breast cancer research. Fewer than 5 percent of breast cancers are genetic and roughly 30 percent are linked to environmental exposures; however, the exact aetiology of breast cancer is still a mystery.

2.13. Breast cancer that runs in families: A woman's chance of developing breast cancer is more than doubled if she has a first-degree relative who has had the disease. In families with a high lifetime risk of breast cancer (60-85 percent) and an elevated risk of ovarian cancer (25 percent), BRCA1 or BRCA2 gene abnormalities account for 20-40 percent of all cases of breast cancer.

Additionally, several inherited disorders are linked to an elevated risk of breast cancer, in addition to the high risk of hereditary breast and ovarian cancer. There is a slew of uncommon cancer predisposition conditions that have been linked to breast cancer as well. Breast cancer is more common in women with certain genetic conditions, such as Li-Fraumeni syndrome, PeutzJeghers syndrome, and Cowen Syndrome.

2.13.1 Potential dangers from the environment Individual lifetime exposure to reproductive factors/choices like early menarche, late menopause, age at first pregnancy, number of pregnancies, child-bearing, hormonal therapies, and endocrine disruptors which have an association with breast cancer; dietary exposures (alcohol, fat intake, and endocrine disruptors) that have an association with breast cancer (like Xenoestrogens, Aromatic Amines, Bisphenol A, Polycyclic Aromatic Hydrocarbons, Vinyl Chlorides, DDT, Ethylene oxide, Lead, Dioxins etc).Over 85,000 carcinogenic chemicals have been detected in the environment and their long-term exposure is linked to breast cancer, according to NBOCC and the Institute of Medicine.

Components such as race and socioeconomic status have the most impact on the level of effect each of the other factors has.

2.13.2 Breast cancer as a result of social inequality: The complex challenges of social inequities, including political, economic, and racial injustices, have been overlooked in the same way that environmental risk factors for breast cancer have been disregarded. Depending on where one live and work, the level and kind of risk factors to which people are exposed will vary.

The burden of hazardous substance exposure falls disproportionately on urban and rural poorer groups. Because low-income women are more likely to be exposed to harmful substances and to be subjected to the strains of social injustice, socioeconomic determinants of breast cancer have been found to have a stronger influence on the development and death of the illness. Women in the lower-income bracket are also less likely to be able to afford or obtain high-quality health care and nutritious meals.

Good research shows people need to understand the intricate web of environmental, social, genetic, and behavioural influences in order to really reduce the incidence and death from breast cancer.

The health cycle requires more effective interventions and more applications connecting the socioeconomic determinants of public health if health data is accessible from a single source [74]. In order to persuade policymakers, will need better visualizing and analytic tools like Geographic Information System (GIS).

Geospatial information systems (GIS) collect, store, analyze, manage and display geographically referenced information (linked to location). GIS mapping in the health sector is now being pursued by a number of organizations around the nation. This includes the design, development, and use of geographic information systems (GIS) to describe health problems, conduct epidemiological analysis, and administer health care. Some of its most common uses include describing and analyzing health occurrences in a particular region, conducting public health surveillance, assessing the health status of a certain population, and planning and implementing health services. Since John Snow's famous illustration of how geography may be utilized in epidemiological research in 1854, the healthcare industry has been using visualization and spatial analysis.

By mapping the locations of people who had been infected, he was able to identify the origins of the cholera epidemic in London. It was in the year 1840 when

Robert Cowan utilized maps to establish the correlation between crowd size and the frequency of yellow fever in Glasgow, England. Eradication of Yellow Fever as a result of Immigration was an obvious conclusion for him. This was done in 1843, and it included all affected people in the home. Data integration, dynamic database querying, and map-based display of results are just a few of the benefits that GIS technology has to offer.

Benefits of GIS usage include the visual impact and data analysis they give. Traditional research and statistical approaches are limited in their capacity to analyze data because of the lack of the ability to overlay data layers. Study participants utilized GIS to map and visualize geographical correlations between various health and health-related factors, as well as to conduct empirical measurements.

Geographic information systems (GIS) may be useful in the examination of many elements of healthcare access as well as health outcomes. There are many health care delivery concerns that affect the world's population, including cancer care, and any analysis must take this into account [173]. Geography is a critical component in addressing concerns of cancer epidemiology that affect cancer treatment and access[66]. As a result, geographic information systems (GIS) have been utilized to store, integrate, visualize, and analyze cancer data, highlighting patterns, dependencies, and inter-relationships. Additionally, GIS enables for interactive searches of information included in the maps, tables or graphs that are generated. Data changes are automatically reflected on maps thanks to a dynamic connectivity between databases and maps provided by this software [177].

In order to educate the public and policymakers, authorities may use map-based strategies that present complicated information in an easier-to-follow style. If you're doing a spatial disease mapping study, you'll want to analyze the disease-covariate relationships, define the illness's geographical variance, and identify high-risk regions. Geographic information systems (GIS) have altered the way academics examine the geography of health. As a result, GIS and health research focuses on the quantitative study of health-related events in geographical contexts, which separates health-related concerns for analysis and interpretation.

Medical geographers use disease mapping as a useful tool. Maps like this one may help researchers find correlations between diseases and other variables, such as pollution. Since disease maps make it possible to examine several elements related with illness at once, they inevitably lead to theories about what causes it. Medical geography has been used to map diseases and conduct spatial correlation research on health-related concerns. By identifying areas with a disproportionately high incidence of cancer, epidemiological research may be used to inform disease control and preventive efforts [75].

Visual analysis techniques strengthened with exploratory analysis has mostly been sufficient for epidemiologists, but quantitative modeling of disease distribution is needed to test certain hypothesis or to estimate the relationship between the measure of disease incidence and the environmental covariates. GIS may be used to generate data for epidemiological models by exhibiting the outcomes of statistical analysis and modelling processes that take place across a large area of the globe.

Geo-epidemiological study has traditionally focused on the detection of clusters. However, these methodologies cannot be utilized to assess whether regions with heightened risks are statistically connected and significant or not. Statistical innovations based on point- and area-based approaches have emerged as a result of these concerns and related statistical issues. It's important to know exactly where the illness occurred, whereas area-based techniques rely on the total number of cases throughout an area. Bayesian approaches allow the borrowing of information from close or adjacent areas when there are just a few occurrences or a tiny population.

GIS may be observed in a variety of health research because of its adaptability and power. SIRs (likelihood statistic) for each ZIP code were established by the New York State Cancer Mapping Project, 2001 in an effort to identify unusually high cancer rates in New York. The spatial proximity GIS tool was used to estimate historical environmental exposure to pesticides and other chemicals as part of the Cape Cod Breast Cancer and Environment Study. In a case-control study on lung cancer, researchers utilized self-reported and GIS-based proxies of residential exposure to environmental pollution to compare their results. They looked at land use patterns, proximity to major roadways, and exposure to industrial pollutants.

According to the results of a Bayesian spatial hierarchical analysis, population-based breast cancer screening in Queensland reduced cancer mortality compared to other detection approaches. California Breast Cancer Mapping: Identifying was created to examine how biomarkers connect with GIS-derived measures of exposures and risk assessment in California. Researchers in 2004 used the Space-Time Systems and Model Transition Sensitivity Analysis to find Space-Time Clustering. This study, conducted in Upper Cape Cod, Massachusetts, used generalized additive models to examine the spatiotemporal patterns of breast cancer.

The Retrospective Space Rime permutation scan statistic was also used by Cao et al., in Shenzhen city to discover a cluster of HINI Influenza A inside the city.
GIS-based cancer atlases might be used to communicate spatial findings, to produce and use statistical results, to design and use statistical models connected with GIS to examine the influence of rurality, area level and individual level socioeconomic status, and temporal changes.

Finding out why gaps exist may be utilized for lobbying, legislation, support, and educational activities. Several combinations of statistical and geostatistical models have been credited with additional recent study. GIS modelling has been utilized to analyze the geographical and temporal clustering of breast cancer cases in Western New York, based on historical exposure to environmental contaminants. Prostate cancer's spatiotemporal patterns and time trends may be evaluated.

Bayesian approaches and the Hierarchical General Linear Model were used. Geographic analysis was performed to identify high-risk locations of breast cancer in order to analyze how racial and ethnic disparities occur at the county level in Texas using Poisson, Bernoulli and multinomial models. Future epidemiological studies using comprehensive exposure models may benefit from spatial-temporal analysis of the breast cancer data. Using GIS-based spatial-temporal analysis, the present research shows how important it is to depict cancer risk, account for known confounders, and test for the statistical significance of place and time in research.

If you have access to thorough, residential histories, this strategy is very valuable. Despite the fact that our results do not provide an explanation, they do reveal regions of greater vulnerability and might therefore promote the formulation of ideas for investigation and eventual action. The primary emphasis of temporal analysis has been the discovery of illness clusters and outbreaks within certain time periods. Spatial patterns of health issues like breast cancer were observed using spatio-temporal analysis in the study of time series.

The mechanism of disease dissemination may be better understood by simulating regional variation. It's important to know if the extra instances start in places with high incidence rates, or whether they originate in areas with low incidence rates before moving elsewhere. In this scenario, linear spatial transfer functions are utilized. For a single research region, spatial-temporal approaches may be used to identify the geographic or temporal death rates for breast cancer Using this technique, the exact position, length, risk ratio, and statistical significance of the observed clusters may be determined.

A study found that the unequal regional distribution of cancer mortality in Texas (USA) highlights the significance of adopting a more data-driven strategy for cancer monitoring. It aims to determine whether the spatial-temporal pattern may put an unequal burden and if the excess tendency has continued into the present decade.

It is common in most geographic studies to include both spatial and temporal aspects. The cluster warning is not only restricted to a certain location, but it is also believed to be present for a set amount of time. The identification, description, and analysis of spatio-temporal processes are crucial to grasping many of geography's foundational notions. A methodology was used to data on children leukemia cases in northeast England and found considerable evidence of spatial and temporal grouping. Other approaches for space-time clustering have been developed, such as the k closest neighbor test.

As a way to find clusters of leukemia cases in New York City, authors used a Bernoulli model to create Space-Time Scan Statistic, which later went on to be implemented. "Cancer Incidence in Five Continents" is a comprehensive study of the global variance in cancer incidence. Genetics was previously thought to have a factor in some of the observed variance, but substantial strides have been made in this area in the last several years. It has been shown that discrepancies in cancer rates across nations may help identify potential causes by highlighting the impact of varying socioeconomic, cultural, and environmental variables.

It's easier to examine significant differences in cancer risk and aetiological variables across nations than it is to research minor differences within countries, which is, what is the need right now. It's increasingly vital, too, to provide viewpoints with a focus on the local context. Cancer incidence and death may be reduced in the future by addressing risk factors now, even when present cancer patterns are based on previous exposure to these variables. Often, simply drawing attention to differences in behaviour can have a positive impact on cancer prevention efforts at the official and individual levels. Geographical differences in cancer incidence, mortality, and survival have been connected to patterns of socioeconomic position and disadvantage.

In order to minimize cancer-related mortality and morbidity, it is important to look at the whole picture of health. Technological advances in medical techniques, diagnosis, and treatment for patient survival and detection of cancer risk are rising on a daily basis, usually at a regional or local scale. In order to determine whether a higher-than-expected number of cancer cases is due to a specific community, researchers must first determine who is at risk. For this, the forms of cancer, the stage of the disease, and the major diagnostic procedures are examined.

As a result, one of the most difficult tasks is figuring out what causes various forms of cancer. In addition to this, it is important to accurately count and record the number of cases in the population that is at risk. Identifying the population at risk for a particular/specific cancer might be critical in determining whether a cancer cluster exists. Analysis of illness incidence and correlation with a person's socioeconomic, environmental, and exposure factors is critical .

A statistically significant discrepancy between the actual and projected number of instances might occur by chance as well, which must be remembered. Additionally, it is necessary to look at the cancer's propensity to spread geographically. As a result, the pace of growth should also be taken into account when assessing the exposure or risk, or both. Currently, researchers are focusing on one or more of the aforementioned issues. One of the most difficult tasks in combating such diseases is integrating GIS technology, socioeconomic relevance, and spatial statistical techniques [181].

CHAPTER 3

A Survey of Machine learning algorithms for Lung cancer detection

3.1 INTRODUCTION

Cancer is a deadly disease which spreads very Fastly. Basically, in Cancer, cancer cells or abnormal cells start growing uncontrollably and damage the body tissues. The broader categorization of the cancer is in 2 types, one is benign tumour and other is malignant cancer. And under malignant cancer, different types of cancers are colon cancer, breast cancer, skin cancer, and brain tumour [119, 182], Lung cancer [176] and rectal cancer [183].

The second most common cancer [184] reported in women and men both is Lung cancer. Studies show that family history increases the possibility of lung cancer. Lung cancer affects the lungs and mostly people about 90% [187] who smoke, it occurs in them. So many tests are there to detect lung cancer.

The biopsy is meddlesome like a successful standard for detection of lung cancer, but scientifically it is complex and expensive, hence, it cannot be frequently opted by patients. Computerized tomography (CT) scan provides a communicable approach for diagnosing irregularities possibly being cancer, but this approach has the drawbacks of ionizing radiation, high priced, and false positives.

Positron emission tomography is an accurate noninvasive imaging test for the detection of larger mass lesions and pulmonary nodules. And it is popularly known as PET scan.

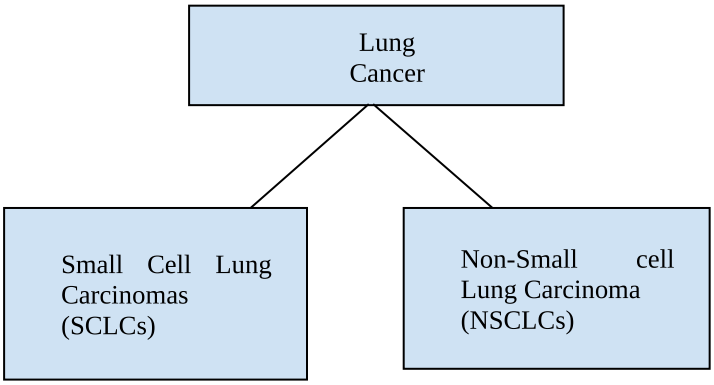
Lung Cancer
Small Cell Lung Carcinomas (SCLCs)
Non-Small cell Lung Carcinoma (NSCLCs)

Fig. 3.1: Types of Lungs Cancer

At 1st and 2nd stage of detection, there are bright chances to recover. If it is detected in the 3rd stage also, it can be recovered because so many advancements have been done but it is a time-consuming process.

Rest of the paper is organized as: Section II defines how the normal cells of a human being converted into malign cancer depending upon the symptoms. Section III is a brief description of machine learning and in section IV all the machine learning algorithms studies have been carried out. Section V explains about parameter analysis and summary had been given and the last section VI depicts the conclusion has been drawn.

3.2 Cancer Development

Fig. 2 depicts that, how the normal growth cells of human being has been converted into Malignant cancer. A human being's normal cells start converting into abnormal cells due to various factors. And these abnormal cells start multiplying at a very pace rate which results in malignant cancer.

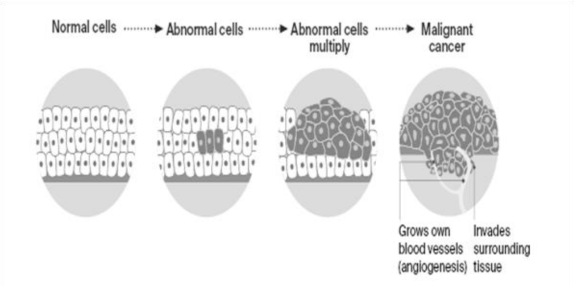

Fig: 3.2 Cancer Growth [200]

3.3 Machine Learning

Machine learning is a technique to learn the computers, where a machine is constructed with the help of the machine learning algorithms, by which machine can select its individual choices and give output to the user [201]. In machine learned computers, artificial intelligence is the subfield. Various techniques have been used to detect and predict any kind of disease. These algorithms help the machine take the input from user and gives the quick output so that appropriate actions should be taken for the critical disease.

Machine learning used for the optimization, because it takes inputs from user and cleanses the data by extracting the required features and gives us the optimized results.

3.4 Literature Survey

Samala et. al. [170], developed a new computer aided diagnosis technique for the observation of lung cancer. This Technique has been divided into two components: (a) CADe (A computer aided detection) and (b) CADx(a computer aided diagnosis). CADe module distinguishes the suspicious lung nodules and segments them for candidate extraction whereas CADx observes both malignancy allocation of patients by considering suspicious lesions from CADe nodule-level and assessment.

In this approach, screening has been done with the help of low dose CT scans and they are providing relevant probability estimates. LUNA 16 for CADe and kaggle data set(3D) for CADx has been used in this technique and this technique achieved malignancy classification task and lung nodule identification. Lung Nodule Analysis dataset contains 1186 lung nodules annotated in 888 CT scans. While Most nodule systems have been made and refined independently, Researchers believe that components involved in diagnosis and detection of lung cancer are critical.

Focus has been given to the false positive reduction so that correct prediction should be drawn. Although LUNA 16 has been designed for optimized results still coupling between diagnosis and detection of lung cancer has been considered in this paper, which produced better results and further to this model uncertainty has been characterized in the deep learning system [76].

CADe and CADx both attain superior outcomes than the best published CADe and CADx systems on the Kaggle Data [203] and LUNA16 [204] benchmarks. The research has been carried out on Stage 1 &2 patients.

Leilei et al. [205] has used an electronic nose technology to detect lung cancer. And for this technology pattern recognition system plays a crucial role. Electronic nose technology has been developed because the tests available for the detection of cancer like biopsy, PET Scan, and Computerized tomography (CT) are very expensive. The Electronic nose is also acknowledged as an olfactory system [206], is a bodily structures system which delivers the sense of smell.

This technology is working in a way to detect lung cancer by the breath that is exhaled [207]-[209]. The basic idea behind the electronic nose system is to find the different compounds (VOCs) in the exhaled breath, but the biggest challenge is class imbalance learning. And to address this challenge, a unique method for the classification named weighted discriminative extreme learning machine (WDELM) has been developed for detection of lung cancer [210].

In the beginning, WDELM empowers a distinct weight to individual samples by adopting a comparable weighting strategy, which empowers it to achieve categorized piece of work with the biased class dispersion. After that, a second repetitive algorithm is used to handle the convex unbiased method with an intellectual investigation presented. At the end, the persuasiveness of the recommended method has been assessed on the dataset available for lung cancer and also on public datasets by differentiating it from traditional methodologies.

The outcomes of WDELM have been compared with DELM, SVM, WELM and ELM in this article which shows this method enhances the specificity by allowing samples having large weights from the small sample sizes class, and hence obtaining an equilibrium recognition result. Experimental results have confirmed that accuracy and G Mean has been calculated on all features and selected features which were 0.854, 0.8433 and 0.8871, 0.8696 respectively. The WDELM excels traditional methods and is good enough for detection of lung cancer.[211]

In this article kaur et al. [206], a comparative study has been performed for the feature extraction and this feature extraction was of 2 types (a) Local feature extraction and (b) Global feature extraction. This feature extraction [207-209] framework has been utilized for the diagnosis of cancer found in lungs and in these images of CT scan has been used. In this investigation, 1000 images of CT scan dataset have been used. This framework works in three basic steps: data collection, local training and investigating, and global training and investigating.

Preprocessing of these images have been completed through image cropping and warping. Then features have been extracted so that feature vectors can be generated that will work in the detection models. 6 different machine learning algorithms were built on which these extracted features have been used and they were Neural network learning method, KNN learning method, Decision tree method, Random Forest method, Naive bayes learning method and SVM. After that comparative study has been performed over all these 6 methods and Support vector machine (SVM) was giving the best results [167].

The outcomes depicted that the Haar Wavelet feature, (HOG) [214] which is also known as histogram of Oriented, and Gabor Filter types exceeded the other seven feature categories. The developed local extracted feature technique excels the accepted worldwide one. The confined part, using SVM with Haar Wavelet attributes attained 88% sensitivity, 91% specificity, and 90% accuracy. Using SVM with HOG attributes attained 85% sensitivity, 89% specificity, and 88% accuracy. Finally, this technique achieved the best 96% sensitivity, 97% specificity, and 97% accuracy.

Masood et al. [176] have developed a new computer aided judgment support technique which is helpful in the detection of the lung nodules. This technique is based upon 3 Dimensional Deep CNN and helps the radiologists for the decision making. Multi region and median intensity have been applied so that the affected area can be automatically selected. The datasets in this research have been used are LIDC IDR, ANODE09, and LUNA16. The results carried out in this research achieved a better performance in the form of AUROC, sensitivity, accuracy, specificity, of 96%, 98.4%, 98.51%, 92%.

Xie et al. [215], have developed a new hybrid approach for the detection of lung cancer is biomarker + metabolomics and some machine learning algorithms. This hybrid approach was tested with 6 different types of machine learning techniques, and they are Neural Network, KNN, AdaBoost, SVM, and Random Forest. Biomarker is a biological molecule that is used to find any abnormality in the body or how well one's body is responding to a particular treatment with the help of the metabolomics which keeps the potential for the early detection of lung cancer.

In these technique sensitivity, AUC and specificity have been calculated and results show that Naive Bayes performed the best and achieved 1.000 in all. So, it shows that Naive Bayes is the best and strongest performer in the early prediction of lung cancer.

Researchers defined a new technique WCBA for the segmentation of the features of lung cancer and that technique is known as Water cycle bat algorithm. Median filter has been used for the feature extraction. On extracted features, Bayesian fuzzy has been applied. And this new hybrid technique attained the accuracy of 0.92 which was far better with compared approach.

3.5 Parameter Analysis Summary

Table 3.1 state that different machine learning algorithm has been used for the diagnosis of lung cancer and some have achieved the better results over another.

Table 3.1. Comparison of Techniques

Algorithm	Data Set	Type of dataset	Precision	Recall	AUC
CADe & CADx	LUNA16 & Kaggle data science bowl	Image	0.25	0.93	0.87
wdelm	Uci	Image	0.9908	0.99	0.97
Svm	TCIA	Image	97%	96%	97%
3ddcnn	LUNA16, ANODE0, and LIDC-IDR	Image	98.4%	98.51%	96%
Biomarker + metbolis	Hubei Taihe Hospital	Image	100%	98.1	0.989
WCBA	lLIDCI	Image	0.92	-	-

CHAPTER 4

Automated Classification of Cancer using Heuristic Class Topper Optimization based Naïve Bayes Classifier

4.1. INTRODUCTION

AI and Machine Learning are recently widely used in health care for the prediction of critical diseases like Colorectal Cancer [97,175], Alzheimer, Fetal Brain Abnormality Detection [71], Diabetes [72]. Cancer tumours are generated by cell growth that breaches the tissues of the human body. The two categories of tumours are benign and malignant tumours. A benign tumour is made up of local noncancerous cells. Cancer is a condition in which cells proliferate uncontrollably.

The malignant cells define the cancer kind. Breast cancer can begin in any part of the breast [73]. Another great approach for identifying cancer is a biopsy. Since the cells vary in size, localizing cancer cells in cancer pictures is difficult. Mastitis, adenopathy, and granuloma are some of the other abnormalities that may be observed on breast imaging.

Machine learning (ML) techniques have a wide range of predictions, including educational, bankruptcy, pattern recognition, image editing, feature reduction, fault, face recognition and micro expression recognition, and medical diagnosis. It has shown tremendous promise in the diagnosis of cancer, in particular. Many academics in recent decades have proposed numerous solutions for automated cell categorization in cancer detection.

Some researchers have focused on nucleus analysis in this setting, identifying nucleus characteristics that might help categorize cells as benign or malignant. The 2D input image structure is used to modify the CNN architecture. A considerable amount of data is required for a CNN training task, which is lacking in medical, particularly in cancer. The application of the TL approach to a natural images dataset, such as ImageNet, and the implementation of a fine-tuning technique, as illustrated in Fig. 1, is one solution to this challenge. By integrating the information of individual CNN designs, the transfer learning (TL) idea may be used to improve their performance [80,87]. The main benefit of TL is that it improves the classification accuracy while also accelerating the training process.

The key contributions of this paper are as follows:

1. HCTO based Naïve Bayes classifier is proposed for automated diagnosis of cancer patients.
2. The comparisons of the proposed model are drawn with various competitive cancer diagnosis models.
3. The paper uses the artificial intelligence approach HCTO for feature selection and to improve classification performance.
4. Overfitting is a problem that has been solved.

The structure of this document is as follows: Section 1 illustrates the introductory part of HCTO and the various optimization techniques, and Section 2 outlines some related and motivational work to develop the proposed method. Section 3 gives a detailed description of the proposed methodology and Section 4 depicts the derived results of the proposed method, and section 5 provides the discussion part and section 6 depicts conclusion of the proposed work.

4.2. RELATED WORK

There are many papers presented in the past for cancer classification and detection. Ting et al. [171] used a deep CNN to classify BC-lesions. Togaçar et al. [83] introduced the BreastNet, for the diagnosis of cancer & for detecting benign and malignant breast regions. Abbas [84] developed a multi-layer DL architecture to extract invariant properties, converting them into deep invariant features, and learning features to make the final decision were the four phases of this network. The MIAS dataset was utilized, and the sensitivity, specificity, accuracy, and AUC were all 92 percent, 84.2 percent, 91.5 percent, and 0.91, respectively.

Sha et al. [172] proposed a technique for automated identification and classification of the malignant area in breast pictures using the same dataset. The grasshopper optimization algorithm and CNNs were used to develop their suggested approach. The findings revealed that this technique could achieve sensitivity, specificity, and accuracy of 96 percent, 93 percent, and 92 percent, respectively.

For cancer detection, Abunaseer et al. [120-123] used CNN. The input image was 224 pixels, and the classification results were applied using the Softmax (SM) algorithm. The MIAS database was used to determine the network's overall accuracy, which was 65 percent. For mitosis classification, Precision, recall, and F-measure were all 0.50, 0.80, and 0.621 for their proposed strategy, respectively. Lotter et al.[88] also built a model for multi-class cancer classification that extracted features. Their model's sensitivity, specificity, and AUC were all 96.2, 90.9, and 0.94, respectively.

When building networks from the ground up, Jiang et al. [89] obtained better BC-classification accuracy than when using TL from a pretrained network. The accuracy was around 0.88 while using GoogleNet. Khan et al. [90] created a model where CNN was used to enhance performance. Without changing the source network layers, Cao et al. [91] improved TL's BC-classification performance (ResNet-125). Instead, they combined various feature groups using random forest dissimilarity.

The classification accuracy was raised to 82.90 percent using the "ICIAR 2018" dataset. Deniz et al. [92] adjusted the model layers to achieve more accuracy. Their model outperformed five other techniques with an accuracy of 91.37 percent. Celik et al. [93, 117] pre-trained the DenseNet 161 model using the same dataset, achieving F-scores of 92.38 and 91.57 percent, respectively.

Wang et al. [183] used the histopathology images for analysis of the whole slide lung cancer and more accuracy has been obtained which was 97.3. In this, a weakly supervised approach has been used to efficiently classify the dataset(TCGA). Chen et al. [188] has been incorporated LDNNET algorithm for lung cancer to achieve more accuracy, sensitivity, and specificity, which is 0.98, 0.99, and 0.98, respectively.

Zhang et al. [191] used TCGA to classify the data and achieved more specificity, accuracy, and sensitivity by 10.11%, 9.25%, 6.45%. and Li et al. [216] had used a new Relief-SVM for the classification of lung cancer and to efficiently perform the diagnosis. They successfully increase the accuracy by 73.91% on two different datasets. Picard et al. [217] had used a gene expression data and deep flexible neural forest model for the classification of cancer and the accuracy achieved is 93.6.

Sun et.al[218] have used the transfer learning approach and the accuracy was improved on TCGA was 0.98. Very recently, researchers have worked on cancer prediction and classification using advanced artificial intelligence tools. The methods used in classifies cancer, but the algorithm used is slow and complex. Furthermore, some of the traditional algorithms such as genetic algorithms and PSO stuck at local optima, thus depriving the optimum solution.

Meta heuristic technique Class Topper Optimization has been applied to many domains and it has produced good results in those domains. To the best of our knowledge, HCTO never used for cancer classification. The suggested approach employs a mix of HCTO and Naive Bayes classifiers for fast and simple cancer classification.

The suggested HCTO algorithm is based on the intellectual conduct of pupils in a given class at a school. There is a competitive conduct among the pupils in one class to be the class leader. The pupils' level of competition helps them perform better in class. It is possible to tackle many complicated optimization issues in real life by modeling the pupils' intelligent behavior.

The naive Bayes classifier is one of the most basic probabilistic classifiers. In spite of the strong assumption that all characteristics are conditionally independent given the class, it frequently works remarkably well in a variety of real-world situations.

In the process of learning this classifier with the given structure, class probabilities and conditional probabilities are computed using training data, and then the values of these probabilities are utilized to classify fresh observations. After extracting the featured with the help of HCTO, Naïve bayes has been implemented.

All above-mentioned approaches use cancer images for data analysis. The complexity and size of such an algorithm can be large due to the large number of image datasets. Furthermore, image classification can only provide a route approximation and it lacks accuracy.

The AI based techniques used in the past use state-of-the-art AI tools that are slow and get stuck at local minima. The proposed approach uses interactive data for analysis and classification of cancer.

The proposed approach is based on CTO which is fast and novel [102, 103, 167-168]. To the best of our knowledge, HCTO has not been used in the past for cancer classification. The proposed work is compared with tradition cancer classifier such as KNN and it has a better performance.

4.3. PROPOSED FRAMEWORK

This paper presents a novel approach for cancer classification. The proposed algorithm operates in various steps as explained below.

4.3.1 Dataset – The first step of the proposed work is data aggregation, the data for cancer patients were taken from Kaggle, and the 22 microarray datasets for 4 types of cancers were acquired from the NCBI Gene Expression Omnibus (GEO) database. The paper performs the classification of lung, kidney, skin, and bladder cancer. In the within-dataset experiment, 9 datasets were used. In cross-dataset studies, the other 13 datasets were utilized as validation.

Patients who died within 5 years of diagnosis were classified as negative samples in the cancer dataset, whereas those who survived were classified as positive samples. The KEGG database was used to get all route information. The data of liver, kidney, breast, and bladder cancers are used for feature selection. Once the feature selection has been done, the model is trained using the 4 cancer datasets.

Testing of the classification is performed and finally the accuracy of the classification is calculated and plotted. The ratio considered for training and testing the dataset was 80:20.

The feature selection process is optimized by HCTO and the classification is done using naïve Bayes classifier.

The implementation is as follows. The hyper-parameters are selected based on the requirements for performance optimization of the data analysis. Only those parameters are selected that contribute to performance enhancement.

4.3.2 Implementation – The dataset is cleaned and feature selection is done using HCTO. The Class Topper Optimization is a fast-converging simple artificial intelligence based algorithm that has smooth convergence. The learning behavior of kids in a certain class of a school inspired this optimization technique. Students in the same class compete to be the best student or class topper in their respective classes. This is a metaheuristic optimization technique based on population. The global optima are attained here by learning at each level. The suggested method ensures that students' learning progresses towards the global optimum point (topper).

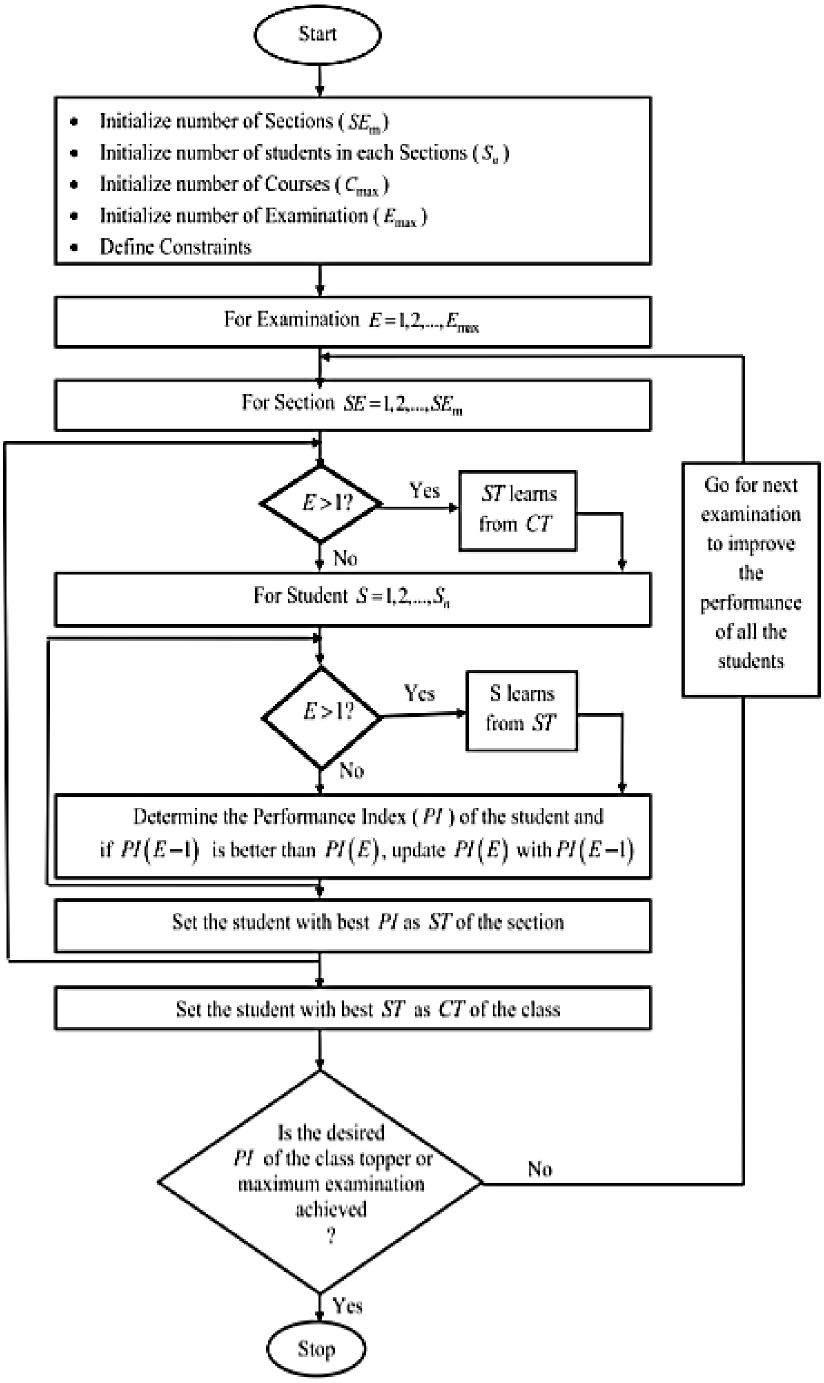
Start
Initialize number of Sections (SE_m)
Initialize number of students in each Sections (S_n)
Initialize number of Courses (C_{max})
Initialize number of Examination (E_{max})
Define Constraints
For Examination $E=1,2,...,E_{max}$
For Section $SE=1,2,...,SE_m$
$E>1?$
Yes
ST learns from CT
No
For Student $S=1,2,...,S_n$
$E>1?$
Yes
S learns from ST
No
Determine the Performance Index (PI) of the student and if $PI(E-1)$ is better than $PI(E)$, update $PI(E)$ with $PI(E-1)$
Set the student with best PI as ST of the section
Set the student with best ST as CT of the class
Is the desired PI of the class topper or maximum examination achieved ?
No
Yes
Go for next examination to improve the performance of all the students
Stop

Fig. 4.1(a) – A flowchart of Class Topper Optimization [210].

In Fig. 4.1(a) shows a graphical depiction of the phases. Let us take a quick look at learning at various levels of education. There are several classes in a school. There are several parts in a lesson. The idea behind dividing a class into parts is to create small groups. One portion of the class is undoubtedly home to the class's topper.

The information transferred from the class topper to the pupils is restricted to the section in which the class topper is a member. However, because the STs are the most knowledgeable members of the section, they will compete with CT and strive to learn from him. Students' performance will increase as a result of their education. Level of Section (SE), A class's section is a subgroup.

Each section's students are in the same class. The section topper is the best student in the class (ST). He might or might not be the class's top student. Apart from the section topper, there is one restriction for students: they can only learn from their own ST. The performance of the pupils in all areas may improve after each test and evaluation procedure. ST may change as a result of this.

A student should ideally learn from the best student or section topper. Learning from the section's best students has an influence on the section's other pupils. Each student improves as a result of this following assessment (evaluation of PI). The modeling of HCTO is shown in equations (1) to (4).

$$\mathrm{I}^{(\mathrm{S,E+1})} = \mathrm{I}_{\mathrm{WF}} * \mathrm{I}^{\mathrm{SE}} + \mathrm{c} * \mathrm{n}_2 * (\,\mathrm{ST}_{\mathrm{pi}}^{(\mathrm{SE,E})} - \;\mathrm{S}_{\mathrm{pi}}^{(\mathrm{S,E})}\,) \quad (1)$$

In equation (1), I stands for improvement which is being calculated, ST is a section topper & SE defines number of sections. Wf represents weight factor and C is the acceleration coefficient whose value is 2. Pi represents performance index.

$$\mathrm{S}_{\mathrm{pi}}^{(\mathrm{S,E+1})} = \mathrm{S}_{\mathrm{pi}}^{(\mathrm{S,E})} + \mathrm{I}^{\mathrm{SE+1}} \qquad (2)$$

In equation (2), student performance has been calculated with the improvement performance in comparison of section topper. S_{pi} depicts the performance of the student.

$$\mathrm{I}_1^{\mathrm{SE,E+1}} = \mathrm{I}_{\mathrm{WF}} * \mathrm{I}_1^{\mathrm{SE,E}} + \;\mathrm{c} * \mathrm{n}_1 * \left(\mathrm{CT}_{\mathrm{pi}}^{(\mathrm{SE,E})} - \mathrm{ST}_{\mathrm{pi}}^{(\mathrm{SE,E})}\right) \qquad (3)$$

In equation (3), CT depicts class topper and performance of the class topper has been evaluated over section topper. C, n are constants. I_1^{SE} represents increased understanding of ST in section SE of the Eth test and CT $_{pi}$ ^{SE,E} represents Performance index of all section's class topper in Eth examination

$$\mathrm{ST}_{\mathrm{pi}}^{(\mathrm{SE,E+1})} = \mathrm{ST}_{\mathrm{pi}}^{(\mathrm{SE,E})} + \mathrm{I}_1^{\mathrm{SE,E+1}} \qquad (4)$$

The $\mathrm{I}_{\mathrm{WFmax}}$ for HCTO is shown in equation 5. This also contributes the algorithm objective.

$$\mathrm{I}_{\mathrm{WF}}^{\mathrm{E}} = \mathrm{I}_{\mathrm{WFmax}} - \left(\frac{\mathrm{I}_{\mathrm{WFmax}} - \mathrm{I}_{\mathrm{WFmin}}}{\mathrm{E}_{\mathrm{max}}} * \mathrm{E}\right) \qquad (5)$$

Each student will be given the opportunity to improve after taking an examination as part of our evaluation process. If the student does not perform better in a specific examination, his prior best score will be kept. Observation (E). Examination is the process of assessing each student's PI. If students' performance isn't up to par, they can improve their PI by taking exams. If a student's performance is determined to be lower than prior performance indices after the examination and assessment procedure, the student's previous best PI will be used to evaluate the student in the future.

Beginning with the second test, ST learns from CT, and students in each section learn from their respective STs. This kind of evaluation (search procedure) will undoubtedly result in the finest answer. As a result, the class topper’s PI may improve in subsequent exams. To increase student performance, a maximum number of examinations should be set. A framework of HCTO is shown in 4.1 (b) respectively.

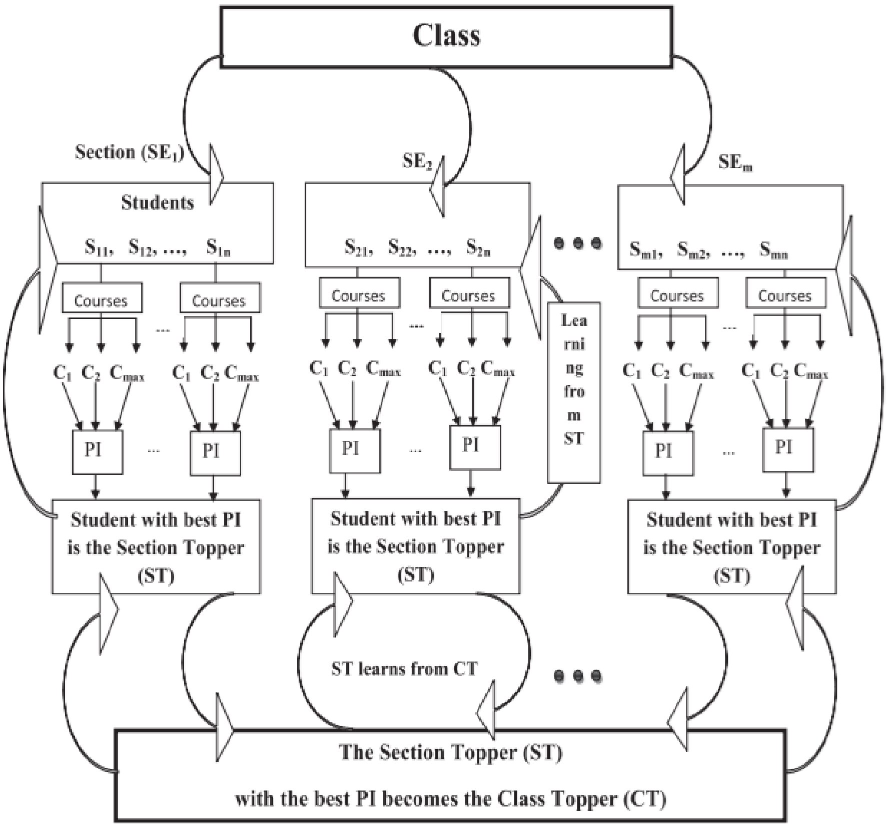
Class
Section (SE$_1$)
SE$_2$
SE$_m$
Students
S$_{11}$, S$_{12}$, ..., S$_{1n}$
S$_{21}$, S$_{22}$, ..., S$_{2n}$
S$_{m1}$, S$_{m2}$, ..., S$_{mn}$
Courses
C$_1$ C$_2$ C$_{max}$
PI
Learning from ST
Student with best PI is the Section Topper (ST)
ST learns from CT
The Section Topper (ST) with the best PI becomes the Class Topper (CT)

Fig. 4.1 (b) – Framework of HCTO.

4.3.3 Training, Testing and Classification – After the feature extraction process is done using HCTO algorithm. Training, testing and classification is done. For this purpose, Naïve Bayes classifier has been used. Classification is a machine learning and data mining technique that is widely utilised.

Different ways to performing classification work may be used depending on the number of target categories utilised to categorise a data collection. Decision trees and support vector machines are widely used for binary classifications; however, these two techniques are constrained by the fact that the number of target categories cannot exceed two.

Because of this strict constraint, they are difficult to apply to a broader sense of real-world classification activity, where the number of target categories is generally greater than two. Weather prediction services, consumer credit assessments, health condition categorizations, and other real-world applications based on the NB classifier have all proven effective.

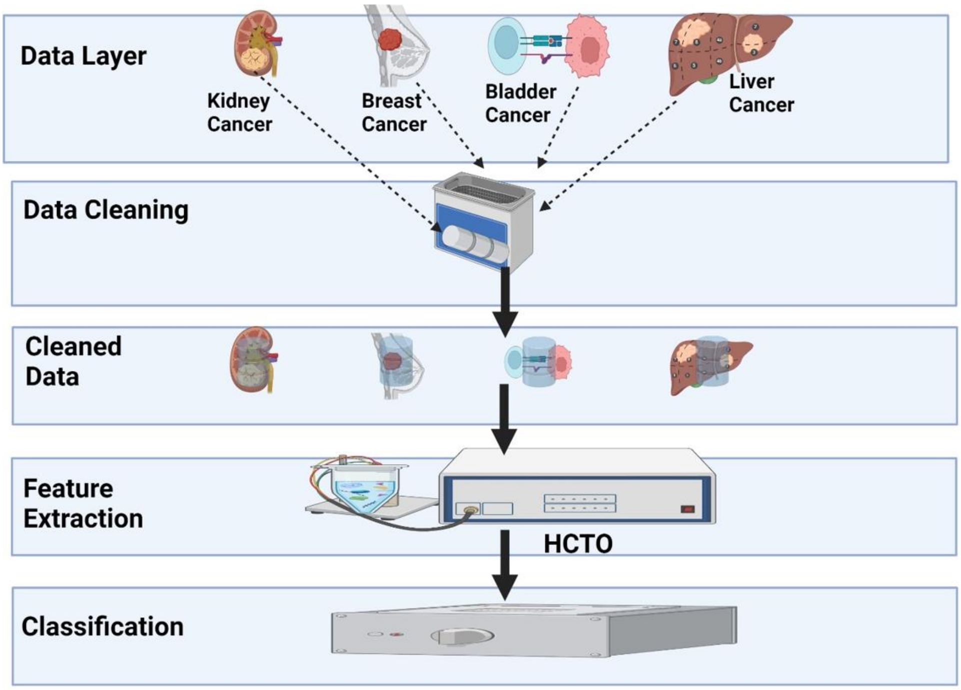

Fig. 4.2 – Model Framework-1

As long as the data set's format is pre-processed into a tabular format inside the issue domain. The validity of fitting a fresh piece of data into each potential categorization may then be computed using this mathematical classifier.

The NB Classifier is a probabilistic classification system based on Thomas Bayes' posthumous Bayesian Theorem. The basic objective of classification is to discover the optimal mapping between fresh data and a collection of categories within a certain problem area. Some mathematical manipulations are made to translate joint probabilities into multiplications of prior probabilities and conditional probabilities so that this mapping may be probabilistically computed. The implementation of this methodology can be easily understood with the help of the model framework Fig. 4.2.

4.4. RESULTS AND ANALYSIS

In this part, performance of the proposed classifier has been evaluated. The proposed approach classifies different types of cancer based on HCTO and naïve bayes artificial intelligence algorithms. The dataset used is explained in the above steps in the paper; training and testing are performed. In cross-dataset research, the initial dataset was used as the training set, while other independent datasets were utilized as the test set. The classification accuracy is shown in Fig. 4.3.

The classification accuracy of Naïve bayes tends to grow monotonically with the number of features. It demonstrates that the NB has low interpretative capabilities since no feature selection is performed; the NB is able to achieve high accuracy when multiple characteristics that are important to the answer are used. Its accuracy increases with increasing iterations.

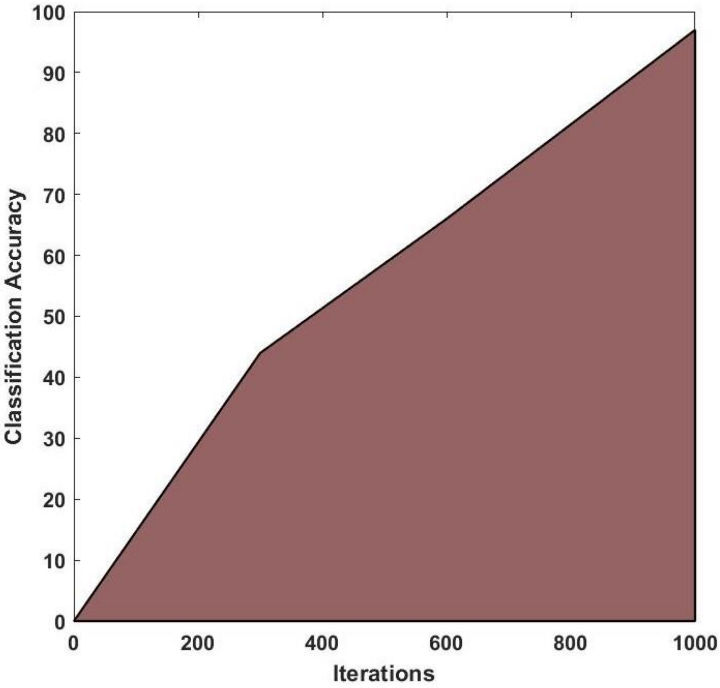

Fig. 4.3 – Accuracy of the proposed classifier.

The results of the NB and KNN classification errors on the dataset are shown in Fig. 4.4. The proposed algorithm, has lower error rate as compared to KNN. The HCTO optimizes the entire process. It is seen that the error rate decreases with number of iterations. NB and HCTO classify the cancer types with more accuracy and less error. This performance improvement is attributed to the fact that the NB classifier has a performance gain over KNN. The process is faster and less complex due to better convergence rate of HCTO as compared to other algorithms.

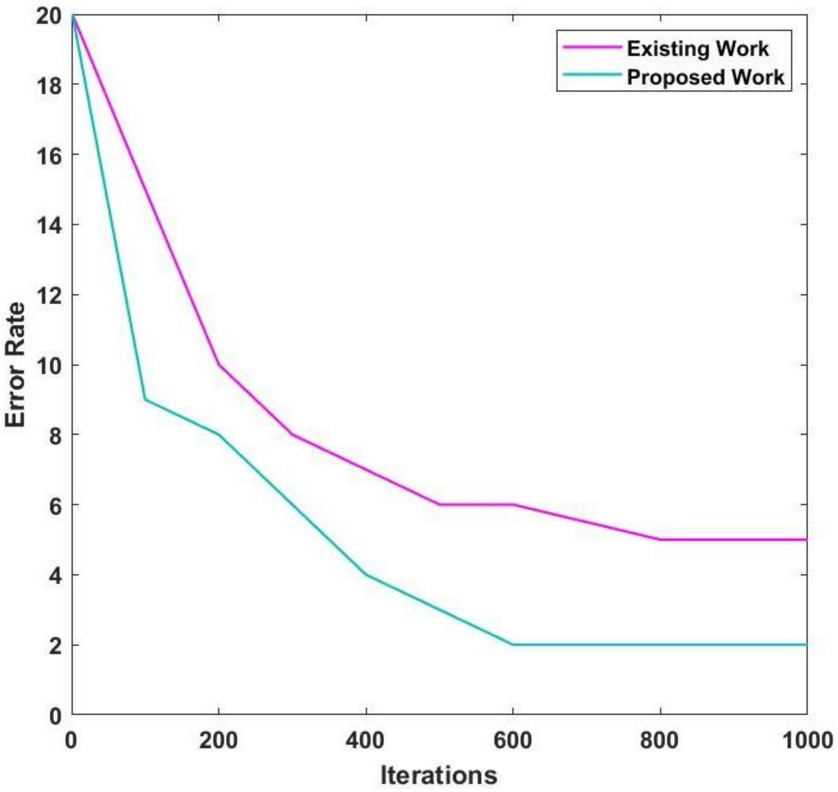

Fig. 4.4 – Error rate for the proposed classifier

The precision is calculated by dividing the number of true positives by the number of true positives plus the number of false positives. Precision is very important in any classifier. In detecting cancer, precision decides the type of cancer and its severity. Accordingly, treatment can be planned. If a classifier has low precision, it can result in wrong cancer diagnosis. In Fig. 4.5, the suggested work is assessed for precision, and it can be observed that the proposed classifier's precision rises as the number of iterations increases.

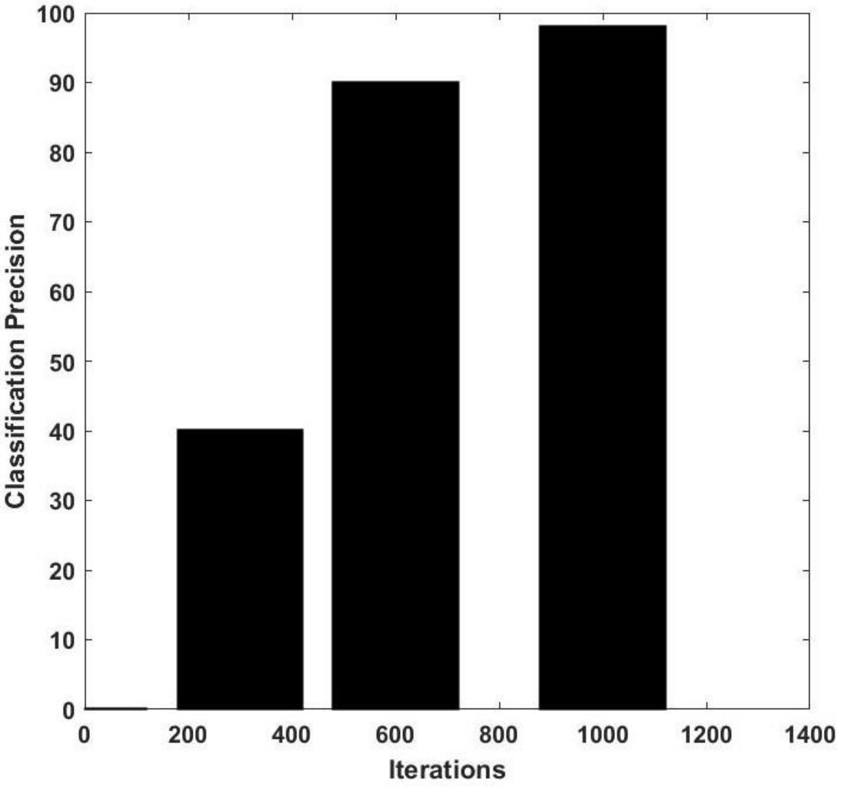

Fig. 4.5 – Precision performance for the proposed classifier

The performance of the classifier is measured in Fig. 4.6. The performance is measured in terms of how efficiently and accurately the proposed framework can classify different cancer types. It is seen that the proposed HCTO-NB artificial intelligence approach has better performance than other AI based classifier such as KNN.

The performance of the proposed framework increases with rising iterations. This is due to the fact that the optimization algorithm HCTO approaches convergence with increasing number of iterations.

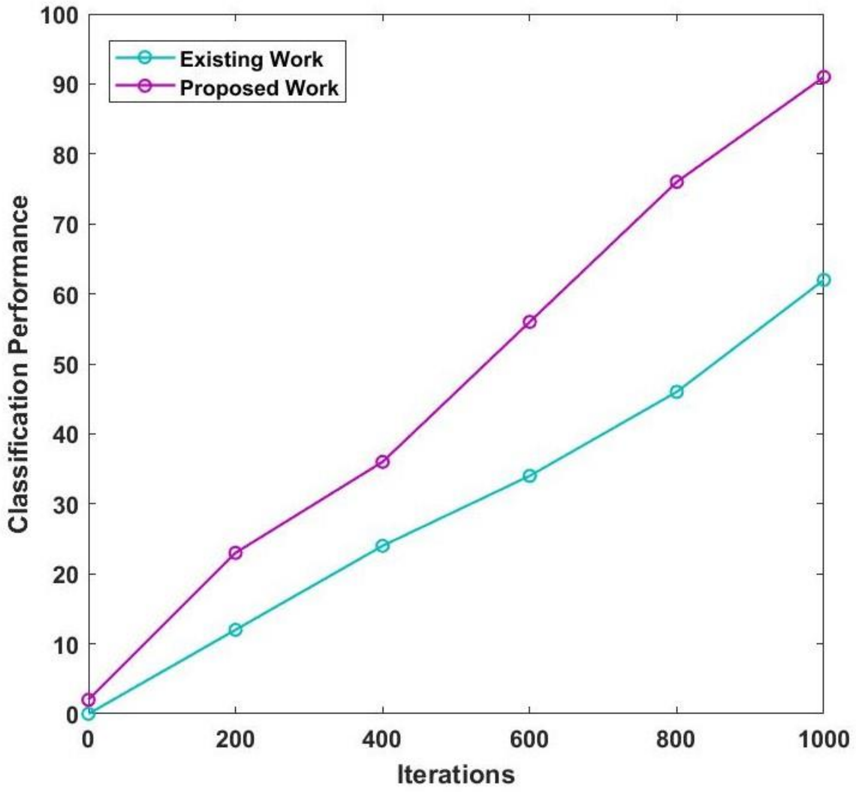

Fig. 4.6 – Classification Performance

Table 4.1 Computational paradigm of Proposed Approach

Computational Paradigm	HCTO+NAÏVE BAYES	KNN
Accuracy	97.6%	95.8%
Precision	98.4%	97.2%
Error Rate	Decreased by 3%	-

Table 4.2 Accuracy analysis between the HCTO-NB and state-of-the-art cancer Classification algorithms

Algorithm	Data Set	Type of dataset	Accuracy	Sensitivity	Specificity
CNN	Kaggle	mammogram images	82.71	82.68	82.73
ML DL	MIAS	mammogram images	91.5	92	84.2
GOA-CNN	MIAS	mammogram images	92	96	93

CNN & Softmax algorithm	MIAS	Mammogram & Histopath images	65	-	-
Annotation based deep learning approach	MIAS	Mammograph Images	-	96.2	90.9
Transfer Learning	ICIAR	Histopath Images	82.90	-	-
Deep Transfer Learning	IDC	Histopath Images	91.37	-	-
WSLM	TCGA	Histopathology Images(Whole slide images)	97.3	-	-
LDNNET	Kaggle	CT images	0.98	0.99	0.98
SVM	TCGA	Images based	9.25	6.45	10.11
RELIEF-SVM	LUSC-ASC	Histopath images	73.91	-	-
DFNForest	Kaggle	Feature based data	93.6	-	-
TL	TCGA	Histopath images	0.98	-	-
DFST	GSE40419	Microarray/Images	0.96	-	0.96
ACGAN-Inception-v4	SqueezeNet	Images based	95	94	97
HCTO	Kaggle-UCI	Microarray/Interactive data	97.6	-	98.4

CHAPTER 5

A Heuristic Machine Learning based Optimization Technique to Predict Lung Cancer Patient Survival

5.1. Introduction

Cancer is the second largest cause of mortality in the United States and a major public health issue globally, with 8.8 million deaths reported in 2015 [177]. Changes in nuclear structure have remained the ‘gold standard' for cancer detection for over 150 years [178] because of the strong link between abnormal nuclei morphology and tumour development.

Nuclear morphology can be visually evaluated under light microscopy with cells stained with reagents (e.g., Hematoxylin and Eosin) in clinical diagnosis [106, 179]. Many cancers have distinct nucleus changes that pathologists can carefully examine in order to make treatment decisions. Modifications in nuclear size, shape, nucleoli appearances, chromatin organisation, and so on are examples of morphological changes [180].

Nuclear structure is widely quantified in computer-aided diagnostic (CAD) systems by a collection of numerical features [181]–[182] that describe intrinsic morphological properties, such as nuclei size and shape features (e.g., perimeter, area, curvature, symmetry) and nuclear texture data (e.g., Haralick, Gabor, Wavelet).

TBM and newly suggested deep learning techniques [79, 85], which have been effectively utilised in nuclei identification [184]–[187], segmentation [100], and detection, are two alternative nuclear quantification methodologies. Pathologists rely on microscopic examination of a collection of nuclei inside a tissue sample for clinical diagnosis. As a result, in the vast majority of cases, a diagnostic label is only provided for the tissue sample, not individual nuclei.

The set detection issue requires a predictive model to learn from sets of nuclei without nuclei level annotations and predict the diagnostic label for a fresh set of nuclei. In addition to cancer diagnosis, the set detection challenge is common in prognosis prediction, where the model must estimate the patient's survival fate using a set of measured nuclei [190-191].

In contrast to traditional image detection, where training and testing samples are labelled single-shot pictures, training and testing samples in the set detection scenario are sets, each of which contains a different number of unlabelled nucleus images. The set detection problem is difficult to solve, and supervised machine learning approaches are ineffective. The following are examples of existing nucleus set detection systems and their limitations.

Many predictive models handle the set detection problem with single picture detection by making particular assumptions about the connection between the distribution of examples and the set label, however this is frequently done implicitly. Many studies use the majority voting approach in set prediction, assuming that at least half of the occurrences in a set represent the set label. Lung cancer, cervical cancer [101, 192], and breast cancer are only a handful of the diseases for which predictive models based on majority voting have been developed.

For hepatocellular carcinoma tumour grading, a threshold-based voting approach was used in [194]. However, for the greatest performance, the voting threshold for a set being classified into a certain class must be pre-defined based on domain expertise. A set is deemed positive in the multiple instance learning (MIL) framework [195] if it contains at least one positive instance; otherwise, it is considered negative. MIL is gaining popularity in medical diagnosis [196-197], but its validity in our situation is debatable since not all nuclei in the tissue sample display the morphological alterations associated with the illness [198]. The goal of single picture detection is to create an instance-level classifier using set-level labels, which infers. However, because to tumour heterogeneity [199], the choice of assumption frequently necessitates previous domain knowledge and has a substantial influence on the prediction model's overall performance.

Set detection takes into account the entire set of data and learns the prediction model at the set level. The concept of categorising nuclei groups rather than individual nuclei is not new. Aggregating several statistics (STATS) regarding nuclear feature characteristics inside a set is the most common and intuitive method [107, 200].

The nuclei set's properties are summarised in the feature vector, which includes statistics like as mean, maximum, minimum, and standard deviation. STATS' success is thus heavily reliant on the pre-designed statistics gleaned from experimental data. Another frequently used technique in set detection is Bag-of-words (BoW) [201-202], which learns a number of sample instances or dictionaries in the training set and then gives a histogram about the composition of one set considering the dictionary.

In feature space, unsupervised clustering, such as k-means, is commonly used to build a dictionary of cluster centres.

The main Objectives of this paper are as follows:

- Provide an efficient Feature selection technique using biomarker genes to find out whether a person will survive or not.
- Establish a new method with SSA, if a patient will survive, then the duration is more than five years or not.
- Design an effective technique to predict the overall survival time with lung cancer.

5.2. Literature Review

Because cancer is one of the most lethal diseases in the world, many researchers have focused on one of the most important cancer biomarkers, DNA methylation, and applied feature selection or feature extraction techniques to its massive data in order to improve prediction accuracy by obtaining the best features set that discriminate biological samples of different tumour types [203].

The study provided by Wu et al. [202] is an example of feature selection approaches being used in research. This work used a three-step feature selection technique based on the properties of clinical DNA methylation data, and the feature selection procedure chose many cancer-related and lymph node [98] metastasis-related gene biomarkers. The findings of the given approach revealed that the accuracy of prediction in identifying LN metastasis was much enhanced.

Furthermore, utilising probe level DNA methylation data, Baur and Bozdag developed a unique feature selection method based on sequential forward selection to calculate gene centric DNA methylation. The results of the proposed method using the K-Nearest Neighbors classification outperformed other algorithms on all metrics, and it was able to accurately predict the expression of particular genes using just DNA methylation data.

These DNA methylation-sensitive genes were also found to be overrepresented in Gene Ontology concepts relevant to the regulation of numerous biological processes. By picking just the beneficial characteristics from the full feature set, Kaur and Kalra [206] verified the relevance of using feature selection to forecast various illnesses such as breast cancer, lung cancer, heart disease, and so on.

The study compared the accuracy and efficiency of several feature selection methods, including F-score, Genetic Algorithm, K-means, ReliefF, and SVM-RFE, and found that SVM-RFE obtained the best accuracy of 97 percent when utilising support vector machine (SVM).

Furthermore, Singh and Sivabalakrishnan's review [213] demonstrated the importance of the feature selection technique in boosting classifier accuracy. This analysis revealed that each feature selection method behaves differently and has its own set of benefits and drawbacks. The research discovered a hybridization of the feature selection elimination technique and a machine learning algorithm based on Random Forest.

This study intended to develop a two-stage computer-aided diagnostic system to identify benign from malignant breast tumours, with the first stage of the proposed system performing a data reduction procedure in preparation for the second stage's learning algorithm.

The suggested technique of this study beat other studies in the test phase, with classification accuracy of 99.82 percent and 99.70 percent, respectively. Recent research, on the other hand, focused only on feature extraction approaches in order to speed up and increase prediction accuracy.

For example, authors [204] proposed a feature extraction approach called Optimal Mean based Block Robust Feature Extraction method (OMBRFE) to extract feature genes to predict colorectal cancer [69] illness based on gene expression data.

To minimise the high dimensionality of the data, the suggested feature extraction technique in this paper relied on singular value decomposition (SVD). The findings showed that the OMBRFE is a useful tool for identifying feature genes, and the extracted genes were found to be closely linked to advanced colorectal cancer in clinical stages.

Liu et al. [205] have developed a unique feature extraction technique for reliable illness prediction called iterative Pearson's correlation coefficient (iPcc).

By iteratively using Pearson's correlation coefficient, this study created a set of new characteristics for samples based on gene expression profiles. Despite the huge number of extracted features and the time it took to get them in this investigation, the number of extracted features was equal to the number of samples.

The following gaps were identified during the literature review and incorporated into this paper:

(a) The current work offers a fundamental SSA framework for low-dimension optimization issues that can be expanded to large-scale optimization and constrained optimization situation.

(b) In addition, multi-objective optimization issues may be solved using SSA. The suggested approach may also be used to resolve NP-hard real-world combinatorial optimization issues.

5.3. Proposed Approach

5.3.1 Squirrel search algorithm

When flying squirrels start foraging, the search process begins. Squirrels hunt for food resources by gliding from one tree to the next during the warm season (fall). They switch locations and explore different parts of the forest while doing so. Because the climatic circumstances are hot enough, they can satisfy their daily energy demands more rapidly by eating acorns, which are plentiful, and they eat acorns as soon as they discover them. They begin looking for the best food source for the winter when they have met their daily energy requirements (hickory nuts Storage of hickory nuts will aid them in meeting their energy needs in adverse weather, reducing the need for costly foraging excursions) and therefore increasing their chances of survival.

In deciduous woods, a decrease of leaf cover during the winter increases the risk of predation, thus they become less active but do not hibernate. Flying squirrels become active again at the conclusion of the winter season. This is a continuous process that lasts the whole lifespan of a flying squirrel and is the basis of SSA. The following assumptions are taken into account when simplifying the mathematical model:

1. In a deciduous forest, there are n flying squirrels, and each squirrel is considered to be on one tree.
2. Each flying squirrel seeks for food on their own and makes the most use of the food resources available to them by showing a dynamic foraging behaviour.

3. There are only three sorts of trees in the forest: regular trees, oak trees (food source for acorn nuts), and hickory trees (hickory nuts food source).
4. Three oak trees and one hickory tree are believed to be present in the woodland region under examination. The number of squirrels in this study, n, is taken to be 50. With 1 hickory nut tree and 3 acorn nut trees, 4 nutritious food resources (Nfs) are evaluated, whereas 46 trees have no food source. That is, 92 percent of the entire squirrel population is on regular trees, with the rest relying on food sources. However, given one ideal winter food source, the number of food resources can be changed according to the restriction 1 Nfs n where Nfs Z>0.
5. A vector identifies the position of a flying squirrel in a d-dimensional search space. With the ability to change their location vectors, flying squirrels are able to glide across one-dimensional and two-dimensional search space. The following diagram depicts the SSA process.

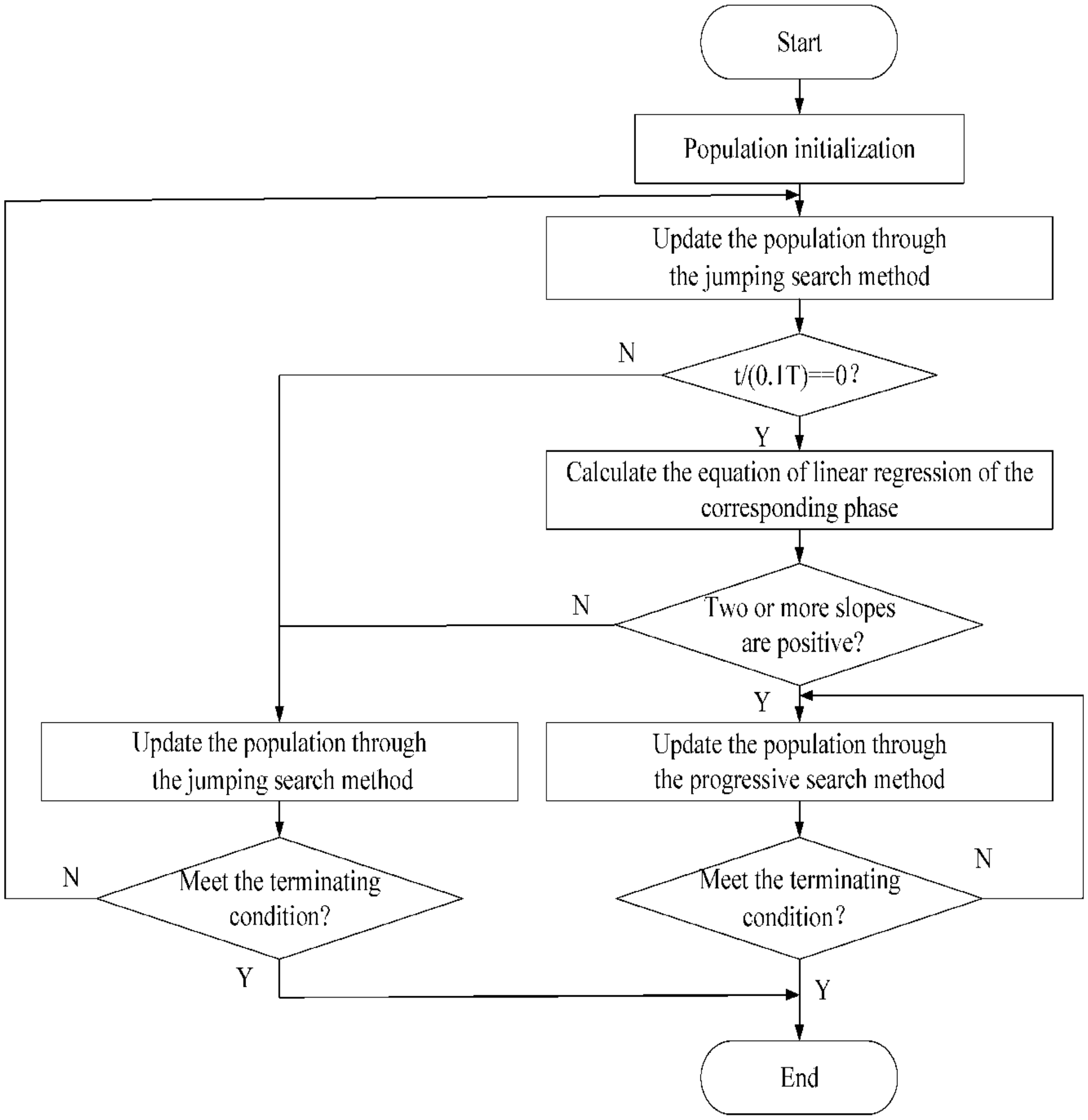
Start
Population initialization
Update the population through the jumping search method
N
t/(0.1T)==0?
Y
Calculate the equation of linear regression of the corresponding phase
N
Two or more slopes are positive?
Y
Update the population through the jumping search method
Update the population through the progressive search method
N
Meet the terminating condition?
Meet the terminating condition?
N
Y
Y
End

Fig. 5.1 – SSA implementation method [210].

Pseudo code for SSA

```
Begin:
Define the input criteria
Random positions for n number of floating squirrels using Eq. (1)
Calculate fitness of each floating squirrel’s position
Sort the positions of floating squirrels in increasing order based upon fitness value
Announce floating squirrels on hickory normal trees, acorn trees and nut tree
At Random elect some floating squirrels moving from normal trees t hickory nut tree
and the rest will move facing acorn trees
```

```
while(the stopping requirement is not met)
    For t=1 to n_1 (n_1=total floating squirrels coming towards hickory nut tree from acorn trees)
        if $R_1 \geq P_{dp}$
            $FS_{at}^{t+1} = FS_{at}^{t} + dg \times Gc \times (FS_{ht}^{t} - FS_{at}^{t})$
        else
            $FS_{at}^{t+1}$ = a random location of search area
        end
    end

For t=1 to n_2 (n_2=total floating squirrels on normal trees travelling in the direction of acorn trees)
        if $R_2 \geq P_{dp}$
            $FS_{nt}^{t} = FS_{nt}^{t} + dg \times Gc \times (FS_{at}^{t} - FS_{nt}^{t})$
        else
            $FS_{nt}^{t+1}$ = a random location of search area
        end
    end

For t=1 to n_3 (n_3=total floating squirrels on normal trees travelling in the direction of hickory nut tree)
        if $R3 \geq P_{dp}$
            $FS_{nt}^{t+1} = FS_{nt}^{t} + dg \times Gc \times (FS_{ht}^{t} - FS_{nt}^{t})$
        Else
            $FS_{nt}^{t+1}$ = a random location of search area
        end
    end

    Evaluate seasonal constant (Sc) using equation (7)
        if (condition for Seasonal monitoring  is met)
    Randomly repositioned floating squirrels
        end
```

```
Update the lowest value of seasonal constant
End
The position of squirrel on hickory tree is the concluding best solution
End
```

5.3.2 Implementation of SSA, like other population-based algorithms, starts with a random beginning position of flying squirrels. In a d-dimensional search space, the location of a flying squirrel is represented by a vector. As a result, flying squirrels may modify their position vectors and glide through 1-D, 2-D, 3-D, or hyper-dimensional search space.

5.3.2.1 Dataset. The dataset used in this paper is imported from Kaggle and other online resources including Wisconsin Prognostic Lung Cancer sub-directory, which over 100 occurrences. The cancer cell nuclei are characterized in terms of their individual properties such as radial distance (mean of distances from the centre to points on the perimeter), opacity (standard deviation of grey scale values), distance from the ground, location, ease of use (local variation in radius lengths).

The degree of compactness is measured in convexity, rounded corners, synchronicity. Findings are the average, standard error, and "worst," or worst (the mean of the three largest values),". One lung cancer patient's follow-up data is included in each record.

5.3.2.2. Algorithm Descriptions for Classification Algorithms: To test their accuracy in predicting lung cancer survival, researchers used the Lung Cancer Dataset to compare three well-known classification algorithms for the prediction model: naive Bayes, quick decision tree learner, and K-nearest neighbour. Algorithms are described in the following paragraphs.

5.3.3.3 An Algorithm by Naive Bayes: Both supervised learning and statistical classification are represented by the Bayesian classification approach. Because it takes a probabilistic model as a foundation, may use probabilities to quantify uncertainty in the model. It is capable of diagnosing and forecasting issues. According to Thomas Bayes (1702-1761), the Bayes theorem is named after this categorization.

The practical learning methods provided by Bayesian classification integrate previous knowledge and observed data. Many learning algorithms may be better understood and evaluated from a Bayesian classification viewpoint. In addition to calculating the probability of a hypothesis, it is resistant to noise in the data that is inputted into it.

5.3.3.4. Quick Decision-Making Algorithm for Tree Learners: "Decision trees learning method based on the premise of calculating the information gain with entropy" and "minimising the error coming from variance" is the reduced error pruning (REP) tree classifier [100]. Iterations of REPTree use regression tree logic to produce numerous trees. Finally, it chooses the finest generated tree among the others. This approach uses variance and information gain to build a regression/decision tree. Additionally, this approach uses a backfitting method to prune the tree.

5.3.3.5. K-Neighbor's Algorithm: A technique known as K-nearest neighbours (KNN) is a supervised classification procedure in which the k-nearest neighbours of a point are selected, determined by minimising a similarity measure (e.g., the Euclidean distance or Mahalanobis distance).

KNN computes the distance of an unlabelled example to the other (labelled) instances and calculates the k-nearest neighbours and their associated labels to identify the class of the sample. If the unlabelled item isn't named, it is then categorised either by majority voting (the dominant class in the area) or by a weighted majority, in which points closer to the unlabelled object are given higher weight.

5.3.3.6 Algorithms for selecting features: Before putting data into a classifier, classification requires careful examination of the dataset. To make the categorization process simpler, it is recommended that just essential characteristics be considered, rather than a large number of useless features. This is why having a wide range of approaches for identifying the most important and relevant characteristics is useful.

Classifiers are also more accurate when they use feature selection to locate the most important feature and reduce their burden, which also increases classification accuracy. SSA outperforms other current feature selection methods in terms of classification accuracy. The following are some of the possible benefits of using SSA for feature selection: Various candidate solutions are able to explore different areas of solution space in order to find the best possible solution.

Due to the fact that SSA's solution has a memory and can retain information about the solution as it travels around the problem space, it makes an excellent feature selection tool. Its computationally low-cost implementation, while yet delivering acceptable performance, contributes to SSA's appeal.

- Rather of focusing on a single answer, SSA considers a wide range of options.
- SSA is able to handle binary and discrete data sets.
- In terms of memory and runtime, SSA is superior to other feature selection methods.

Using SSA is quick and straightforward, and the results are encouraging. SSA's effectiveness is virtually completely unaffected by the problem's size.

Considering the population is N and Upper bound in the search space is represented by FSu whereas lower bound has been represented by FS_i. FSi depicts the population and i ranges from 1 to N. D represents dimensions and rand represents random number. With the help of equation (1) population is initialized.

$$FS_i = FS_l + rand(1, D) * (FS_u - FS_l) \qquad (1)$$

Equation (2), (3) & (4) is used to identify the position of the squirrel, whether it is on hickory tree, oak tree or regular tree and it can be done with the help of

$$FS_{at}^{t+1} = FS_{at}^{t} + dg \times Gc \times (FS_{ht}^{t} - FS_{at}^{t}) \quad \text{if } \mathrm{R}_1 > \mathrm{P}_{\mathrm{dp}} \qquad (2)$$

$$FS_{nt}^{t} = FS_{nt}^{t} + dg \times Gc \times (FS_{at}^{t} - FS_{nt}^{t}) \quad if\ \mathrm{R}_2 > \mathrm{P}_{\mathrm{dp}} \qquad (3)$$

$$FS_{nt}^{t+1} = FS_{nt}^{t} + dg \times Gc \times (FS_{ht}^{t} - FS_{nt}^{t}) \quad if\ \mathrm{R}_3 > \mathrm{P}_{\mathrm{dp}} \qquad (4)$$

Here R is a random variable which lies in between 0 and 1 whereas Pdp depicts predator probability of appearance. If r > Pdp it means predator will not appear and vice versa, t depicts the current cycle & Gc is 1.9. FSat represents floating squirrels on acorn tree, FSnt represents floating squirrels on normal tree and FSht represents floating squirrels on hickory tree.

In equation (5), d$_g$ is the floating space which can be calculated with the help of

$$d_g = \frac{h_g}{\tan(\varphi) * \ sf} \tag{5}$$

In equation (6), hg and sf depict constant value which is 8 and 18 respectively. Now $\tan(\varphi)$ which is gliding angle will be calculated as

$$\tan(\varphi) = D/L \tag{6}$$

Where D is the pull strength and L is the lift strength.

Equation (7) is used to calculate seasonal constant Sc

$$S_c^t = \sqrt{\sum_{k=1}^{d} (FS_{at,k}^t - FS_{ht,k})^2} \tag{7}$$

Some of the advantages of selecting features with SSA include the following: In order to discover the greatest potential solution, various candidate solutions might explore different sections of solution space. SSA's solution is an outstanding feature selection tool because it has a memory and can keep knowledge about the solution as it moves across the issue space. Because of its computationally low-cost implementation and good performance SSA has become a popular choice for many businesses.

As opposed to concentrating on a single response, the Social Security Administration evaluates a broad variety of possibilities. SSA is capable of working with both discrete and binary data. SSA is more efficient in terms of memory and performance than other feature selection approaches. SSA is easy to use, and the results are promising. The scale of the issue has no bearing on SSA's efficacy.

The implementation of this methodology can be easily understood with the help of the model framework Fig. 5.2. It explains that dataset of lung cancer has been taken initially and data preprocessing has been performed to clean the dataset. After cleaning, Features have been extracted with the help of SSA and on this Naïve bayes has been implemented to find the optimal solution whether the lung cancer patient could survive for more than 5 years or not.

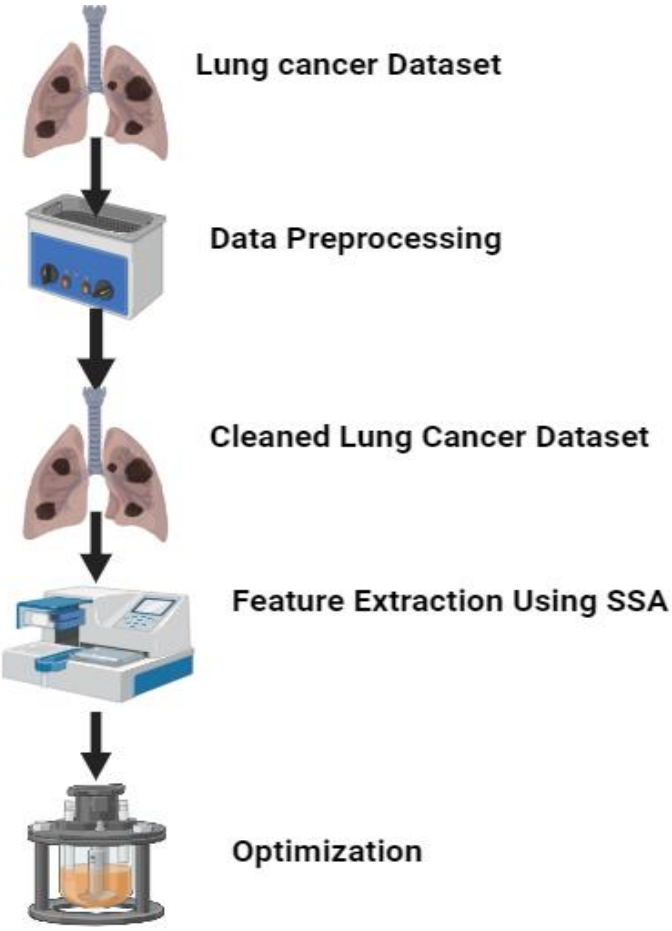

Fig. 5.2 – Model Framework-2

5.4. Experimental Results

The dataset is also available at https://archive.ics.uci.edu/ml/datasets/lung+cancer [212][access time – 17 Jan 2021] , which was randomly split into two sets: a training set and test set in the proportion 80:20. Experiments on each dataset were conducted five times in order to ensure the fairness and robustness of the research methodology.

Fig. 5.3 shows the error percentage value of the proposed work in comparison to the existing algorithm. It is seen from Fig. 5.3 that error rate decreases with increasing number of iterations. The number of iterations has been taken as an input on the X- axis from 0-1000, while error value has been taken on the Y-axis. As illustrated, the proposed hybrid approach attained 0.3 less error rate than other existing methods. This is due to the optimization of SSA.

At each step, proposed algorithm has a better performance as compared to the existing algorithm.

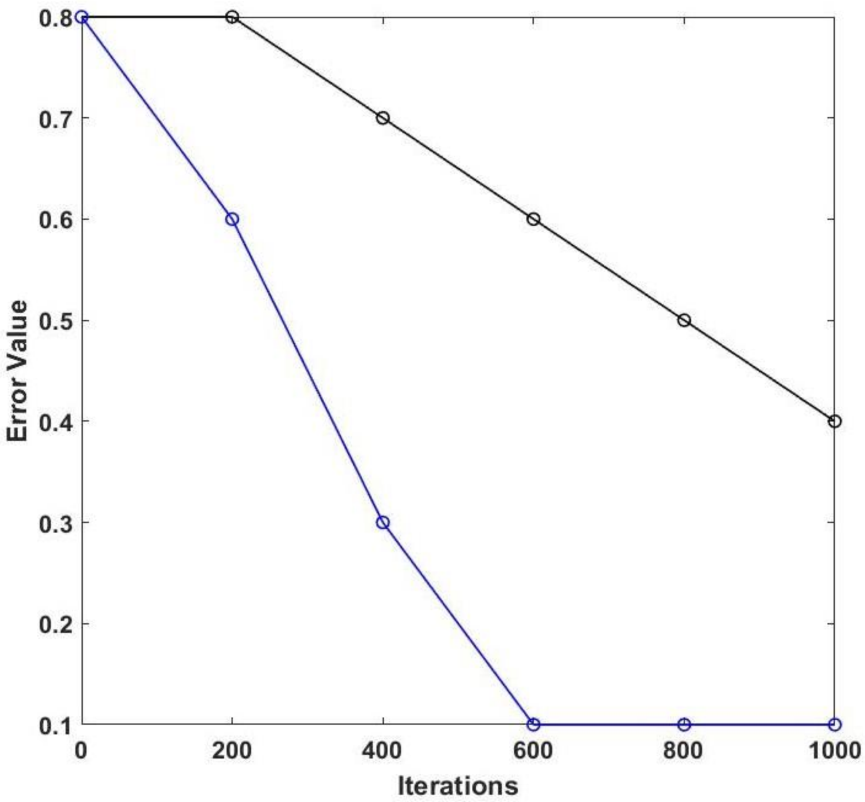

Fig. 5.3 – Error rate comparison

Table 5.1 Error rate Comparison

Error Comparison		
Rounds	Random Forest	SSR
0	0.8	0.6
200	0.8	0.3
400	0.7	0.1
600	0.6	0.1
800	0.5	0.1
1000	0.4	0.1

It is shown in Fig. 5.4 that the suggested approach is more accurate than the current method. Increasing the number of iterations leads to an improvement in accuracy. A large part of this may be attributed to SSA's improved performance. Comparing the suggested method to the current one, it is better at each stage and achieved better accuracy in comparison with existing method.

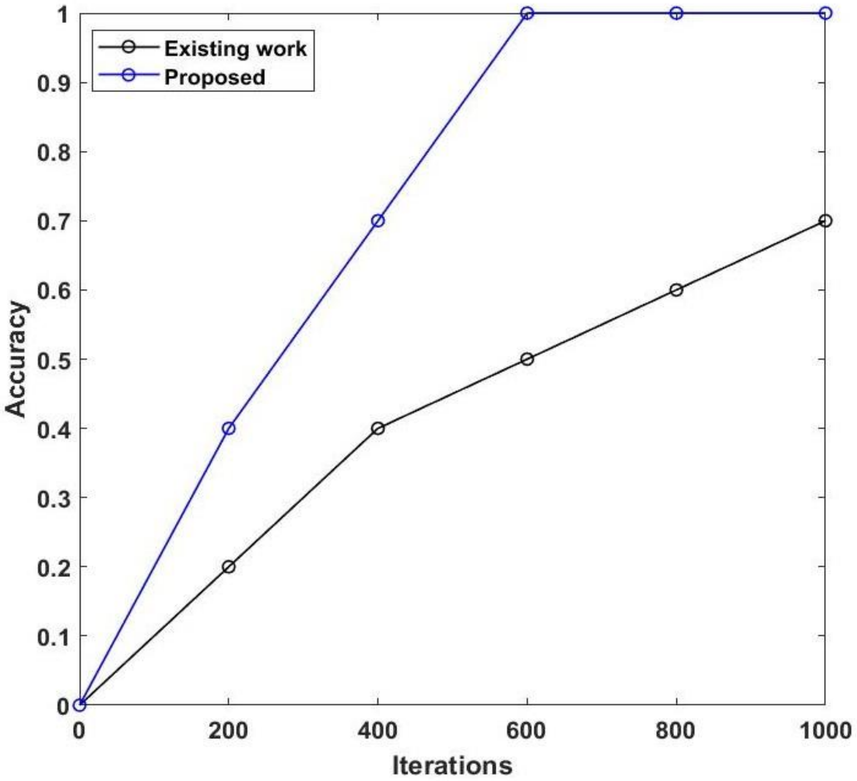

Fig. 5.4 – Accuracy comparison of the proposed work.

Table 5.2 Accuracy Comparison

Accuracy		
Rounds	Random Forest	SSA
0	0	0
200	0.2	0.4
400	0.4	0.7
600	0.5	1
800	0.6	1
1000	0.7	1

True positive rate comparison of the proposed work is shown in Fig. 5.5. It is evident that the true positive rate shows a gradual increase with number of rounds. It shows a sudden rise at 600 rounds. The proposed approach has a better true positive rate in comparison to the past approach.

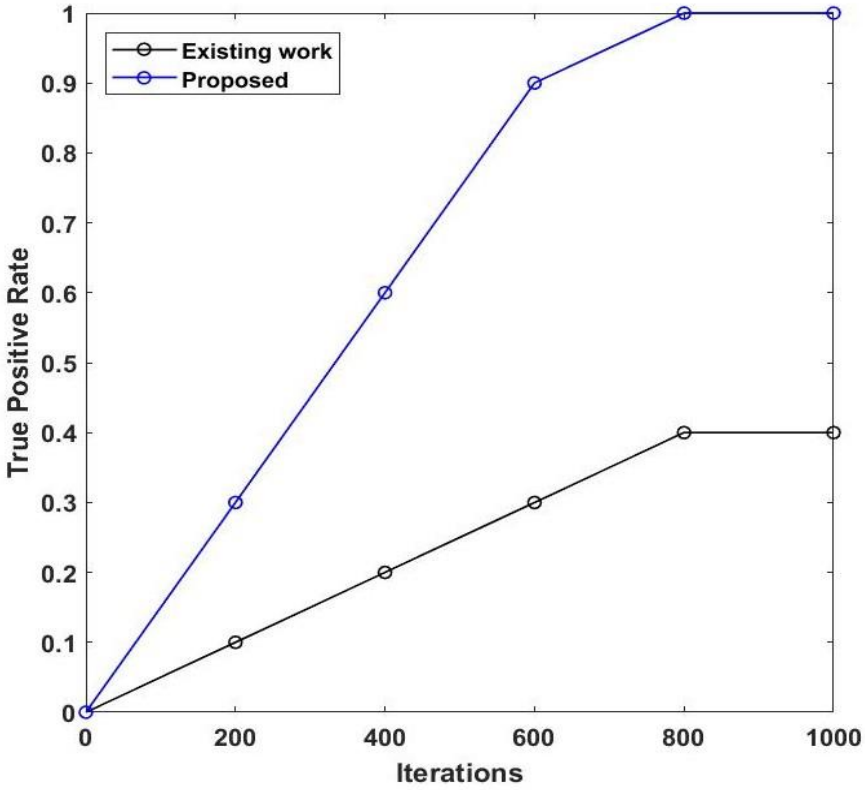

Fig. 5.5 – True Positive rate of prediction

Table 5.3 True Positive Rate

True Positive Rate		
Rounds	Random Forest	SSR
0	0	0
200	0.1	0.3
400	0.2	0.6
600	0.3	0.9
800	0.4	1
1000	0.4	1

In the similarity network fusion process of two malignancies, quick convergence may be achieved, however in order to meet the criterion for iterating termination, 1,500 times of iterations are required.

The precision and recall are crucial parameters in judging the performance of a prediction model. Precision and recall values of the proposed approach are shown in Fig. 5.6 and Fig. 5.7 respectively.

It is clear that as the number of iterations increases, so do the precision which is increased. However, the suggested technique outperforms the current strategy in terms of accuracy and recall. This is due to the application of SSA. It is evident that the Precision rate shows a gradual increase with number of rounds.

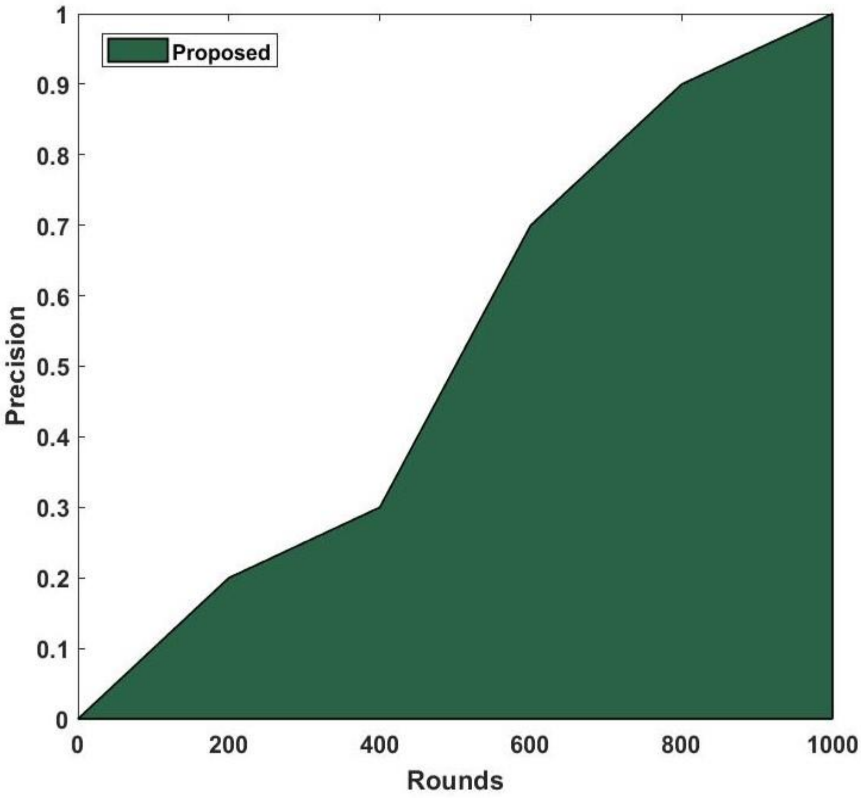

Fig. 5.6 – Precision value of the proposed approach.

Table 5.4 Precision

Precision	
Rounds	SSR
0	0
200	0.2
400	0.3
600	0.7
800	0.9
1000	1

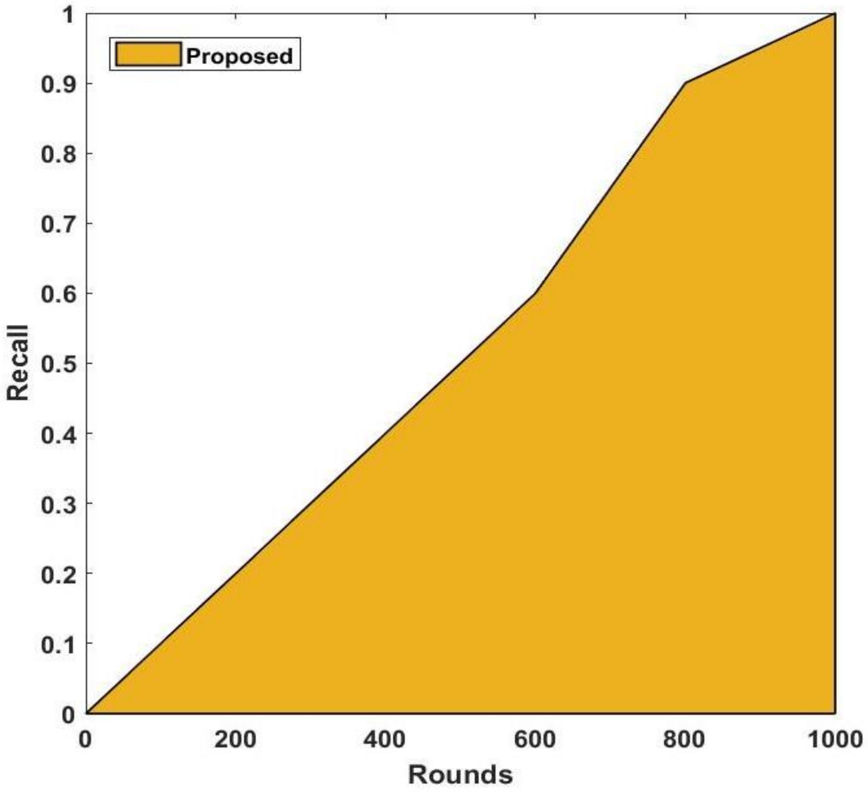

Fig. 5.7 – Recall of the proposed algorithm

Table 5.5 Recall

Recall Value	
Rounds	SSR
0	0
200	0.2
400	0.4
600	0.58
800	0.9
1000	1

CHAPTER 6

Conclusion & Future Scope

The thesis presents a novel approach for cancer classification. An extensive dataset has been considered for various cancers and classify them using Naïve Bayes classifier. The feature extraction is done using HCTO algorithm. The HCTO algorithm is a new and fast converging artificial intelligence tool.

The proposed approach is simple, less complex, accurate and have a very low error percentage, which is crucial in cancer classification. The performance of the proposed approach is plotted for accuracy, precision, error rate and classification efficacy. The performance is also compared with KNN classifier which is used by many researchers in the past for cancer classification. The output shows that the proposed approach is fast and more accurate, making it a perfect choice to real life cancer detection.

As a part of our investigation into lung cancer prognosis, this research integrated a feature selection method with a classification system. Using feature selection approaches to minimise the number of features, it is believed that most classification systems may be improved. There are certain factors that have a greater impact on the categorization algorithms than others.

The findings of our tests using a well-known classification technique, namely naive Bayes, SSA, have been provided. As a result, naïve Bayes provided superior output without SSA, but SSA enhanced performance. New algorithms and feature selection strategies will be tested in the future as part of this research. Our experiments will include both cluster and ensemble methods. A machine learning based web application or Android application can be developed for the user benefits.

References

[1]. Almotairi, S., Kareem, G., Aouf, M., Almutairi, B., & Salem, M. A. M. (2020). Liver tumor segmentation in CT scans using modified SegNet. Sensors, 20(5), 1516.

[2]. Abdel-massieh, N. H., Hadhoud, M. M., & Amin, K. M. (2010, December). Automatic liver tumor segmentation from CT scans with knowledge-based constraints. In 2010 5th Cairo International Biomedical Engineering Conference (pp. 215-218). IEEE.

[3]. Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., ... & Feuer, E. J. (2020). The effect of advances in lung-cancer treatment on population mortality. New England Journal of Medicine, 383(7), 640-649.

[4]. ArneMilitzer, Tobias Hager & Florian Jager & Christian Tietjen 2010, "Automatic Detection and Segmentation of Focal Liver Lesions in Contrast Enhanced CT Images‟, Pattern Recognition (ICPR), 2010 20th International Conference on 23-26.

[5]. Bakas, S, Chatzimichail, K, Hunter, G, Labbe, B, Sidhu, PS & Makris, D 2015, „Fast semi-automatic segmentation of focal liver lesions in contrast enhanced ultrasound, based on a probabilistic model‟, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, pp. 1–10.

[6]. Nooreldeen, R., & Bach, H. (2021). Current and future development in lung cancer diagnosis. International journal of molecular sciences, 22(16), 8661.

[7]. Changyang Li, Xiuying Wang, Stefan Eberl, Michael Fulham, Yong Yin, Jinhu Chen & David Dagan Feng 2013, „A Likelihood and Local Constraint Level Set Model for Liver Tumor Segmentation from CT Volumes‟, IEEE Transactions on Biomedical Engineering, vol. 60, issue. 10, pp. 2967 – 2977

[8]. Anu Sharma, RC Jalandhar Mandeep Kaur, "Comparative Analysis of Particle Swarm Optimization and Particle Swarm Optimization with Aging Leader and Challengers towards Benchmark Functions", International Journal of Computer Applications (0975 – 8887) Vol. 120, No.24, June 2015.

[9]. Oudkerk, M., Liu, S., Heuvelmans, M. A., Walter, J. E., & Field, J. K. (2021). Lung cancer LDCT screening and mortality reduction—evidence, pitfalls and future perspectives. Nature reviews Clinical oncology, 18(3), 135-151.

[10]. Wadowska, K., Bil-Lula, I., Trembecki, Ł., & Śliwińska-Mossoń, M. (2020). Genetic markers in lung cancer diagnosis: A review. International journal of molecular sciences, 21(13), 4569.

[11]. Ammu Prasanna Kumar and Preeja Valsala, "Feature Selection for High Dimensional DNA Microarray Data Using Hybrid Approaches", Bio information, Vol. 9, No. 16, pp. 824-828, 2013.

[12]. K. Bhargavi and S. Jyothi, "Classification of DNA Sequence Using Soft Computing Techniques: A Survey", Indian Journal of Science and Technology, Vol. 9(47), December- 2016.

[13]. P. Bhargavi, K. Lohitha Lakshmi, "A Study on Different Gene Expressions Using an Evolutionary Optimization", Proceedings of the 10th INDIACOM; INDIACom-2016; IEEE Conference ID: 37465 2016 3rd International Conference on Computing for Sustainable Global Development, Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA), 16th-18th March, 2016.

[14]. Bhargavi. P, Lohitha Lakshmi. K, Jyothi. S," Gene Sequence Analysis of Breast Cancer using Genetic Algorithm", In: Venkata Krishna P, Obaidat M. (eds) Emerging Research in Data Engineering Systems and Computer Communications.

Advances in Intelligent Systems and Computing, vol. 1054, Springer, Singapore, 11 February 2020.

[15]. Berger, MF, Lawrence, MS, Francesca Demichelis, Yotam Drier, “The Genomic Complexity of Primary Human Prostate Cancer”, Nature, Vol.470, No.7333, pp. 214–220. 2011.

[16]. Ben-Dor, A, Laurakay Bruhn, Nir Friedman, Iftach Nachman, Michel Schummer& Zohar Yakhini, “Tissue Classification with Gene Expression Profiles”, Journal of Computational Biology, Vol.7, No.3.4, pp.559-584, 2000.

[17]. Barman (Mandal). Roy, M. Biswas and Saha. S, “Prediction of Cancer Cell using Digital Signal Processing”, International Journal of Engineering, Vol.9, No.3, pp.91-9, 2011. 94

[18]. Zhang. Biaobiao, Wu.Yue, Lu, Jiabin and Du. K.L, “Evolutionary Computation and its Applications in Neural and Fuzzy Systems”, Applied Computational Intelligence and Soft Computing, Source: DBLP , Vol. 2011, Article ID 938240, 20 pages, 2011.

[19]. Wang, M., Herbst, R. S., & Boshoff, C. (2021). Toward personalized treatment approaches for non-small-cell lung cancer. Nature medicine, 27(8), 1345-1356.

[20]. Mathios, D., Johansen, J. S., Cristiano, S., Medina, J. E., Phallen, J., Larsen, K. R., ... & Velculescu, V. E. (2021). Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nature communications, 12(1), 5060.

[21]. Carbone, A. (2020). Cancer classification at the crossroads. Cancers, 12(4), 980.

[22]. Goodall, G. J., & Wickramasinghe, V. O. (2021). RNA in cancer. Nature Reviews Cancer, 21(1), 22-36.

[23]. Cox, T. R. (2021). The matrix in cancer. Nature Reviews Cancer, 21(4), 217-238.

[24]. Hanahan, D. (2022). Hallmarks of cancer: new dimensions. Cancer discovery, 12(1), 31-46.

[25]. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. Ca Cancer J Clin, 73(1), 17-48.

[26]. Hegde, P. S., & Chen, D. S. (2020). Top 10 challenges in cancer immunotherapy. Immunity, 52(1), 17-35.

[27]. Preethi, K. A., Lakshmanan, G., & Sekar, D. (2021). Antagomir technology in the treatment of different types of cancer. Epigenomics, 13(07), 481-484.

[28]. Cao, W., Chen, H. D., Yu, Y. W., Li, N., & Chen, W. Q. (2021). Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chinese medical journal, 134(07), 783-791.

[29]. Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome and cancer. Cancer Cell, 39(10), 1317-1341.

[30]. Nia, H. T., Munn, L. L., & Jain, R. K. (2020). Physical traits of cancer. Science, 370(6516), eaaz0868.

[31]. Eckerling, A., Ricon-Becker, I., Sorski, L., Sandbank, E., & Ben-Eliyahu, S. (2021). Stress and cancer: mechanisms, significance and future directions. Nature Reviews Cancer, 21(12), 767-785.

[32]. Nagy, Á., Munkácsy, G., & Győrffy, B. (2021). Pancancer survival analysis of cancer hallmark genes. Scientific reports, 11(1), 6047.

[33]. Hong-Hee Won; Sung-Bae Cho; "Paired neural network with negatively correlated features for cancer classification in DNA gene expression profiles", Proceedings of the International Joint Conference on Neural Networks, Vol. 3, Pp. 1708 – 1713, 2003.

[34]. Hiam-Galvez, K. J., Allen, B. M., & Spitzer, M. H. (2021). Systemic immunity in cancer. Nature reviews cancer, 21(6), 345-359.

[35]. Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. Contemporary Oncology/Współczesna Onkologia, 25(1), 45-52.

[36]. J Debela, D. T., Muzazu, S. G., Heraro, K. D., Ndalama, M. T., Mesele, B. W., Haile, D. C., ... & Manyazewal, T. (2021). New approaches and procedures for cancer treatment: Current perspectives. SAGE open medicine, 9, 20503121211034366.

[37]. Martínez-Reyes, I., & Chandel, N. S. (2021). Cancer metabolism: looking forward. Nature Reviews Cancer, 21(10), 669-680.

[38]. Britt, K. L., Cuzick, J., & Phillips, K. A. (2020). Key steps for effective breast cancer prevention. Nature Reviews Cancer, 20(8), 417-436.

[39]. L.D. Miller, P.M. Long, L. Wong, S. Mukherjee, L.M. McShane, and E.T. Liu. Optimal gene expression analysis by microarrays. Cancer Cell, 2:353–361, November 2002.

[40]. C.A. Harrington, C. Rosenow, and J. Retief. Monitoring gene expression using DNA microarrays. Current Opinion in Microbiology, 3(3):285–291, 2000.

[41]. D.V. Nguyen and D.M. Rocke. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 18(1):39–50, 2002.

[42]. Mizrahi, J. D., Surana, R., Valle, J. W., & Shroff, R. T. (2020). Pancreatic cancer. The Lancet, 395(10242), 2008-2020.

[43]. Briukhovetska, D., Dörr, J., Endres, S., Libby, P., Dinarello, C. A., & Kobold, S. (2021). Interleukins in cancer: from biology to therapy. Nature Reviews Cancer, 21(8), 481-499.

[44]. Anwanwan, D., Singh, S. K., Singh, S., Saikam, V., & Singh, R. (2020). Challenges in liver cancer and possible treatment approaches. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1873(1), 188314.

[45]. Fitzgerald, R. C., Antoniou, A. C., Fruk, L., & Rosenfeld, N. (2022). The future of early cancer detection. Nature medicine, 28(4), 666-677.

[46]. C.C. Chang and C.J. Lin. Training v-support vector classifiers: theory and algorithms, 2001. http://www.csie.ntu.edu.tw/~cjlin/ libsvm.

[47]. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5):631, 2005.

[48]. K.Y. Yeung and R.E. Bumgarner. Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biology, 4(12):83–83, 2003.

[49]. Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., ... & Wei, W. (2021). Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. Cancer Communications, 41(11), 1183-1194.

[50]. Ayadi, W., Elhamzi, W., Charfi, I., & Atri, M. (2021). Deep CNN for brain tumor classification. Neural processing letters, 53, 671-700.

[51]. Wang, L., Lankhorst, L., & Bernards, R. (2022). Exploiting senescence for the treatment of cancer. Nature Reviews Cancer, 22(6), 340-355.

[52]. JayashreeDev, Sanjit Kumar Dash, Sweta Dash and Madhusmita Swain, "A Classification Technique for Microarray Gene Expression Data using PSO-FLANN", International Journal on Computer Science and Engineering, Vol. 4, No. 09, pp. 1534- 1539, Sep 2012.

[53]. Alok Sharma, SeiyaImoto, and Satoru Miyano, “A top-r Feature Selection Algorithm for Microarray Gene Expression Data”, IEEE/ACM Transactions on Computational Biology And Bioinformatics, Vol. 9, Vol. 3, pp. 754-764, 2012.

[54]. Castaño F., Fernández-Navarro, C. Hervás-Martínez and P. A. Gutierrez, “Neuro-logistic Models Based on Evolutionary Generalized Radial Basis Function for the Microarray Gene Expression Classification Problem”, Neural Process Letter, Vol. 34, pp. 117–13, 2011. 200

[55]. Chhanda Ray, “Cancer Identification and Gene Classification using DNA Microarray Gene Expression Patterns”, International Journal of Computer Science Issues, Vol. 8, No. 2, pp. 155-160, 2011.

[56]. Venkatesh and Thangaraj, "Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, Vol. 02, No. 06, pp. 2114-2116, 2010

[57]. Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes and Ian H. Witten Data mining in bioinformatics using Weka, 2004 Vol. 20, Pg 2479–2481

[58]. Tomaž Curk, Janez Demšar, Qikai Xu, Gregor Leban, UrošPetrovič, Ivan Bratko, Gad Shaulsky, Blaž Zupan, Microarray data mining with visual programming. Bioinformatics

[59]. Brian T. Wilhelm, Josette-Renée Landry, “RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing”, Methods, Vol. 48, No. 3, pp. 249–257, July 2009.

[60]. Gad Getz, Erel Levine and EytanDomany, “Coupled two-way clustering analysis of gene microarray data”, Proceeding of the National Academy of Sciences of the United States of America, Vol. 97, No. 22, 2000.

[61]. Adi L. Tarca, Roberto Romero and SorinDraghici, “Analysis of microarray experiments of gene expression profiling”, American Journal of Obstetrics and Gynecology, Vol. 195, No. 2, pp. 373–388, 2006

[62]. M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill, “Analysis of Variance for Gene Expression Microarray Data”, Journal of Computational Biology, Vol. 7, No. 6, pp. 819-837, December 2000.

[63]. Laura J. vant Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards and Stephen H. Friend “Gene expression profiling predicts clinical outcome of breast cancer”, Nature, Vol. 415, pp. 530-536, 2002.

[64]. Matthews, H. K., Bertoli, C., & de Bruin, R. A. (2022). Cell cycle control in cancer. Nature Reviews Molecular Cell Biology, 23(1), 74-88.

[65]. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.

[66]. Joshi, S. S., & Badgwell, B. D. (2021). Current treatment and recent progress in gastric cancer. CA: a cancer journal for clinicians, 71(3), 264-279.

[67]. Tran, L., Xiao, J. F., Agarwal, N., Duex, J. E., & Theodorescu, D. (2021). Advances in bladder cancer biology and therapy. Nature Reviews Cancer, 21(2), 104-121.

[68]. Katsura, C., Ogunmwonyi, I., Kankam, H. K., & Saha, S. (2022). Breast cancer: presentation, investigation and management. British Journal of Hospital Medicine, 83(2), 1-7.

[69]. Sushama Tanwar, S. Vijayalakshmi, Munish Sabharwal 2021. Using Novel Method with Convolutional Neural Network for Colorectal Cancer Classification. Annals of the Romanian Society for Cell Biology. (Mar. 2021), 2653–2671.

[70]. Savita, S. Vijayalakshmi, Munish Sabharwal, “Alzheimer’s disease Detection through Machine Learning”, Annals of RSCB, pp. 2782–2792, Mar. 2021

[71]. P.Durgadevi, S.Vijayalakshmi, Munish Sabharwal, “Fetal Brain Abnormality Detection through PSO (Particle Swarm Optimization) and Volume Estimation”, Annals of RSCB, pp. 2700–2714, Mar. 2021.

[72]. Kour H., Sabharwal M., Suvanov S., Anand D. (2021) An Assessment of Type-2 Diabetes Risk Prediction Using Machine Learning Techniques.

[73]. K. P. Lowry, ‘‘Long-term outcomes and cost-effectiveness of breast cancer screening with digital breast tomosynthesis in the United States,’’ J. Nat. Cancer Inst., vol. 112, no. 6, pp. 582–589, 2020.

[74]. S. A. George, ‘‘Barriers to cancer screening: An integrative review,’’ Health Care Women Int., vol. 21, no. 1, pp. 53–65, Jan. 2000, doi: 10.1080/073993300245401.

[75]. P. H. Viale, ‘‘The American Cancer society’s facts & figures: 2020 edition,’’ J. Adv. Practitioner Oncol., vol. 11, no. 2, p. 135, Mar. 2020, doi: 10.6004/jadpro.2020.11.2.1.

[76]. J. Zhou, L. Luo, Q. Dou, H. Chen, C. Chen, G. Li, Z. Jiang, and P. Heng, ‘‘Weakly supervised 3D deep learning for cancer classification and localization of the lesions in MR images,’’ J. Magn. Reson. Imag., vol. 50, no. 4, pp. 1144–1151, Oct. 2019, doi: 10.1002/jmri.26721.

[77]. N. Gu, M. Fan, L. Du, and D. Ren, ‘‘Efficient sequential feature selection based on adaptive eigenspace model,’’ Neurocomputing, vol. 161, pp. 199–209, Aug. 2015.

[78]. W. Lotter, A. Rahman Diab, B. Haslam, J. G. Kim, G. Grisot, E. Wu, K. Wu, J. O. Onieva, J. L. Boxerman, M. Wang, M. Bandler, G. Vijayaraghavan, and A. G. Sorensen, ‘‘Robust breast cancer detection in mammography and digital breast

tomosynthesis using annotation efficient deep learning approach,’’ 2019, arXiv:1912.11027. [Online]. Available: http://arxiv.org/abs/1912.11027

[79]. Q. Zhang, L. T. Yang, Z. Chen, and P. Li, ‘‘A survey on deep learning for big data,’’ Inf. Fusion, vol. 42, pp. 146–157, Jul. 2018

[80]. K. Weiss, ‘‘A survey of transfer learning,’’ J. Big Data, vol. 3, no. 1, pp. 1–40, 2016, doi: 10.1186/s40537-016-0043-6.

[81]. R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, ‘‘Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms,’’ Phys. Med. Biol., vol. 62, no. 23, pp. 8894–8908, Nov. 2017, doi: 10.1088/1361-6560/aa93d4.

[82]. F. F. Ting, Y. J. Tan, and K. S. Sim, ‘‘Convolutional neural network improvement for cancer classification,’’ Expert Syst. Appl., vol. 120, pp. 103–115, Apr. 2019.

[83]. M. Toğaçar, K. B. Özkurt, B. Ergen, and Z. Cömert, ‘‘BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of cancer,’’Phys. A, Stat. Mech. Appl., vol. 545, May 2020, Art. no. 123592, doi: 10.1016/j.physa.2019.123592.

[84]. Q. Abbas, ‘‘DeepCAD: A computer-aided diagnosis system for mammographic masses using deep invariant features,’’ Computers, vol. 5, no. 4, p. 28, Oct. 2016, doi: 10.3390/computers5040028.

[85]. Z. Sha, L. Hu, and B. D. Rouyendegh, ‘‘Deep learning and optimization algorithms for automatic cancer detection,’’ Int. J. Imag. Syst. Technol., vol. 30, no. 2, pp. 495–506, Jun. 2020, doi: 10.1002/ima.22400.

[86]. S. Charan, M. J. Khan, and K. Khurshid, ‘‘Cancer detection in mammograms using convolutional neural network,’’ in Proc. Int. Conf. Comput., Math. Eng. Technol. (iCoMET), Mar. 2018, pp. 1–5, doi: 10.1109/icomet.2018.8346384.

[87]. N. Wahab, A. Khan, and Y. S. Lee, ‘‘Transfer learning based deep CNN for segmentation and detection of mitoses in cancer histopathological images,’’ Microscopy, vol. 68, no. 3, pp. 216–233, Jun. 2019, doi: 10.1093/jmicro/dfz002.

[88]. W. Lotter, A. Rahman Diab, B. Haslam, J. G. Kim, G. Grisot, E. Wu, K. Wu, J. O. Onieva, J. L. Boxerman, M. Wang, M. Bandler, G. Vijayaraghavan, and A. G. Sorensen, ‘‘Robust cancer detection in mammography and digital breast tomosynthesis using annotation efficient deep learning approach,’’ 2019, arXiv:1912.11027. [Online]. Available: http://arxiv.org/abs/1912.11027

[89]. F. Jiang, H. Liu, S. Yu, and Y. Xie, ‘‘Breast mass lesion classification in mammograms by transfer learning,’’ in Proc. 5th Int. Conf. Bioinf. Comput. Biol., Jan. 2017, pp. 59–62, doi: 10.1145/3035012.3035022.

[90]. Khan, S., Islam, N., Jan, Z., Din, I. U., & Rodrigues, J. J. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1-6.

[91]. H. Cao, Improve the Performance of Transfer Learning Without FineTuningUsing Dissimilarity-Based Multi-View Learning for Breast Cancer Histology Images (Lecture Notes in Computer Science Image Analysis and Recognition). 2018, pp. 779–787, doi: 10.1007/978-3-319-93000- 8_88.

[92]. E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, ‘‘Transfer learning based histopathologic image classification for cancer detection,’’ Health Inf. Sci. Syst., vol. 6, no. 1, pp. 1–7, Dec. 2018, doi: 10.1007/s13755-018-0057-x.

[93]. Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, ‘‘Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images,’’ Pattern Recognit. Lett., vol. 133, pp. 232–239, May 2020, doi: 10.1016/j.patrec.2020.03.011

[94]. M. Ogino, Z. Li and A. Shimizu, "Augmented Radiology: Feature Space Transfer Model for Prostate Cancer Stage Prediction," in IEEE Access, vol. 9, pp. 102559-102566, 2021, doi: 10.1109/ACCESS.2021.3098038.

[95]. Z. Chen, X. Sun and L. Shen, "An Effective Tumor Classification With Deep Forest and Self-Training," in IEEE Access, vol. 9, pp. 100944-100950, 2021, doi: 10.1109/ACCESS.2021.3096241.

[96]. L. C. Soto-Ayala and J. A. Cantoral-Ceballos, "Automatic Blood-Cell Classification via Convolutional Neural Networks and Transfer Learning," in IEEE Latin America Transactions, vol. 19, no. 12, pp. 2028-2036, Dec. 2021, doi: 10.1109/TLA.2021.9480144.

[97]. M. Mulenga, S. A. Kareem, A. Q. M. Sabri and M. Seera, "Stacking and Chaining of Normalization Methods in Deep Learning-Based Classification of Colorectal Cancer Using Gut Microbiome Data," in IEEE Access, vol. 9, pp. 97296-97319, 2021, doi: 10.1109/ACCESS.2021.3094529.

[98]. T. D. Pham, "From Raw Pixels to Recurrence Image for Deep Learning of Benign and Malignant Mediastinal Lymph Nodes on Computed Tomography," in IEEE Access, vol. 9, pp. 96267-96278, 2021, doi: 10.1109/ACCESS.2021.3094577.

[99]. S. Rajpal, M. Agarwal, V. Kumar, A. Gupta and N. Kumar, "Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification," in IEEE Access, vol. 9, pp. 103347-103364, 2021, doi: 10.1109/ACCESS.2021.3093616.

[100]. H. -R. Kim, K. -J. Kim, K. -T. Lim and D. -H. Choi, "Histological Image Segmentation and Classification Using Entropy-Based Convolutional Module," in IEEE Access, vol. 9, pp. 90964-90976, 2021, doi: 10.1109/ACCESS.2021.3091578.

[101]. G. Senthilkumar et al., "Incorporating Artificial Fish Swarm in Ensemble Classification Framework for Recurrence Prediction of Cervical Cancer," in IEEE Access, vol. 9, pp. 83876-83886, 2021, doi: 10.1109/ACCESS.2021.3087022.

[102]. Das, Pranesh, Dushmanta Kumar Das, and Shouvik Dey. "A new class topper optimizationalgorithm with an application to data clustering." IEEE Transactions on Emerging Topicsin Computing 8, no. 4 (2018): 948-959.

[103]. Gupta, P.K., Das, D.K., Rai, A. and Srivastava, A., 2019, March. Solving economic load dispatch problem using class topper optimization algorithm. In 2019 Innovations in Power and Advanced Computing Technologies (i-PACT) (Vol. 1, pp. 1-5). IEEE.

[104]. El Tekle, G., & Garrett, W. S. (2023). Bacteria in cancer initiation, promotion and progression. Nature Reviews Cancer, 23(9), 600-618.

[105]. Davalos, V., & Esteller, M. (2023). Cancer epigenetics in clinical practice. CA: a cancer journal for clinicians, 73(4), 376-424.

[106]. Ye, F., Dewanjee, S., Li, Y., Jha, N. K., Chen, Z. S., Kumar, A., ... & Tang, H. (2023). Advancements in clinical aspects of targeted therapy and immunotherapy in breast cancer. Molecular cancer, 22(1), 105.

[107]. Ahmedin Jemal, Freddie Bray, Melissa M. Center, Jacques Ferlay, Elizabeth Ward, David Forman (2011). Global Cancer Statistics 2011, CA Cancer J Clin, Vol. 61, pp.69-90.

[108]. Frankell, A. M., Dietzen, M., Al Bakir, M., Lim, E. L., Karasaki, T., Ward, S., ... & Swanton, C. (2023). The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature, 1-9.

[109]. Nogueira, A., Ferreira, A., & Figueiredo, M. (2023). A machine learning pipeline for cancer detection on microarray data: the role of feature discretization and feature selection. BioMedInformatics, 3(3), 585-604.

[110]. Osama, S., Shaban, H., & Ali, A. A. (2023). Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. Expert Systems with Applications, 213, 118946.

[111]. MS, K., Rajaguru, H., & Nair, A. R. (2023). Evaluation and exploration of machine learning and convolutional neural network classifiers in detection of lung cancer from microarray gene—a paradigm shift. Bioengineering, 10(8), 933.

[112]. Anne Rosenberg, Douglas Arthur Kieper, Mark B Williams, Nathalie Johnson and Leora Lanzkowsky (2012). The Role of Molecular Imaging Technologies in Breast Cancer Diagnosis and Management, Imaging of the Breast - Technical Aspects and Clinical Implication, Edited by Dr. Laszlo Tabar, ISBN 978-953-51-0284-7, 8, pp.179- 196.

[113]. Alharbi, F., & Vakanski, A. (2023). Machine learning methods for cancer classification using gene expression data: A review. Bioengineering, 10(2), 173.

[114]. Saheed, Y. K. (2023). Effective dimensionality reduction model with machine learning classification for microarray gene expression data. In Data Science for Genomics (pp. 153-164). Academic Press.

[115]. Alromema, N., Syed, A. H., & Khan, T. (2023). A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data. Diagnostics, 13(4), 708.

[116]. Balakumaran T, ILA Vennila and Gowri Shankar C (2010). Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering, International Journal of Computer Science and Information Security, Vol.7, No.1, pp.

[117]. Asif, S., Zhao, M., Tang, F., & Zhu, Y. (2023). An enhanced deep learning method for multi-class brain tumor classification using deep transfer learning. Multimedia Tools and Applications, 1-28.

[118]. Cheng H.D and Xu H (2002). A novel fuzzy logic approach to mammogram contrast enhancement, Information Science, Vol.148, No.1- 4, pp.167-184.

[119]. Tandel, G. S., Tiwari, A., Kakde, O. G., Gupta, N., Saba, L., & Suri, J. S. (2023). Role of Ensemble Deep Learning for Brain Tumor Classification in Multiple Magnetic Resonance Imaging Sequence Data. Diagnostics, 13(3), 481.

[120]. Abunasser, B. S., Al-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023). Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning. Asian Pacific journal of cancer prevention: APJCP, 24(2), 531.

[121]. Fatima, M., Khan, M. A., Shaheen, S., Almujally, N. A., & Wang, S. H. (2023). B2C3NetF2: Breast cancer classification using an end-to-end deep learning feature fusion and satin bowerbird optimization controlled Newton Raphson feature selection. CAAI Transactions on Intelligence Technology.

[122]. Demir, F., Akbulut, Y., Taşcı, B., & Demir, K. (2023). Improving brain tumor classification performance with an effective approach based on new deep learning model named 3ACL from 3D MRI data. Biomedical Signal Processing and Control, 81, 104424.

[123]. Dhawan A.P, Buelloni G and Gordon R (1986). Enhancement of mammographic features by optimal adaptive neighborhood image processing, IEEE Transactions on Medical Imaging, Vol. 5, No.1, pp. 8-15.

[124]. Ali, W., & Saeed, F. (2023). Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data. Processes, 11(2), 562.

[125]. Dheeba J and Tamil Selvi S (2012). A swarm optimized neural network system for classification of microcalcification in mammograms, Journal of Medical Systems, Vol.36, No.5, pp.3051-306.

[126]. Dheeba J and Wiselin Jiji G (2010). Detection of microcalcification clusters in mammograms using neural network, International Journal of Advanced Science and Technology, Vol.19, pp.13-22.

[127]. Dheeba J, Albert Singh N and Tamil Selvi S (2014).Computer aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach, Journal of Biomedical Informatics, Vol.49, pp.45-52.

[128]. Taha, A. M., Ariffin, D. S. B. B., & Abu-Naser, S. S. (2023). A Systematic Literature Review of Deep and Machine Learning Algorithms in Brain Tumor and Meta-Analysis. Journal of Theoretical and Applied Information Technology, 101(1), 21-36.

[129]. Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: a systematic review. Frontiers in Oncology, 12, 893972.

[130]. Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R., Jawarneh, M., & Asenso, E. (2022). Lung cancer classification and prediction using machine learning and image processing. BioMed Research International, 2022.

[131]. Amin, J., Sharif, M., Haldorai, A., Yasmin, M., & Nayak, R. S. (2022). Brain tumor detection and classification using machine learning: a comprehensive survey. Complex & intelligent systems, 8(4), 3161-3183.

[132]. Rinesh, S., Maheswari, K., Arthi, B., Sherubha, P., Vijay, A., Sridhar, S., ... & Waji, Y. A. (2022). Investigations on brain tumor classification using hybrid machine learning algorithms. Journal of Healthcare Engineering, 2022.

[133]. Venkatesan, C., Balamurugan, D., Thamaraimanalan, T., & Ramkumar, M. (2022, March). Efficient Machine Learning Technique for Tumor Classification Based on Gene Expression Data. In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1982-1986). IEEE.

[134]. Shaikh, F. J., & Rao, D. S. (2022). Prediction of cancer disease using machine learning approach. Materials Today: Proceedings, 50, 40-47.

[135]. Hage Chehade, A., Abdallah, N., Marion, J. M., Oueidat, M., & Chauvet, P. (2022). Lung and colon cancer classification using medical imaging: A feature engineering approach. Physical and Engineering Sciences in Medicine, 45(3), 729-746.

[136]. Rukhsar, L., Bangyal, W. H., Ali Khan, M. S., Ag Ibrahim, A. A., Nisar, K., & Rawat, D. B. (2022). Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification. Applied Sciences, 12(4), 1850.

[137]. Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R., Jawarneh, M., & Asenso, E. (2022). Lung cancer classification and prediction using machine learning and image processing. BioMed Research International, 2022.

[138]. Günaydin, Ö., Günay, M., & Şengel, Ö. (2019, April). Comparison of lung cancer detection algorithms. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-4). IEEE.

[139]. Kom G, Tiedeu A and Kom M (2007). Automated detection of masses in mammograms by local adaptive thresholding, Computers in Biology Medicine, Vol.37, No.1, pp. 37-48.

[140]. Kuan-Yuei Li and Yu-Kun Huang (2006). Detection of microcalcifications in digital mammograms using wavelet filter and Markov random field model, Computerized Medial Imaging and Graphics, Vol.30, No.3, pp.163-173.

[141]. Lannin, T. B., Thege, F. I., & Kirby, B. J. (2016). Comparison and optimization of machine learning methods for automated classification of circulating tumor cells. Cytometry Part A, 89(10), 922-931.

[142]. Castillo T, J. M., Arif, M., Niessen, W. J., Schoots, I. G., & Veenland, J. F. (2020). Automated classification of significant prostate cancer on MRI: a systematic review on the performance of machine learning applications. Cancers, 12(6), 1606.

[143]. Sharma, S., Gupta, S., Gupta, D., Juneja, S., Gupta, P., Dhiman, G., & Kautish, S. (2022). Deep learning model for the automatic classification of white blood cells. Computational Intelligence and Neuroscience, 2022.

[144]. Masud, M., Sikder, N., Nahid, A. A., Bairagi, A. K., & AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. Sensors, 21(3), 748.

[145]. Saber, A., Sakr, M., Abo-Seida, O. M., Keshk, A., & Chen, H. (2021). A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. IEEE Access, 9, 71194-71209.

[146]. Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J., & Maria Vanegas, A. (2020). Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors, 20(16), 4373.

[147]. Lo S.C.B, Li H, Wong Y, Kinnard L and Freedman M.T (2002). A multiple circular path convolution neural network system for detection of mammography masses, IEEE Transactions on Medical Imaging, Vol.21, No 2, pp.150-158.

[148]. Malik MAN, Salahudin O, Azar M, Irshad H, Sadia, SalahudinA(2010). Breast diseases: spectrum in Wah Cantt, POF Hospital Experience. Professional Med J, Vol.17, No, pp.366-372.

[149]. Naeem, S., Ali, A., Qadri, S., Khan Mashwani, W., Tairan, N., Shah, H., ... & Anam, S. (2020). Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images. Applied Sciences, 10(9), 3134.

[150]. Amin, J., Sharif, M., Haldorai, A., Yasmin, M., & Nayak, R. S. (2022). Brain tumor detection and classification using machine learning: a comprehensive survey. Complex & intelligent systems, 8(4), 3161-3183.

[151]. Mario Mustra, MislavGrgic and KresimirDelac (2012). Enhancement of microcalcifications in digital mammograms, Intelligent Image Features Extraction in Knowledge Discovery Systems, IWSSIP, pp.248-251.

[152]. Mencattini A, Salmeri M, Lojacono R, Frigerio M and Caselli F (2008). Mammographic images enhancement and denoising for breast cancer detection using dyadic wavelet processing, IEEE Transactions on Instrumentation and Measurement, Vol. 57, No.7, pp.1422- 1430.

[153]. Cuocolo, R., Caruso, M., Perillo, T., Ugga, L., & Petretta, M. (2020). Machine learning in oncology: a clinical appraisal. Cancer letters, 481, 55-62.

[154]. Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. Journal of Infection and Public Health, 13(9), 1274-1289.

[155]. Mohamed Lagzouli and YoussfiElkettani (2014). A New Morphology Algorithm for Microcalcifications Detection in Fuzzy Mammograms Images, International Journal of Engineering Research & Technology, Vol. 3, No.1, pp.729 -733.

[156]. Mohd. Khuzi A, Besar R, Wan Zaki W, and Ahmad N (2009). Identification of masses in digital mammogram using gray level co-occurrence matrices, Biomedical Imaging Interv Journal, Vol.5, No.3, pp.1-13.

[157]. Khan, M. A., Ashraf, I., Alhaisoni, M., Damaševičius, R., Scherer, R., Rehman, A., & Bukhari, S. A. C. (2020). Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. Diagnostics, 10(8), 565.

[158]. Papandrianos, N., Papageorgiou, E., Anagnostis, A., & Feleki, A. (2020). A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. Applied Sciences, 10(3), 997.

[159]. Mussarat Yasmin, Muhammad Sharif and Sajjad Mohsin (2013) Survey Paper on Diagnosis of Breast Cancer Using Image Processing Techniques, Research Journal of Recent Sciences ,Vol. 2, No.10, pp.88-98.

[160]. Suh, Y. J., Jung, J., & Cho, B. J. (2020). Automated breast cancer detection in digital mammograms of various densities via deep learning. Journal of personalized medicine, 10(4), 211.

[161]. Abd El Kader, I., Xu, G., Shuai, Z., Saminu, S., Javaid, I., & Salim Ahmad, I. (2021). Differential deep convolutional neural network model for brain tumor classification. Brain Sciences, 11(3), 352.

[162]. Nalini Singh, Ambarish G Mohapatra and Gurukalyan Kanungo (2011). Breast Cancer Mass Detection in Mammograms using K-means and Fuzzy C-means Clustering, International Journal of Computer Applications. Vol. 22, No.2, pp.15-21.

[163]. Kang, J., Ullah, Z., & Gwak, J. (2021). Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. Sensors, 21(6), 2222.

[164]. Netsch T and Peitgen H.O (1999). Scale-space signatures for the detection of clustered microcalcifications in digital mammograms, IEEE Transactions on Medical Imaging, Vol.18, No.9, pp.774-786.

[165]. Niblack W. An Introduction to Digital Image Processing. Prentice-Hall, Englewood Cliffs, NJ, pp.115-116 Nitin S Lingayat and Manoj R Tarambale (2013). A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol.3, No.6, pp.624-629.

[166]. Nizar Ben Hamad, Khaled Taouil and Med Salim Bouhlel (2013). Mammographic Microcalcifications Detection using Discrete Wavelet Transform, International Journal of Computer Applications, Vol.64, No.21, pp.17-22.

[167]. B. A. Soomro, G. R. Lakhan, S. Mangi, and N. Shah, ‘‘Predicting entrepreneurial intention among business students of public Class Topper Optimization universities of Pakistan: An application of the entrepreneurial event model,’’ World J. Entrepreneurship, Manage. Sustain. Develop., vol. 16, no. 3, pp. 219–2301, 2020.

[168]. J. Tu, A. Lin, H. Chen, Y. Li, and C. Li, ‘‘Predict the entrepreneurial intention of fresh graduate students based on an adaptive support Class Topper Optimization machine framework,’’ Math. Problems Eng., vol. 2019, pp. 1–16, Jan. 2019.

[169]. Y. Wei, N. Ni, D. Liu, H. Chen, M. Wang, Q. Li, X. Cui, and H. Ye, ‘‘An improved grey wolf optimization strategy enhanced SVM and its application in predicting the second major,’’ Math. Problems Eng., vol. 2017, pp. 1–12, Feb. 2017.

[170]. R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, ‘‘Multi-task transfer learning deep convolutional neural network:

Application to computer-aided diagnosis of breast cancer on mammograms,’’ Phys. Med. Biol., vol. 62, no. 23, pp. 8894–8908, Nov. 2017, doi: 10.1088/1361-6560/aa93d4.

[171]. F. F. Ting, Y. J. Tan, and K. S. Sim, ‘‘Convolutional neural network improvement for cancer classification,’’ Expert Syst. Appl., vol. 120, pp. 103–115, Apr. 2019.

[172]. Z. Sha, L. Hu, and B. D. Rouyendegh, ‘‘Deep learning and optimization algorithms for automatic cancer detection,’’ Int. J. Imag. Syst. Technol., vol. 30, no. 2, pp. 495–506, Jun. 2020, doi: 10.1002/ima.22400.

[173]. Haney, P. Tandon, R. Divi, M. R. Ossandon, H. Baker and P. C. Pearlman, "The Role of Affordable, Point-of-Care Technologies for Cancer Care in Low- and Middle-Income Countries: A Review and Commentary," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 5, pp. 1-14, 2017, Art no. 2800514, doi: 10.1109/JTEHM.2017.2761764.

[174]. M. T. Islam, E. Tasciotti and R. Righetti, "Non-Invasive Imaging of Normalized Solid Stress in Cancers in Vivo," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 7, pp. 1-9, 2019, Art no. 4300209, doi: 10.1109/JTEHM.2019.2932059.

[175]. Saribudak, H. Kucharavy, K. Hubbard and M. Ü. Uyar, "Spatial Heterogeneity Analysis in Evaluation of Cell Viability and Apoptosis for Colorectal Cancer Cells," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 4, pp. 1-9, 2016, Art no. 4300209, doi: 10.1109/JTEHM.2016.2578331.

[176]. Masood et al., "Cloud-Based Automated Clinical Decision Support System for Detection and Diagnosis of Lung Cancer in Chest CT," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 8, pp. 1-13, 2020, Art no. 4300113, doi: 10.1109/JTEHM.2019.2955458.

[177]. K. H. Chow, R. E. Factor, and K. S. Ullman, “The nuclear envelope environment and its cancer connections,” Nature Rev. Cancer, vol. 12, pp. 196–209, Mar. 2012. Authorized licensed use limited to: Shakuntla Educational & Welfare Society Galgotias. Downloaded on February 12,2021 at 08:57:15 UTC from IEEE Xplore. Restrictions apply. LIU et al.: SETSVM: AN APPROACH TO SET CLASSIfiCATION IN NUCLEI-BASED CANCER DETECTION 361

[178]. H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, “Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—Current status and future potential,” IEEE Rev. Biomed. Eng., vol. 7, pp. 97–114, Jan. 2014.

[179]. Morgado, J., Pereira, T., Silva, F., Freitas, C., Negrão, E., de Lima, B. F., ... & Oliveira, H. P. (2021). Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. Applied Sciences, 11(7), 3273.

[180]. Yuan, F., Lu, L., & Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1866(8), 165822.

[181]. Chen, J. W., & Dhahbi, J. (2021). Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. Scientific reports, 11(1), 13323.

[182]. Raj, R. J. S., Shobana, S. J., Pustokhina, I. V., Pustokhin, D. A., Gupta, D., & Shankar, K. J. I. A. (2020). Optimal feature selection-based medical image classification using deep learning model in internet of medical things. IEEE Access, 8, 58006-58017.

[183]. W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde, “A linear ˇ optimal transportation framework for quantifying and visualizing variations in sets of images,” Int. J. Comput. Vision, vol. 101, no. 2, pp. 254–269, Sep. 2012.

[184]. K. Sirinukunwattana, S. E. A. Raza, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, “Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images,” IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1196–1206, May 2016.

[185]. J. Xu et al., “Stacked sparse autoencoder (SSAE) for nuclei detection on lung cancer histopathology images,” IEEE Trans. Med. Imag., vol. 35, no. 1, pp. 119–130, Jan. 2016.

[186]. F. Xing, Y. Xie, and L. Yang, “An automatic learning-based framework for robust nucleus segmentation,” IEEE Trans. Med. Imag., vol. 35, no. 2, pp. 550–566, Feb. 2016.

[187]. Janowczyk, S. Doyle, H. Gilmore, and A. Madabhushi, “A resolution adaptive deep hierarchical (RADHical) learning scheme applied to nuclear segmentation of digital pathology images,” Comput. Methods Biomech. Biomed. Eng. Imag. Vis., vol. 1163, pp. 1–7, 2016.

[188]. H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. A. Heng, “DCAN: Deep contour-aware networks for object instance segmentation from histology images,” Med. Image Anal., vol. 36, pp. 135–146, Feb. 2017.

[189]. Z. Gao, L. Wang, L. Zhou, and J. Zhang, “HEp-2 cell image classification with deep convolutional neural networks,” IEEE J. Biomed. Health Informat., vol. 21, no. 2, pp. 416–428, Mar. 2017.

[190]. K. -H. Yu et al., “Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features,” Nature Commun., vol. 7, 2016, Art. no. 12474.

[191]. X. Zhang, F. Xing, H. Su, L. Yang, and S. Zhang, “High-throughput histopathological image analysis via robust cell segmentation and hashing,” Med. Image Anal., vol. 26, no. 1, pp. 306–315, Dec. 2015.

[192]. H. A. Phoulady, M. Zhou, D. B. Goldgof, L. O. Hall, and P. R. Mouton, “Automatic quantification and classification of cervical cancer via adaptive nucleus shape modeling,” in Proc. 2016 IEEE Int. Conf. Image Process., 2016, pp. 2658–2662.

[193]. P. Filipczuk, T. Fevens, A. Krzyzak, and R. Monczak, “Computer-aided lung cancer diagnosis based on the analysis of cytological images of fine needle biopsies,” IEEE Trans. Med. Imag., vol. 32, no. 12, pp. 2169–2178, Dec. 2016.

[194]. Atupelage et al., “Computational hepatocellular carcinoma tumor grading based on cell nuclei classification.” J. Med. Imag., vol. 1, no. 3, 2014, Art. no. 034501.

[195]. J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artif. Intell., vol. 201, pp. 81–105, Aug. 2013.

[196]. Y. Xu, J. -Y. Zhu, E. I. -C. Chang, M. Lai, and Z. Tu, “Weakly supervised histopathology cancer image segmentation and classification,” Med. Image Anal., vol. 18, no. 3, pp. 591–604, Apr. 2014.

[197]. G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, “Multiple-instance learning for anomaly detection in digital mammography,” IEEE Trans. Med. Imag., vol. 35, no. 7, pp. 1604–1614, Jul. 2016.

[198]. Lopez-Rincon, A., Mendoza-Maldonado, L., Martinez-Archundia, M., Schönhuth, A., Kraneveld, A. D., Garssen, J., & Tonda, A. (2020). Machine learning-based

ensemble recursive feature selection of circulating mirnas for cancer tumor classification. Cancers, 12(7), 1785.

[199]. M. Kandemir and F. A. Hamprecht, “Computer-aided diagnosis from weak supervision: A benchmarking study,”Comput. Med. Imag. Graphics, vol. 42, pp. 44–50, Jun. 2015.

[200]. M. B. Terry, J. A. McDonald, H. C. Wu, S. Eng, and R. M. Santella, ‘‘Epigenetic biomarkers of lung cancer risk: Across the lung cancer prevention continuum,’’ Adv. Exp. Med. Biol., vol. 882, pp. 33–68, Mar. 2016.

[201]. Baur and S. Bozdag, ‘‘A feature selection algorithm to compute gene centric methylation from probe level methylation data,’’ PLoS One, vol. 11, no. 2, p. e0148977, 2016.

[202]. J. Wu et al., ‘‘Identification of biomarkers for predicting lymph node metastasis of stomach cancer using clinical DNA methylation data,’’ Disease Markers, vol. 2017, pp. 1–7, Aug. 2017.

[203]. M. Pouliot, Y. Labrie, C. Diorio, and F. Durocher, ‘‘The role of methylation in lung cancer susceptibility and treatment,’’ Anticancer Res., vol. 35, pp. 4569–4574, Sep. 2015.

[204]. Z. M. Hira and D. F. Gillies, ‘‘A review of feature selection and feature extraction methods applied on microarray data,’’ Adv. Bioinform., vol. 2015, p. 198363, Jun. 2015.

[205]. J. Liu, Y. Cheng, X. Wang, L. Zhang, and H. Liu, ‘‘An optimal mean based block robust feature extraction method to identify colorectal cancer genes with integrated data,’’ Sci. Rep., vol. 7, p. 8584, Aug. 2017.

[206]. S. Kaur and S. Kalra, ‘‘Feature extraction techniques using support vector machines in disease prediction,’’ in Proc. IJARSE, May 2016, p. 5.

[207]. R. K. Singh and M. Sivabalakrishnan, ‘‘Feature selection of gene expression data for cancer classification: A review,’’ Proc. Comput. Sci., vol. 50, pp. 52–57, Jan. 2015.

[208]. Nguyen, Y. Wang, and H. N. Nguyen, ‘‘Random forest classifier combined with feature selection for lung cancer diagnosis and prognostic,’’ J. Biomed. Sci. Eng., vol. 6, no. 5, pp. 551–560, 2013.

[209]. X. Ren, Y. Wang, X.-S. Zhang, and Q. Jin, ‘‘iPcc: A novel feature extraction method for accurate disease class discovery and prediction,’’ Nucl. Acids Res., vol. 41, no. 4, p. e143, 2013

[210]. Mohit Jain et. Al. A novel nature-inspired algorithm for optimization: Squirrel search algorithm, Swarm and Evolutionary Computation Volume 44, February 2019, Pages 148-175

[211]. Chunyu Wang, Junling Guo, Ning Zhao, Yang Liu, Xiaoyan Liu, Guojun Liu, Maozu Guo. "A Cancer Survival Prediction Method Based on Graph Convolutional Network", IEEE Transactions on NanoBioscience, 2020

[212]. https://archive.ics.uci.edu/ml/datasets/lung+cancer

[213]. Singh, R. K., & Sivabalakrishnan, M. J. P. C. S. (2015). Feature selection of gene expression data for cancer classification: a review. Procedia Computer Science, 50, 52-57.

[214]. Ristiana, R., Kusumandari, D. E., Simbolon, A. I., Amri, M. F., Sanhaji, G., & Rumiah, R. (2021, November). A Comparative Study of Thermal Face Recognition Based on Haar Wavelet Transform (HWT) and Histogram of Gradient (HoG). In 2021 3rd International Symposium on Material and Electrical Engineering Conference (ISMEE) (pp. 242-248). IEEE.

[215]. Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. Translational oncology, 14(1), 100907.

[216]. Li, M., Ma, X., Chen, C., Yuan, Y., Zhang, S., Yan, Z., ... & Ma, M. (2021). Research on the auxiliary classification and diagnosis of lung cancer subtypes based on histopathological images. Ieee Access, 9, 53687-53707.

[217]. Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal, 19, 3735-3746.

[218]. Sun, C., Xu, A., Liu, D., Xiong, Z., Zhao, F., & Ding, W. (2019). Deep learning-based classification of liver cancer histopathology images using only global labels. IEEE journal of biomedical and health informatics, 24(6), 1643-1651.