

AN INTELLIGENT LUNG CANCER PREDICTION USING DEEP LEARNING

Thesis Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in

Computer Science & Engineering

By

Ms. Kanchan Sitaramji Pradhan

41800587

Supervised By

Dr. Priyanka Chawla

22046



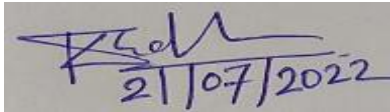
LOVELY PROFESSIONAL UNIVERSITY

PUNJAB

2023

DECLARATION

I, hereby declared that the presented work in the thesis entitled “**An Intelligent Lung Cancer Prediction Using Deep Learning**” in fulfillment of the degree of **Doctor of Philosophy (Ph. D.)** is outcome of research work carried out by me under the supervision **Dr. Priyanka Chawla**, working as Professor, in the School of Computer Science & Engineering of **Lovely Professional University, Punjab, India**. In keeping with general practice of reporting scientific observations, due acknowledgments have been made whenever work described here has been based on findings of another investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.

A rectangular box containing a handwritten signature in blue ink and the date '21/07/2022' written below it.

Name of the scholar: Kanchan Sitaramji Pradhan

Registration No.:41800587

Department/School: School of Computer Science & Engineering,
Lovely Professional University, Punjab, India

CERTIFICATE

This is to certify that the work reported in the Ph. D. thesis entitled “An Intelligent Lung Cancer Prediction Using Deep Learning” submitted in fulfillment of the requirement for the reward of the degree of **Doctor of Philosophy (Ph.D.)** in Computer Science & Engineering, is a research work carried out by **Kanchan Sitaramji Pradhan**, 41800587, is bonafide record of her original work carried out under my supervision and that no part of thesis has been submitted for any other degree, diploma or equivalent course.



Name of supervisor: Dr. Priyanka Chawla

Designation: Professor

Department/School: School of Computer Science & Engineering

University: Lovely Professional University, Punjab, India

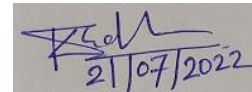
Acknowledgment

I am delighted to express a deep sense of gratitude and thanks to my respected supervisor **Dr. Priyanka Chawla** who has offered tremendous support in the completion of this research. Her unparalleled knowledge, judgment, and moral fibre were together with their expertise. I would like to thank the administrative staff members of Centre of Research Degree Programmes for their support and guidance. I would also thank the Chancellor, Vice-Chancellor of Lovely Professional University, Phagwara for providing me with this research opportunity.

I would like to express my gratitude toward the entire **Lovely Professional University** family for providing a suitable infrastructure and environment for completing my research work in a time-bound manner. Also, I like to thank the Division of Research & Development and School of Computer Engineering for their help and encouragement in my entire Ph.D. journey.

I would also thank God, my husband, parents, friends, and contemporaries for their co-operation and compliance. I cannot cherish a greater fortune other than having them in my life. Their care and love are indispensable for my achievements.

I express my sincere thanks to everyone who has helped me in various capacities to carry out this research and prepare the report.

A rectangular box containing a handwritten signature in blue ink and the date '21/07/2022' written below it.

Ms. Kanchan Sitaramji Pradhan

Abstract

Cancer is a very serious disease that can be life-threatening. Among all types of cancer, lung cancer is the main reason many people around the world die from cancer. If lung cancer isn't found early, the chance of surviving can be low. Moreover, there aren't many radiologists available. In India, for instance, there's only one radiologist for every one lakh patients. This is why we need automatic ways to help doctors look at patients quickly and decide on treatment. To combat this disease, advanced techniques are being employed to foresee its future implications and show complicated relationships and patterns within extensive and complex datasets. The healthcare sector is observing a notable transformation driven by the enhanced computational capabilities of machine learning and deep learning methods, which allow accurate data analysis. Deep learning techniques, which play a pivotal role in predicting the significance of various diseases as well as their potential outcomes.

Primary goal of this research is to propose an intelligent lung cancer prediction using deep learning. To address this problem Self Adaptive Sea Lion Optimization (SA-SL_nO) algorithm along with Recurrent Neural Network (RNN) has been used to increase classification. proposed model contains of different phases: Data Gathering, Feature Extraction, New attribute correlation-based optimized weighted feature extraction, and Prediction. In the initial phase, two distinct attribute extraction methods are utilized: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). The model incorporates an original input, along with an enhanced correlation-based biased feature elimination process referred to as SA-SL_nO. Advanced high-level independent algorithms are connected to refine and enhance the overall performance of the model. The SA-SL_nO algorithm is designed to optimize number of hidden nerve cells within a Recurrent Neural Network (RNN), with the primary objective of minimizing the Mean Squared Error (MSE) among actual and projected outputs in the context of lung tumor detection. The main focus of this lung tumor detection model effectively reduces the MSE among actual and projected output. New outcomes of the proposed model display significantly improved results in comparison to existing methods across various evaluation metrics, including False Positive Rate (FPR), False Negative Rate (FNR), F1-score, Accuracy, Sensitivity, and Precision.

Ensuing to new results, accuracy achieved by the SA-SL_nO-based RNN exceeds that of other models by 1.7%, 2%, 5%, and 2.7% in relation to PSO-RNN, GWO-RNN, GSO-RNN, and SL_nO-RNN, respectively. Similarly, when evaluating the model's performance on two distinct datasets, the proposed lung cancer analysis model reliably shows enhanced performance. The

experimental findings validate the efficiency of the proposed model by showing its superiority in comparison to existing methodologies. Specifically, the model's impact is apparent in optimizing hidden neurons within RNNs, outperforming traditional techniques. In conclusion, through a complete comparative analysis with conventional methods, the experimental results provide a convincing indication that the proposed model characterizes a momentous improvement in this area.

List of Figures

Figure 1. 1 Human Lung Anatomy [3]	2
Figure 1. 2 Fundamental Deep Learning Architecture [15].....	13
Figure 3. 1 PCA feature space	38
Figure 3. 2 Scree Plot.....	38
Figure 4. 1 Proposed Lung Cancer Detection Architecture	46
Figure 4. 2 The proposed optimized RNN.....	48
Figure 4.3 Flowchart of the proposed SA-SLnO algorithm.....	53
Figure 5. 1 Comparative Analysis of Accuracy	58
Figure 5. 2 Comparative Analysis of Precision	58
Figure 5. 3 Comparative Analysis of FPR	59
Figure 5. 4 Comparative Analysis of FNR	59
Figure 5. 5 F1-score Comparative Analysis	60
Figure 5. 62 Impact lung cancer diagnosis on attribute correlation-based optimized weighted feature extraction for rescission.	64

List of Tables

Table 1. 1 Comparison between Benign and Malignant Tumor	3
Table 2. 1 Review of The State of The Art	30
Table 2. 2 Comparative Analysis of Research Work.....	34
Table 3. 1 Attributes of Database.....	43
Table 5. 1 Overall Performance Analysis of The Developed and Traditional Meta-Heuristic-Based RNN [29].....	70
Table 5. 2 Overall Performance Analysis of the Developed and Traditional Classifiers [29] 71	
Table 5. 3 Performance evaluation of the lung cancer diagnostic model using several optimization-based techniques regarding complexity [29].....	83
Table 5. 4 Performance evaluation of the built lung cancer diagnostic model using several classifiers in terms of time complexity	83
Table 5. 5 Comparing the lung cancer diagnosis model's overall performance in terms of time complexity to several traditional models	83

Table of Contents

DECLARATION	ii
CERTIFICATE	iii
Acknowledgment	iv
Abstract	v
List of Figures	vii
List of Tables	viii
Chapter 1	1
Introduction	1
1.1 Overview of Lung Cancer	1
1.2 Screening of Lung Cancer	3
1.3 Lung Cancer Causes	4
1.3.1 Lung Cancer Symptoms	5
1.3.2 Lung Cancer Risk Factor	6
1.4 Types of Lungs Cancer	7
1.4.1 Small Cell Lung Cancer	7
1.4.2 Non-Small Cell Lung Cancer (NSCLC).....	8
1.5 Tests for Lung Cancer	9
1.5.1 Medical History and Physical Exam.....	9
1.5.2 Imaging Tests to Look for Lung Cancer	9
1.6 Introduction to Deep Learning	10
1.7 Machine Learning Approaches	11
1.8 Fundamental of Deep Learning Architectures	12
1.9 Applications of Deep Learning	14
1.10 Challenges	16
1.11 Motivation	17
1.12 Research Objectives	18
1.13 Major Contribution of Thesis	18
1.14 Thesis Outline	19
Chapter 2	21
Literature Review	21
2.1 Introduction	21
2.2 Lung Cancer Diagnosis Approach	21

2.3	Research Gap.....	35
Chapter 3.....		37
Development of Feature Selection and Classification Model		37
3.1	Principal Component Analysis.....	37
3.2	t- Distributed Stochastic Neighbor Embedding (t-SNE).....	39
3.3	Feature Extraction Model.....	39
3.3.1	Innovative Weighted Feature Extraction Model Based on Attribute Correlation	42
3.4	Dataset.....	43
Chapter 4.....		45
Proposed Methodology		45
4.1	Proposed SA-SLnO-RNN Deep Learning Model for Prediction of Lung Cancer....	45
4.1.1	System Architecture.....	45
4.1.2	Objective Model for Proposed Lung Cancer Diagnosis.....	47
4.1.3	Optimized RNN.....	47
4.1.4	Proposed SA-SLnO Algorithm and Model	49
Chapter 5.....		55
Verification and Validation of Proposed Model		55
5.1	Work Environment.....	55
5.2	Performance Measures	56
5.3	Performance Analysis on Optimization Algorithms for Dataset 1 using SA-SLnO Algorithm	57
5.4	Comparative analysis of Model Performance	60
5.5	Optimized Feature Extraction with Weighted Attributes Based on Correlation.....	63
5.6	Performance Evaluation of Existing Techniques	66
5.7	Performance Evaluation for Dataset 2	67
5.8	Overall Performance Evaluation on Existing Models.....	69
5.9	Performance Analysis on the Time Complexity of Proposed Model.....	82
Chapter 6.....		85
Conclusion and Future Scope		85
References		86
List Of Publications		95

1.1 Overview of Lung Cancer

The most common cause of mortality is lung cancer. The American Cancer Society's most current statistics show that there are around 228 thousand new instances of lung cancer, of which approximately 135 thousand results in death [1]. In the majority of instances, delayed therapy due to late detection results in mortality. Early lung cancer identification, therefore, has a crucial role in improving the patient's chances of survival. Therefore, quick localization and delineation of the cancer site in high-resolution images may aid in diagnosis and planning. Chest computed tomography (CT) scans play a significant character in detection and analysis of lung cancer. Recent studies have shown that CT screening can be effective in community settings and that low-dose CT screening lowers lung cancer fatalities in dangerous persons. The broad adoption of lung cancer screening benefit millions of individuals. Due to the many CT scan pictures that radiologists get from their patients; they are overworked. There has also been evidence of intra-team conflict. Computer-aided diagnostic tools have previously been shown to improve the diagnosis of lung nodules in computed tomography (CT) scans [1]. Medical machine learning models may be tested for robustness, and a wide range of processing methods for chest CT scans can be compared to see which one performs better. The wealth of a nation is determined by the health of its citizens. In the medical world, heart, cancer, and orthopedic illnesses are among the top three killers. Cells in the body grow improperly, and these aberrant cells proliferate at an excessive rate, resulting in cancer. Normally, healthy cells divide at a far slower rate than this. Cancer cells, on the other hand, are deemed abnormal because they develop rapidly and do not die for a long period. Early sickness and death are often the result of abnormal cell proliferation. Metastasis is the term used to describe the growth and development aberrant cells. According to World Health Organization (WHO), over four million people in the world are diagnosed with malignancies each year, and treatment is critical. Therapeutic imaging techniques can be used in a few areas of medicine, such as computer-aided pathology diagnosis, surgical planning and guidance, and longitudinal study. Computed tomography (CT) and magnetic resonance imaging (MRI) are the tomography methods of choice for lung cancer screening. Because medical images typically contain a great quantity of data and sometimes a few artefacts due to the patient's limited acquisition time and sensitive tissue boundaries, frequently not well defined, segmentation of these images is a difficult task.

It's vital to note that the lungs are the most significant respiratory organ in humans. Lungs are spongy body part in the chest that is the breathing system. Thus, to accommodate the heart, there are three lobes in right lung and left side lung has two. Nose or mouth and the trachea and each bronchus are two methods to obtain oxygen in lungs. Cone-shaped organs called the lungs and it found in the body organ (lung). When the air is inhaled, the lungs are filled with oxygen. The lungs pass oxygen all over body via the bloodstream. As they take in oxygen, cells exhale carbon dioxide. Carbon dioxide is returned to the lungs by circulation and expelled when the air is taken in. Figure 1.1 depicts the lungs' anatomy [2].

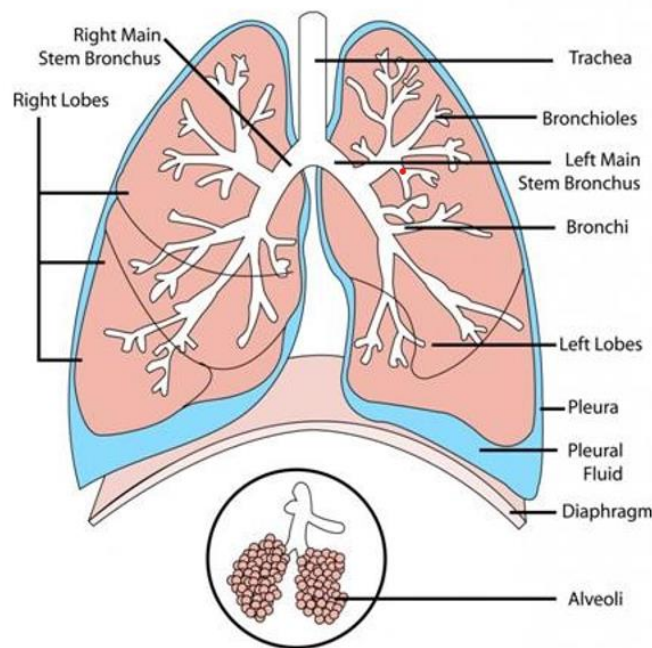


Figure 1. 1 Human Lung Anatomy [3]

The lung, which is where gas exchange takes place, is filled with low-density (1000 HU) air. The primary elements of the lungs, aside from air, are the pulmonary arteries and airways. The left and right lungs are two separate sections of the lungs. The upper and lower lobes of the left lung are distinguished by an oblique fissure. Similarly, the right lobe is segmented into three portions through oblique and horizontal fissures, namely, the upper lobe, middle lobe, and lower lobe. The right and left lungs are supplied through two basic bronchial branches, the left and right primary bronchi, respectively, originate in trachea. Bronchi are entered in each lobe of the body primary bronchi. Each lung region has one or more of these 8–10 segmental bronchi trees. The pulmonary artery's bronchial branch always travels parallel to the airways.

1.2 Screening of Lung Cancer

Every year, lung cancer claims more than a million lives as a result of this disease. When it comes to today's clinical world, detecting lung nodules in CT imaging is critical. Figure 1.2 illustrates the morphology of lung cancer. The treatment of this disease and the probability of survival of the patient largely depends on the stage of the disease. Patients must be diagnosed with the disease as early as possible to maximize their chances of recovery. Lung cancer early detection can be facilitated through a computer-aided diagnosis (CAD) system [4].

There exist two primary categories of lung tumors: benign and malignant. A benign tumor is characterized by well-defined, smooth, and consistent borders, while a malignant tumor exhibits irregular borders and demonstrates a more rapid growth rate than its benign counterpart. Benign versus malignant tumor characteristics are shown in comparison in Table 1. These tumors can be removed surgically and do not affect the rest of the body. Tumor cells can enter the lymphatic system or circulation and spread to other regions of the body through the aggressive growth of tumors. A process known as metastasis occurs when a cancerous tumor spreads throughout the body [5].

It is difficult for a doctor to detect a small cell cancer seen on imaging scan, and in most cases these small cell cancers are difficult to distinguish from one another. Many studies have given several recommendations. Eight-millimeter nodules have been found suitable for follow-up and are usually classified as malignant. Currently, cancer can only be detected by the size of the nodule, which complicates the doctor's work. Our aim is to detect lung cancer in its early stages from an unknown nodule using CT images. [5].

Table 1. 1 Comparison between Benign and Malignant Tumor

Characteristics	Malignant	Benign
Rate of Growth	Rapid growth rate	Slow growth rate
Nature of spreading	Recurrent	Not recurrent
Prognosis	Bad	Good
Metastasis	Frequently present	Absent
Mimics	No Mimics	Normal

Differentiation	Undifferentiating; the cell structure is not nearly as distinct as at conception.	A well-differentiated cell structure; the cell structure is identical to the origin.
-----------------	---	--

1.3 Lung Cancer Causes

Lung cancer is mainly caused by genetic mutations in lung cells that cause uncontrolled growth. These mutations can be triggered by a number of factors, the most common of which is smoking. Below are several frequent causes and risk factors connected with lung cancer:

- Smoking: smoke contains many harmful chemicals that can damage the DNA of lung cells, causing genetic changes and the development of cancer.
- Secondhand smoke: Cigarettes, cigars, or pipes, heighten the likelihood of developing lung cancer, particularly among nonsmokers and children.
- Radon Gas: A naturally occurring radioactive gas has the potential to permeate homes and structures from the ground. Prolonged exposure to elevated radon levels can escalate chance of rising lung cancer.
- Asbestos: Being exposed to asbestos in workplace, a collection of minerals frequently utilized in construction and insulation materials can notably amplify the probability of rising lung cancer.
- Air pollution: Long-term experience to outdoor air pollutants to fine particles and chemicals can make more likely for someone to get lung cancer.
- Family history and genetics: Some people have a genetic predisposition to lung cancer due to inherited gene mutations.
- Personal history of lung disease: Individuals who have previously experienced certain lung conditions like chronic obstructive pulmonary disease (COPD) or pulmonary fibrosis are at a greater chance of developing lung cancer.
- Exposure to carcinogens: Experience to definite carcinogens in office, like arsenic, chromium, nickel, and chemicals used in the manufacture of plastics and textiles, can increase the risk of lung cancer.
- Radiation: High doses of radiation, such as those used in medical treatments such as radiation therapy, chances to developing lung cancer.

- Diet and nutrition: A diet low in fruits and vegetables and high in processed foods may increase the risk of lung cancer.

1.3.1 Lung Cancer Symptoms

The initial indications of lung cancer can vary based on the specific type, but they might involve a cough that worsens or difficulty in breathing. These signs may worsen or intensify as the malignancy progresses. Early lung cancer may not show any signs. However, there are several indicators to watch out for, such as those listed below [7]:

- A bloody cough
- Chest pain that becomes worse when laugh, cough, or take a big breath
- Back pain that gets worse when cough, laugh, or take deep breaths
- Increasing discomfort in the shoulder when coughing, laughing, or breathing deeply.
- Quick, rapid-onset tininess of snuffle while doing daily tasks
- Weight reduction
- Weakness or exhaustion
- Loss of appetite
- Persistent bronchitis or pneumonia
- Hoarseness or wheezing
- Changes in the look of the fingers
- Swelling of the face or neck
- Pain or difficulty swallowing
- Bone pain
- Headaches
- Dizziness
- Limbs
- Jaundice
- Lumps in the neck or collarbone region
- Coughing up rust-colored mucus or blood
- Breathing difficulties
- Weakness or fatigue
- Recurrent infections like bronchitis or pneumonia

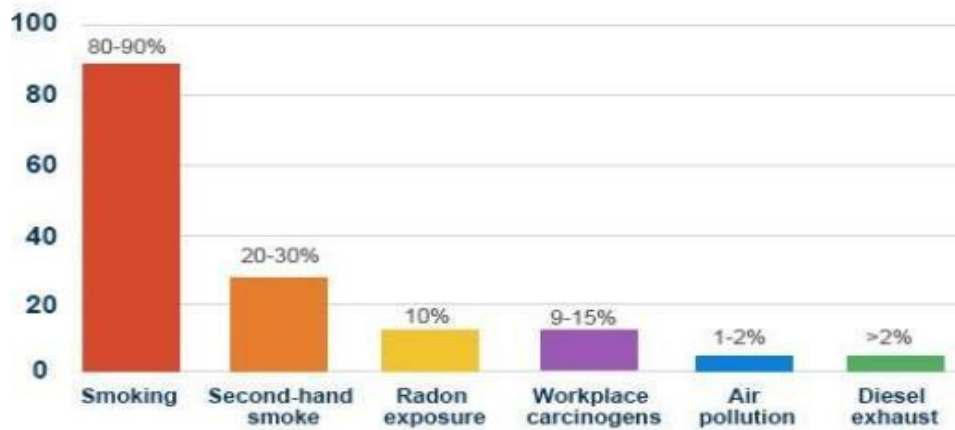


Figure 1. 2 Lung Cancer Causes [6]

1.3.2 Lung Cancer Risk Factor

There is different risk factor which is responsible for not only lung cancer but other types of cancers as well. The cancer gets involved in human body because there may be the possibilities of one or more types of such kind of risks. Following are some kinds of risks:

a) Using tobacco or smoking regularly [7]

Passive smoking is not considered smoking tobacco; it must be done actively.

b) Passive Smoke

Smokers who are not actively smoking inhale tobacco smoke, which is known as passive smoking.

c) Issues in Environmental and Workplace

i. Radon

Radon, a radioactive gas occurs naturally. Uranium miners constitute the majority of those who are exposed to radon at work; they have been shown to have a 1.6–3.8-fold increase the risk of lung cancer.

ii. Arsenic

Inhalation of particles containing arsenic is the main way mining and industrial workers are exposed to this chemical element.

iii. Polycyclic aromatic hydrocarbons exist.

Polycyclic aromatic hydrocarbons (PAHs) form when organic materials like tobacco, coal, crude oil, natural gas, and garbage are burned. Food, cigarette smoking, industrial air pollution, and urban areas are all sources of PAHs.

iv. Cadmium

Seawater and the crust of the Earth contain the element cadmium naturally. A wide range of industrial operations uses it.

v. Asbestos

In the past, asbestos, a logically occurring lifeless, often used in building supplies. Lung cancer risk is more when breathing in asbestos fibers, especially for long time.

vi. Silica

Silica is a natural source that contains silicon and oxygen. Numerous of these natural sources emit minute dust particles that are inhaled and have been related to lung cancer.

vii. Iron & Steel's foundation

Individuals employed in the iron and steel industry encounter multiple factors that increase the risk of lung cancer while they engage in tasks like pouring molten metal to produce new shapes or items.

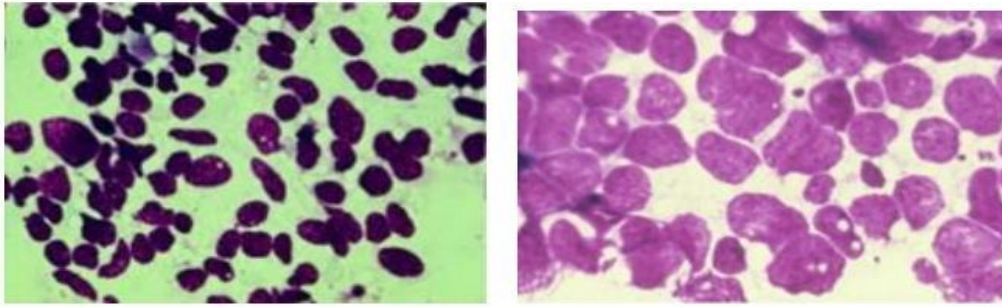
1.4 Types of Lungs Cancer

Lung cancer divided in two parts: non-small cell lung cancer and small cell lung cancer. These are differences based on whether cancer cells are present when looked at under a microscope and how they act.

1.4.1 Small Cell Lung Cancer

The growth rate of small cell tumors is often greater than that of non-small cell cancers. Figure 1.4 (a) displays scleroderma cancer cells. When first diagnosed, tumors in the mediastinum, supraclavicular lymph nodes, or hemithorax of origin are only sometimes observed in people with small-cell lung cancer. Restricted stage and Expansive stage are two categories.

- **Limited stage:** An area of the chest, the respiratory organ, or nearby lymph nodes is affected by cancer at this stage.
- **Extensive stage:** Cancer has progressed to other parts of the chest or the anatomy at this point.



(a)

(b)

Figure 1. 4 SCLC and NSCLC cells [8]

1.4.2 Non-Small Cell Lung Cancer (NSCLC)

The group of lung tumors called non-small cell lung cancer includes cancer of squamous cells and adenocarcinoma. Thinness of breath, weight loss, a persistent cough, and blood in the cough are all signs. Surgery, chemotherapy, and radiation are all forms of treatment. Figure 1.4 (b) shows the NSCLC cancer cells. Non-small cell carcinoma accounts for 75-80% of all lung cancers. It's divided into four unique stages (I, II, III, and IV). Patients with non-small cell lung cancer (NSCLC) are more than 70% in stage III or IV at the time of their diagnosis. Adenocarcinoma, SCC, and LCC are all types of NSCLC carcinoma [8].

- **Adenocarcinoma:** Adenocarcinoma, the most frequent form of carcinoma in women, is the most common form of non-small cell lung cancer in women. This type of cancer is becoming more common. Adenocarcinoma typically emerges in the mucus-producing glands encircling the airways of the lungs. As with other types of cancer, it able to grow on its own.
- **Squamous Cell Carcinoma (SCC):** Among men, squamous cell carcinoma stands as the most common form of cancer, if it's always within the broader airways. Squamous carcinoma, like other non-small cell lung cancers (NSCLCs), may be slow-growing.
- **Large Cell Carcinoma (LCC):** LCC occurs less frequently than other NSCLCs, and its cells are bigger. In the beginning, the disease appears to be limited to the smallest of the breathing tubes. When other forms of lung cancer have been ruled out, large-cell carcinoma is found. Large-cell carcinoma grows and spreads (metastasizes) more rapidly than other forms of non-small cell carcinoma and does so at an earlier stage.

1.5 Tests for Lung Cancer

Although some lung cancers are detectable through screening, most lung cancers are first diagnosed when they are causing significant symptoms. Lab analysis of lung cells is used to make the final determination of lung cancer status [9].

1.5.1 Medical History and Physical Exam

To learn more about symptoms and other risk factors, a doctor inquiry about medical history. Besides checking for indications of lung cancer, a doctor also does a physical examination and need to undergo additional testing if medical history and physical examination suggest that might have lung cancer. Lung biopsies and tomography testing are two options.

1.5.2 Imaging Tests to Look for Lung Cancer

Utilizing X-rays, magnetic fields, sound waves, or radioactive substances, it's possible to generate visuals of the body's internal structures. Imaging studies are conducted before and after diagnosing lung cancer for a range of purposes, including:

- To examine potentially cancerous growths in areas of concern.
- To determine the extent of cancer spread.
- To assess the effectiveness of the treatment.
- Keep an eye out for any signs that the cancer has returned following therapy.

a) Chest X-ray

To check for any abnormalities in the lungs, a doctor most likely orders an X-ray of the chest. A doctor may prescribe more testing if something unusual is found.

b) Computed Tomography (CT) Scan

X-radiation is used in a CT image to provide highly detailed cross-sectional images of the human anatomy. A CT scanner, as opposed to a traditional X-ray, collects a slew of images, which are then stitched together on a computer to produce an image of the part of the body being examined. Routine chest X-rays are less likely to detect lung cancers than a CT scan. Radiation therapy may also be used to identify any lung tumors, as well as enlarged lymph nodes that may be harboring cancerous cells. It is also possible to perform this test to screen for tumors in other organs such as the adrenal glands, liver, and brain [10].

c) Magnetic resonance imaging (MRI) Scans

Similar to CT scans, MRI scans offer highly detailed visuals of the body's soft tissues. Instead of using X-rays, MRI scans utilize radio waves and strong magnets. These scans are frequently employed to identify the spread of lung cancer to the brain and spinal cord.

d) Positron emission tomography (PET) Scan

It is possible to have both a PET and a CT scan done at the same time on a single scanner. CT and PET images can be compared to identify areas with elevated radioactivity on PET scans. This PET scan is the most commonly used for patients with lung cancer.

PET/CT scans can be useful:

- Physician suspect's cancer has spread but is unsure of the exact location of the cancer's spread. In some cases, they can indicate the progression of malignancy to the liver, bones, or adrenal glands. The brain and spinal cord cannot be seen as well with these methods.
- However, there is little evidence to support their use in determining if treatment is effective to predict cancer. PET/CT scans are not routinely recommended by clinicians for individuals who have undergone treatment for lung cancer.

e) Bone scan

It is necessary to inject the patient with a little amount of low-level radioactive material, which accumulates mostly in aberrant sections of the bone, to perform a bone scan. Using a bone scan, doctors can determine if a malignancy has migrated to the bones. Nevertheless, PET scans can typically tell if the disease has progressed to the bones [11].

1.6 Introduction to Deep Learning

The widespread accessibility of advanced calculating resources, the popularity of deep learning approaches utilizing deep neural networks has been on the rise. When dealing with unstructured data, deep learning can process a huge number of features, which gives it greater power and flexibility. Each layer of the deep learning algorithm can extract information sequentially and pass them on to the next layer. First, the lowest-level features are extracted, and then the higher-level features are combined to create a comprehensive image representation [12].

The initial iteration of artificial neural networks (ANNs) featured neural layers with perceptron's that had limitations on their computational capacity. These perceptron's were capable of executing only a finite number of computations per second. In the subsequent generation, error rates were reevaluated and retroactively propagated, with the breakthrough of addressing backpropagation challenges emerging alongside the advent of the Restricted Boltzmann machine. This milestone paved the way for the development of various other network architectures. A chronological representation illustrating the evolution of deep models alongside the classical model showcases this progression. When considering the volume of data they can analyze, deep learning classifiers significantly outperform traditional learning methods on an extensive scale. While conventional machine learning algorithms tend to stabilize as training data increases, deep learning algorithms continue to advance and improve in efficacy as data expands.

1.7 Machine Learning Approaches

Machine learning is a subset of Artificial intelligence (AI) allows computer programs to predict events added precisely, even lacking explicit programming. Machine learning algorithms predict original crop ethics by processing ancient information as involvement. Deep neural networks find applications across various learning scenarios, such as supervised, unsupervised, reinforcement, and hybrid approaches.

A. Supervised Learning

A function f for translating input variables (expressed as letters or numbers) to output variables (represented as letters or numbers) is taught through the process of supervised learning.

$$Y = f(X) \quad (1)$$

In order to predict the output (Y) for a new input, the learning approach approximates the mapping function (X). During training, the predictions generated can help refine the output. Learning can be halted once all inputs have been trained to achieve the intended output. Techniques like regression, Support Vector Machines for classification, and the utilization of Random Forests can be applied to address both classification and regression challenges [13].

B. Unsupervised Learning

In unsupervised learning output cannot be plan; simply have input data. This course

seeks to teach students about data by teaching them how to model data distribution. It is possible to identify interesting patterns in data using algorithms. For clustering and association challenges, unsupervised learning proves to be the most suitable approach. Techniques such as K-means for clustering and Apriori for association are employed within unsupervised learning to address these specific problems.

C. Reinforcement Learning

Reinforcement learning uses a reward and punishment mechanism to teach an algorithm. Learned from its environment using an algorithm or a robot. Agents are rewarded for good work and penalized for bad work. Think of a self-driving car as an example; when it gets to its goal, it gets a reward, and when it gets off course, it gets fined. There may be two possible reward states in the case of a computer program that plays chess: winning and getting checked. The agent's goal is to achieve the best possible balance between reward and penalty. Rather than being explicitly instructed on how to solve a problem, a reinforcement learning algorithm solves it by itself [13].

D. Hybrid Learning

The term "hybrid learning" refers to designs that include both generative and discriminative components. Multiple architectures can be combined to create a hybrid deep neural network. By utilizing action bank features, the recognition of human actions can be greatly enhanced in terms of accuracy.

1.8 Fundamental of Deep Learning Architectures

Despite the longer training times of deep structures, they outperform a simple ANN in terms of accuracy. However, technologies such as transfer learning and Graphics Processing Unit (GPU) computing can shorten the training period. Network architecture has a significant role in the success or failure of neural networks. The following section provides an overview of some of the most relevant deep learning architectures [14]. The growth of several scientific domains, including speech and picture recognition, gaming, image-related, song, etc., has been greatly influenced by deep learning (DL). Some of the issues that the deep learning approach resolved were thought to be intractable by other computerized techniques. As a result, this technique has become very important for medical imaging, as seen by the rising trend in the number of publications in the area. Before knowing the characteristics of deep learning, it is crucial to comprehend the linked fundamental subjects. These subjects include neural networks, ML and PR (Pattern recognition). This section

summarizes a quick introduction to deep learning for medical image processing.

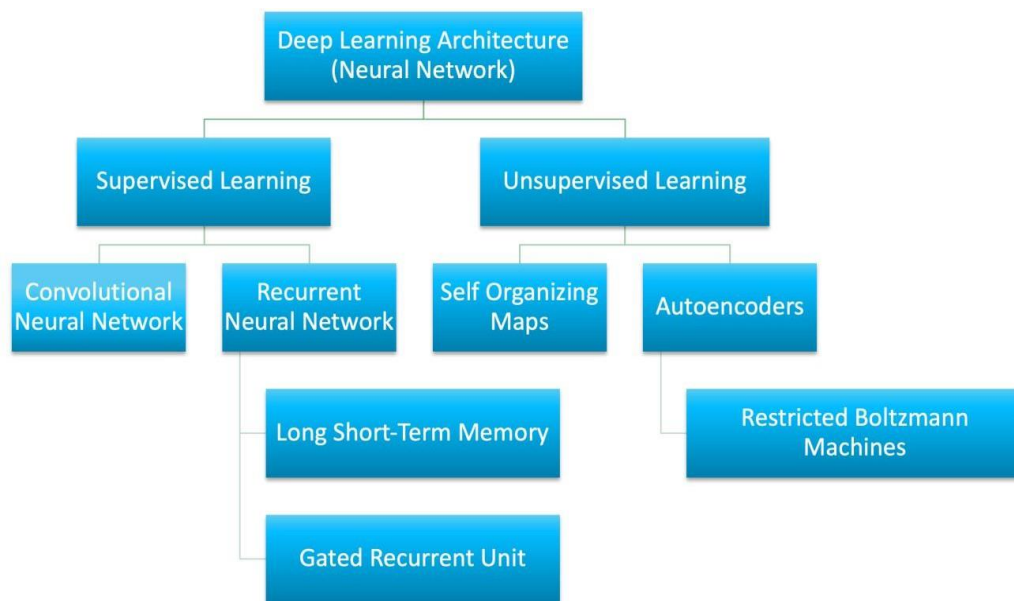


Figure 1.2 Fundamental Deep Learning Architecture [15]

A. *Unsupervised Pre-Trained Networks*

Before being utilized for prediction, a model is trained unsupervised using unsupervised pre-training. Following are a few examples of unsupervised pre-training architectures [16].

- **Autoencoders:** These can be used for a variety of purposes, including data reduction, novelty discovery, and anomaly detection. The initial layer of an autoencoder is designed as an encoding layer, and transposition of that layer serves like decoding layer. After that, may use the unsupervised technique to teach it how to recreate the input. Fix the weights of that layer after training. After that, go to the next layer until the deep net has been pre-trained through its entirety. Re-visit our original goal of employing a deep net and use stochastic gradient descent for optimization, initializing with weights acquired during pre-training.
- **Deep Belief Network (DBN):** Training the deep belief network begins with learning the features of the first layer of data. Then activate the features learned in the following layer. Continue until the last layer is applied. To train the DBNs, Restricted Boltzmann Machines (RBM) are used, and the feed-forward network is used for fine-tuning. As opposed to other deep networks, DBN learns hidden patterns at a global level.

- **Generative Adversarial Network (GAN):** It is Ian Good fellow who introduced the term Generative Adversarial Networks (GAN). Networks for generators and discriminators are both part of it. The discriminator verifies the generated content while the generator creates it. Images are generated by the generator and then judged by the discriminator based on how realistic they appear to be. There are two players in GAN and it is thought of as having an optimal minimax strategy. GANs employ convolutional and feed-forward Neural Nets in their computations [17].

B. Convolutional Neural Networks

CNNs are mostly used to analyze images, but they can also be used to analyze other types of data. It does this by giving different parts of the image varying weights and biases and then using that information to make distinctions. Compared to other techniques for classifying data, this one necessitates less preparation. The spatial and temporal relationships in a picture are captured by the CNN using appropriate filters. Some of the CNN architectures are LeNet, AlexNet, VGGNet, GoogleNet, ResNet, and ZFNet. CNNs are primarily employed in the detection, segmentation, and captioning of objects [18].

C. Recurrent Neural Networks

Output from previous states is used as an input in RNN. The RNN's hidden layers are capable of storing data. Using the output from the previous state, the concealed state is continually being updated. Time series prediction is possible because RNNs can remember prior inputs, known as Long-Short Term Memory (LSTM) [19].

1.9 Applications of Deep Learning

The development of deep learning (DL) has helped a wide range of scientific fields, including voice recognition, image identification, gaming technology, music, and many more. Deep learning has proven to be more effective than conventional computer methods in several problems that were previously assumed to be intractable. As a result, the trend of more publications in this field is a sign of the technology's increasing significance for medical imaging. One must first have a clear understanding of the underlying ideas to truly understand deep learning. Deep learning techniques are occasionally denoted as "deep neural networks" due to their frequent utilization of neural network architectures. The depth of a network is often indicated by the number of concealed neural network layers. Deep networks can encompass as many as 150 hidden layers and old neural network contains two

or three layers only. All labeled data along with neural network structures that autonomously extract features while directly learning from the data, are employed to train model. Deep learning used in wide range of industries, from healthcare to manufacturing. In this section, we'll go through several of them in greater detail [20].

- **Self-Driving Cars:** There is a lot of interest in self-driving cars among techies, and deep learning is the driving force behind that interest. Machine learning models are built from a variety of data sets and then tested in a controlled environment to ensure that they are ready to learn. There is an Uber Artificial Intelligence Lab in Pittsburgh that isn't simply preparing driverless vehicles but also includes many inventive ideas. For instance, they're working on making driverless food delivery a reality. This data is being used to create short and sophisticated models to direct traffic, identify routes, and real-time components such as volume of traffic and road stoppages. Additionally, self-driving car developers are concerned about dealing with unusual situations. Deep learning algorithms follow a standard sequence of testing and execution that ensures safe driving in an increasingly diverse range of conditions. As a result, may predict that deep learning lead to the development of many more intelligent devices in the future [21].
- **Virtual Assistants:** Using deep learning in the form of virtual assistants is one of the most prominent uses of this technology. Our daily lives are increasingly reliant on voice-activated assistants like Alexa, Siri, and Google Assistant. It's possible to have an optional human interaction experience with these helpers every time speak with them. Deep learning use virtual assistants to learn themes, such as favorite restaurants, frequented locations, and favorite music artists. To carry out instructions, they analyze everyday human speech patterns to decipher instructions. In reality, virtual assistants are at service, as they can do everything from getting things done and answering phone calls to coordinating tasks among co-workers. Virtual assistants also can translate text, take notes, and book appointments. Virtual assistants can also help with email duplications thanks to deep learning applications like text generation and record synopses [22].
- **Virtual Recognition:** When someone have a lot of old photos, but only a few that want, run into difficulties. As a result, had to spend a lot of time and energy picking out photos. When it comes to arranging photos based on specific elements of the image, such as faces or a mix of people in the photo or the day and time they were taken, deep learning comes into play. Visual recognition algorithms that are state-of-the-art are needed to find a specific photograph in a large database (such as Google's image library), which

requires several layers from basic to advanced. The use of convolutional neural networks, TensorFlow, and Python in this area of advanced media management is enhancing the development of this portion of advanced media management [23].

- **Natural Language Processing:** When it comes to human-computer communication, natural language processing (NLP) is a key tool. NLP is the natural manipulation of normal languages, such as speech and text, by software. Deep learning has a wide range of applications in the field of NLP. It is attempting to do something quite similar using Deep Learning by training robots to pick up on linguistic nuances and formulate appropriate responses. Deep learning is gaining momentum in NLP subdomains such as question answering, text classification, Twitter analysis, and sentiment analysis [24].
- **Entertainment:** Deep learning is being used by Netflix and Amazon to provide viewers with a more personalized experience, taking into account factors such as a viewer's preferences, the amount of time they have to watch a show, and so on. Using Deep Learning AI, cameras can analyze human nonverbal communication to capture virtual characters. Because of deep learning and its focus on expression and pattern recognition, content alteration and auto-content production are real [25].
- **Fraud Detection:** Fraud detection is yet another key use of deep learning. Financial sector and Bank can greatly benefit from this technology since individuals increasingly rely on digital transactions. Financial institutions stand to save billions of dollars in recovery and protection costs because of the development of autoencoders in Keras and TensorFlow that can detect credit card fraud. Trust in banking and Internet transactions is a result of this superior technology. Preventing and detecting fraud requires the ability to spot outliers in customer communications and credit scores, as well as a keen eye for odd behavior and other irregularities. Therefore, ML approaches and neural networks are active for fraud recognition, which includes classification and other regression techniques [26].

1.10 Challenges

Some of the challenges confronted by various lung cancer detection techniques are deliberated as follows:

Challenges to Detect Lung Cancer:

Most of the time, humans identify the signs such as coughing, shortness of breath and chest pain on a delayed note or probably they ignore it at an early stage which leads to further

creating complications in treating lung cancer.

- Although medical science has grown in terms of technology in recent few years, still computed tomography (CT) and chest X-rays are only two techniques which are commonly used for identifying lung cancer. So, one of the biggest challenges faced is that there is no system which help doctors to identify lung cancer at an early stage.
- Another challenge is that most of the Artificial intelligence (AI) system developed are only identifying whether the patient has lung cancer or not, but there is a need to design a system which not only predicts lung cancer, but should also identify the size of lung cancer tumor which help to identify the lung cancer stage. That help to treat the patient with the correct diagnosis and that too in early stage.
- Current techniques had limitations in reducing the error margins and creating a more reliable and dependent lung cancer prediction system.
- Current data acquisition techniques for lung cancer detection like CT scan or X-Ray images need to be pre-processed for better and more accurate results.

1.11 Motivation

The lungs are the primary respiratory organs. It is leading cause of cancer mortality in people. Deaths of lung cancer outnumber those are colon, cervix, breast cancer shared. Coughing is the most common sign of lung cancer. It's important to pay extra attention to it because many people with lung cancer have a cough since they smoke and also have a lung problem called chronic obstructive pulmonary disease. Both of these can make them cough more.

Hence special attention needs to be given to identify the lung cancer by using advance Artificial Intelligence techniques become a need of today in the medical field. Only identification and prediction are not sufficient for lung cancer diseases but tumor size detection is also important and needs to be focused in today's scenario to fight against this lung cancer disease. Lung cancer must be staged correctly for doctors to decide on the patient's course of therapy. It takes a lot of time and effort to complete these activities manually.

This research gap motivates us to take up this research. Advanced deep learning methods with optimization concepts, one detects lung cancer early. Detecting lung cancer initially help Doctors/Surgeons to give the right kind of treatment at an early stage, so that the survival rate of patients on the higher side. Additionally, after administering a specific

medicine or therapy to a patient with lung cancer, the doctors evaluate the patient's response to treatment using the RECIST (Response evaluation criteria in solid tumors) criteria. RECIST is a standard way of measuring tumor response to therapy. It provides objective criteria for determining whether a tumor disappears, shrinks, stays the same, or grows.

1.12 Research Objectives

The primary goal of this study is to create a lung cancer prediction model using deep learning. To accomplish this task following objectives were proposed [27].

- I. To review various state-of-the-art lung cancer prediction models and develop a new weighted feature extraction.
- II. To design and develop a deep-learning model to predict lung cancer.
- III. To validate the proposed model by comparing it with other conventional models.

1.13 Major Contribution of Thesis

- I. To propose a novel weighted feature extraction deep learning model based on feature linking and optimization of prediction of Lung Cancer. For this proposed learning, two datasets from Kaggle have been derived.
- II. T-Distributed Stochastic Neighbor Embedded(t-SNE) and Principal Component Analysis (PCA) feature selection methods were combined via weighted optimization technique to propose the novel attribute correlation-based technique using Deep Learning. The proposed SA-SL_nO methodologies optimized the weights to increase the effectiveness of Lung Cancer Prediction.
- III. Bringing an innovative self-adaptive SL_nO algorithm to improve weights in innovative feature correlation-based weighted feature extraction stage & augment RNN's hidden neurons to lower MSE among actual and predicted results.
- IV. Compare the planned method with current optimization and various performance measures to confirm the concert of the projected lung cancer analysis prototypical.
- V. The proposed SA-SL_nO model improves the performance of the prediction of Lung Cancer compared to different machine learning algorithms & Conventional Optimization techniques.

1.14 Thesis Outline

The thesis has been organized into six chapters. An overview of each chapter is given below:

Chapter 1: This chapter introduced Lung Cancer along with a discussion of its various types and signs of lung cancer. The motivation for the research has been discussed. Furthermore, the chapter briefly overviews about ML and DL and how they apply to medical image analysis. The problem statement, objectives, and significant thesis contributions are discussed further. The thesis organization is provided at the end of the chapter.

Chapter 2: This chapter presents a comprehensive literature review of the research on lung cancer diagnosis and treatment analysis. It includes a literature review on how deep learning and various machine learning algorithms are used to find lung cancer. Lastly, the chapter summarizes the literature presented and discusses the research gaps.

Chapter 3: This chapter describe feature selection methodologies, and techniques like t-SNE and PCA. It also introduces novel approaches that Weighted Feature Extraction Model based on Attribute Correlation. The description of relevant mathematical equations within the context of feature selection and the weighted feature extraction model. Moreover, the lung tumor datasets used for the proposed studies are discussed in the chapter.

Chapter 4: This chapter describes the lung cancer diagnosis based on optimized deep learning methods, optimized weighted feature mining based on the novel characteristic, and proposed lung cancer detection architecture. Proposed model objective for lung cancer diagnosis to enhance the efficiency of hidden neurons within the Recurrent Neural Network (RNN) during the classification phase. This is complemented by an optimization of characteristic-correlation-based weighted feature extraction within the proposed RNN framework. This chapter also describes algorithm and flowchart of proposed model.

Chapter 5: This chapter covers the results obtained from proposed models implemented using deep learning. The performance measures of the proposed model are accuracy, precision, recall, and F1-Score and it compared with the recent existing model. Using an optimized deep learning approach and attribute correlation-based optimized weighted feature extraction, the proposed model has sent a benchmark for prediction of lung cancer.

Chapter 6: This chapter, conclude research work shown of thesis, and the limitations of proposed work and the future scope are also discussed.

Chapter Summary

In this chapter, introduced the Lung Cancer and how it impacting on human body. It is also focused on different types of cancer, test of cancer, introduction to deep learning, fundamental of deep learning approaches, application of deep learning, motivation and challenges, the contribution of thesis, research objective and thesis outline. Also discussed about thesis chapter summary, research objectives, Advance machine learning and deep learning techniques.

Chapter 2

Literature Review

Literature review on deep learning-based approaches for lung cancer prediction, lung cancer classification. For this, done a thorough research on various articles in recent years on lung cancer prediction system. Concepts like features extraction, lung cancer different stage classification methodologies and patients' response to lung cancer treatment has been analyzed in this literature review.

2.1 Introduction

Recurrence is a problem for a considerable number of lung cancer patients, even after they have had their tumors surgically removed. Imaging characteristics in analytical modelling of NSCLC recurrence-free subsistence has been attempted. For identifying mutation subgroups in malignancies, radio mic features have been used as well. However, it is uncertain whether these features have any effect on the final prognosis for patients. CT scans can detect genetic mutation subtypes in lung cancer with imaging correlates, according to studies. CT radionics-based deep learning is used in this work to establish the connection between quantitative imaging parameters indicative of mutation status and recurrence-free survival [28].

2.2 Lung Cancer Diagnosis Approach

Wang et. al [31], the author used two phases, one for node identification and another for risk estimate phase. The 3D-CNN & RNN were used for the node identification to reduce the loss and improve the recall. In the risk estimate step, the proposed 3D-CNN Model was tested with real-time medical data on lung cancer. Next, the regression model was integrated to align with the ground truth labels, which represented the level of morbidity for each CT scan. To guarantee the recommended model work as intended, it verified using the lung cancer module with the greatest risk

Suresh et. al [32] proposed end-to-end CNN learning model removed the self-learned structure model. Subsequently, results of proposed CNN model were compared through conventional CNN model. For this study, 1018 images from the Lung Cancer dataset were utilized. Image processing techniques were active to detect Region of Interest (ROI) and accurately pinpoint the Lung Cancer region, thus enhancing accuracy. Generative Adversarial Networks (GANs) were then implemented on this preprocessed image dataset. The outcomes derived from these proposed CNN models demonstrated an enhancement in

the performance of Lung Cancer prediction.

Palani et. al [29] state that survival rates may increase with proper treatment. The incidence of lung cancer varies by ethnicity and geographic location. In affluent nations, the five-year prevalence of lung cancer is roughly 15%, compared to 5% in impoverished nations. A very small proportion of computer tomography (CT) screening in high-risk groups results in a false-positive rate of 96% but there is a 20% reduction in mortality growth in lung cancer.

Pradhan et al. [27], implied a lung cancer detection model to lower the mean square error. The feature extraction process has been carried out using PCA (principal component analysis) and t-distributed stochastic neighbor embedding (t-SNE). Additionally, a self-adaptive sea lion optimization method (SA-SL_nO) described for an enhanced correlation-based weighted feature extraction approach is used most current meta-heuristic techniques is used enhance weights, hidden neurons increased through SA-SL_nO that is provided.

Abdul et al. [30], state that lung cancer in humans can have a major impact on an individual's quality of life. Lung cancer tumors might be difficult to identify when they are still small. Detecting tumors early on has the potential to save a great many lives. Even though CT scan pictures can be used to create an ALCDC system that accurately detects and classifies lung cancer, the task of creating such a system is difficult. Hand-engineered systems for detecting and classifying lung cancer are currently in use, but their accuracy and other performance metrics are limited. A system for ALCDC based on images from CT scans is now available as a consequence of DL's impressive performance in several recognition-related tasks. To detect and categorize whether lung cancers are malignant or benign, an ALCDC system is constructed utilizing a CNN model. According to the findings, the SA-SL_nO system has a precision rate of 97.2%. A side-by-side evaluation demonstrates that the SA-SL_nO system is superior to currently available cutting-edge technologies. In medical diagnosis research and health care systems, the proposed SA-SL_nO beneficial.

Mukherjee et al. [31], have implemented lung disorders have always been a source of fascination for scientists in the medical field, and that hasn't changed. An early detection system like this may only reduce the risk to human life by detecting cancerous development at an early stage. Over time, a few structures have been proposed, although the vast majority are merely speculative plans. and to establish an AI-based deep neural system framework for the detection of lung cancer. This approach relies heavily on the precision that can be achieved through DL. The framework contains a variety of approaches, such as

classification, pre-processing, enhancement, segmentation, feature extraction, and neural network identification. Deep Learning model in today's era has possible to detect and predict lung cancer disease affordable cost which very useful for medical practitioners.

S. R. Jena et al. [32] state that scientific studies should be geared toward finding a cure for lung cancer because it is the deadliest disease. Cancer can be completely cured if diagnosed early enough. To increase accuracy of cancer finding, a new method was developed i.e., Support vector machines and neural networks, together with image processing techniques, are employed extensively for cancer detection.

Haga A. et al. [33] have developed a deep learning-based method for predicting mortality risk in patients with lung cancer that incorporates both chest CT scans and coronary artery risk scores. End-to-end frameworks that use hybrid imaging features instead of relying solely on automated feature extraction are referred to as Hybrid Risk Networks (HyRiskNet). It has been shown that deep learning approaches may be used to predict mortality from chest CT scans. When compared to neural networks with only picture inputs and other typical semi-automated scoring systems, the experiment's results show that HyRiskNet can achieve superior performance. Feature extraction can further have used to extract more features using CNN.

Sharma et al. [38] highlighted that microarray data classification has gained significant attention due to the widespread availability of gene expression assays online. The main focus has been on achieving high classification accuracy and identifying the smallest set of genes for each population. To address this, C-HMOSHSSA was introduced as a gene selection method, utilizing a combination of MOSHO and SSA to achieve similar objectives. The challenge of balancing convergence and diversity is a common issue in real-life optimization problems with multiple objectives. SSA is effective in preserving diversity, although the requirement to maintain necessary data, while MOSHO demands less computational effort and is employed for data storage. Consequently, a hybrid algorithm was developed by combining SSA and MOSHO to facilitate exploration and extraction. In proposed model, a new method of feature selection method discovered that can deliver the greatest results with the best classification accuracy and the lowest expenses.

Cao et al. [34] proposed an ensemble learning architecture based on a three-layer CNN. Feature extraction is accomplished with the use of an Offline hard mining approach. False positive rate calculated. Multi-branch ensemble learning architecture is detected using a

three-dimensional convolutional neural network (CNN). For judgement of lung cancer recognition algorithms by Gunaydin et al. [35] has been given. As a statistical method for discovering new patterns, PCA is employed. The extraction process makes use of eigenvector and entropy properties. Algorithms like KNN, SVM, or naive Bayes are used to classify data. However, the truly significant weighted characteristics for diagnosing lung cancer were uncovered in the proposed investigation.

Senthil et al. [42] developed image segmentation algorithms for medical images, particularly focusing on enhancing the clarity of CT scans. Both accuracy and convergence time play significant roles in the effectiveness of these segmentation algorithms. To overcome the challenges connected with medical image separation, imperative to explore and incorporate novel evolutionary algorithms. Lung cancer diagnoses among males have reached alarmingly high rates globally. In the proposed approach, the SA-SL_nO algorithm is employed for lung cancer detection through weighted feature extraction. This method has demonstrated superior accuracy compared to the other examined methods.

Shanthi et al. [43] introduced a symbolic method for data analysis that incorporates a range of quantitative data. The author explores diverse feature selection strategies for predicting histological subtypes of lung cancer, utilizing symbolic data or radio mic features. The features were extracted using GLCM and Gabor filters. After normalizing with the Z-score, the features were combined through concatenation. Experimental results verified the advantage of proposed system over alternative approaches. The initial phase involved the use of VGG16 and ResNet50 to extract images. The classification was achieved through the employment of the region-CNN method.

Rahman et al. [44] explored the application for detecting lumps. Pre-processing data initiates with blurring and thresholding procedures. Zuo and colleagues delved into multi-resolution identification of pulmonary nodules, employing CNN and candidate classification techniques. CNN is utilized to extract lung imaging characteristics [45]. Classification is executed using a multi-resolution CNN model. The process involves detecting and classifying pulmonary nodules in chest CT scans using CNN. Extracted images are fed into the CNN, and nodule identification employs metrics like false positive and false negative rates, along with attributes such as size and position [46]. The segmentation process is facilitated through clustering, and the Faster R-CNN approach is employed for data classification. In terms of future directions, it is suggested that a novel feature selection technique be implemented to enhance the accuracy of classification.

Yu et al. [47] investigated imaging biomarkers applicable for diagnosing and predicting the pathological stage of non-small cell lung cancer (NSCLC). The study encompassed participants across various stages, ranging from Stage I to IV of NSCLC. The participant cohort was segregated into three distinct groups: a training group, a testing group, and a validation group. To address the issue of imbalanced datasets, a new dataset was generated and the SMOTE Algorithm was employed to achieve a balanced distribution among classes. The training and testing groups were randomly divided dataset. The significance of CT scan features was assessed using the mean reduction in Gini impurity induced through the Random Forest (RF) algorithm. The study employed the Kaggle dataset for research and experimental analysis. The proposed model incorporates the PCA and t-SNE approaches for feature categorization. The model's performance evaluated using various performance parameters, including precision, recall, and accuracy.

Palani et al. [48] introduced an innovative Internet of Things (IoT)-based predictive model for continuous monitoring and early detection of lung cancer, along with providing medical guidance. The approach involves employing the fuzzy clustering technique to effectively segment images through the extraction of transition regions. Lung cancer imaging data was used for applying the Fuzzy C-Means Clustering Algorithm, which facilitated the classification of transition area features. The study utilized Otsu threshold techniques to extract transition zones from lung cancer images, aided for segmentation. Furthermore, edge lung cancer images underwent morphological cleaning and were filled with object regions. The paper also proposed a novel incremental classification algorithm that integrates with methods. For experimentation, standard images from databases were combined with real-time health data collected from IoT devices worn by patients. The results of the experiments conclusively demonstrated that the performance of the proposed model surpassed that of existing similar studies, confirming its efficacy in lung cancer detection and prediction.

El-Askary et al. [49] introduced an early detection strategy for lung nodules, aimed at justifying the likelihood of lung cancer progressing to advanced stages. The team employed the Random Forest (RF) machine learning classifier for both the identification of lung nodules and the classification of soft tissues as containing nodules or not. A five-stage model was developed and tested using 165 samples from the Lung Image Database Consortium (LIDC). The initial stage involved image acquisition and pre-processing. Subsequently, 119 attributes from CT scans were extracted and subjected to analysis. LIDC maintains a repository lung images. Improvement feature vectors included deletion of duplicate instance

of execution of growth of abnormal tissue. Fine-tuning of RF parameters was conducted afterward. The extracted feature sets were evaluated across various configurations to determine the optimal classification scores. This study utilized image attributes and wavelet transformation to achieve the most effective results.

Fotouhi et al. [50] determined predictions regarding individuals with non-small cell lung cancer (NSCLC) can be made utilizing phenotypic radiomic characteristics extracted from CT images. 186 ill persons of NSCLC underwent CT image analysis using Pyradiomics. To create training and validation datasets, a 3:1 ratio was randomly allocated using the Synthetic Minority Oversampling Technique (SMOTE). Various Random Forest (RF) models were trained through a hyper-parameter grid search employing ten-fold cross-validation, with precision or recall as the evaluation criteria. The decision threshold of innovated prototype was an evaluation of the final model conducted using a ROC plot and a prediction accuracy test. The segmentation of scans from the 186 individuals resulted in a total of 1218 features. Model performance was assessed using recall, and an optimal decision threshold of 0.56 was identified. The model achieved an accuracy of 89.33% and an AUC score of 0.9296. RF classifiers with hyper-parameter tuning demonstrated superior evaluation expecting the existence status of NSCLC patients, suggesting their potential as automated classifiers for patient stratification. The proposed model's evaluation encompassed accuracy, precision, and recall. The study utilized the Kaggle dataset for research purposes.

Alam et al. [36], studied that SVM classifiers used effectively distinguish and forecast lung cancer. Cancer detection was made possible with the use of a multi-stage categorization Lung cancer risk might be estimated using the technique. At each step of the categorization process improved and segmented the images. To improve the photos, used image scaling, contrast enhancement, and color space transformation. Watershed-based segmentation controlled by thresholds and markers was used for segmentation. When it came to making the distinctions, an SVM binary classifier was used.

Lobo et al. [37] studied that classification and segmentation techniques for the detection of lung cancer from CT images. It's utilized for pre-processing. An algorithm known as GLCM is used to extract texture features such as dissimilarity and dissimilarity probability, as well as features such as contract and correlation. Classification is accomplished through the application of an SVM classifier. Segmentation is accomplished via FCM, to identify the best attributes. The author concluded that a better feature selection method can expand evaluation of innovative technique.

As per Kido et al. [53], Convolutional Neural Network containing CNN-specific characteristics have been employed within CNN framework identification and sorting of lung irregularities. The process begins with feature extraction using CNN. The classification task is executed by employing a Support Vector Machine classifier. The validation set is subjected to evaluation using a multiclass SVM. Additionally, the study explored the application of a random forest classifier for the classification of imbalanced lung nodules. For this purpose, structures of lung lumps were taken out by methods such as relief, genetic algorithm, and particle swarm optimization.

Netto et al. [55] recently published research on lung lesion classification and temporal analysis using a modified version of Quality Threshold Clustering (QTC). The approach involves employing statistical characteristics for feature extraction. Lung sounds' observation rate was utilized to distinguish between disordered respiration and normal breathing patterns. Furthermore, acoustic and segmental features were used to derive lung imaging features from the input data. The classification task was carried out using a hidden Markov model. However, the proposed model incorporates weighted feature extraction, providing it an advantage and enabling the prediction of lung cancer with enhanced accuracy.

Raweh et al. [56] developed grouping of feature selection and extraction procedures to predict cancer. Given the high dimensionality of DNA methylation data, an F-score filtering algorithm was utilized. Specifically, the innovation lies in the extraction model's components, which encompass extraction approaches designed for precise cancer classification. The dependability of the approach was assessed by evaluating various cancer types, such as breast, kidney, colon, lung, and uterine cancers. The evaluation included scenarios with or without the implementation of a hybrid method.

CiOMPI et al. [38] developed model that detect lung cancer using CT images using deep learning. The t-Distributed stochastic neighbour embedding approach is used to represent each nodule using a multidimensional scaling approach. It examined two classifiers for lung nodule detection: an artificial neural network with large training and a CNN. CNN examines a wide range of architectures, each having a unique combination of kernel size, stride, and pad.

Saritas and Yasar [39] utilized Artificial Neural Networks (ANN) in combination with Naive Bayes classifiers to evaluate disease conditions based on data. This approach involved

providing input data and generating corresponding output. Their study elucidated the performance of these two algorithms concerning the accurate and erroneous categorization of samples. The experiments demonstrated that anthropometric data and routine blood analysis parameters are especially valuable for breast cancer diagnosis.

Prabukumar et al. [40] developed a diagnostic approach for early lung cancer detection using Fog Computing. They stored high-volume CT scans in Fog for privacy, low latency, and mobility support. They used Fuzzy C-Means algorithm for precise ROI segmentation and documented nodule characteristics. The CS optimization algorithm was used to classify nodules as malignant or benign.

Sun et al. [57] developed the influence of machine learning techniques in predicting the overall survival of non-small cell lung cancer (NSCLC) patients. To gather information, 339 radio mic attributes were generated since segmented tumor sizes in CT scans taken before treatment. The medical scans' tumor phenotypic characteristics were quantified using parameters like tumor shape, size, intensity statistics, and texture, which collectively define the tumor's phenotypic attributes. The study encompassed the evaluation of five different feature variety approaches and eight distinct ML techniques to explore impact on overall survival prediction. The concordance index, reflecting the agreement between the predicted and actual overall survival for NSCLC patients, was employed to assess prediction performance. While the accuracy, average precision, and precision-recall curve were utilized for evaluation, they weren't applied in the current model to gauge the model's performance on training and testing sets. The study utilized samples from the Kaggle database to assess the model's effectiveness in forecasting future outcomes. The SA-SL_nO proposed model has been identified as more accurate than the alternatives in terms of validation.

Darwish et al. [58] introduced a two-step procedure utilizing four distinct swarm algorithms, including floral pollination, grey wolf optimization, and moth flame optimization. These algorithms were employed in combination with classifiers like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees. Result analysis compared against those of other frequently used algorithms for breast cancer detection. Analysis of result demonstrated the efficacy of the proposed approach in effectively classifying lung cancer data and selecting relevant characteristics.

Firouzabadi et al. [41] developed a reliable method for predicting eight different types of cancer. Several normalization approaches, including PSO, were used to select the most

appropriate normalization technique for each feature for database normalization. For the classification system, statistical feature selection approaches to separate the discriminative features and are upgrade through PSO over proper weighting. Message Passing Neural Networks (MPNN) and SVM were used as classifiers in the end. It concludes that proposed model has considerable promise fast recognize cancer through using SA-SL_nO algorithm.

as et al. [61] presented a methodology for detecting lung cancer in CT images of the lungs. The classification task is performed utilizing an Artificial Neural Network (ANN). Feature extraction was carried out using a residual network. A Faster R-CNN (Region-Based Convolutional Neural Network) architecture, leveraging two network layers, facilitated the classification process. For diagnosing lung cancer, radiograph image enhancement and a radial basis function classifier were utilized. Specifically, a neural network classifier incorporating a radial technology for classification. In proposed methodology, weighted features were employed to efficiently identify lung cancer cells.

Lynch et al. [42] using SEER (Surveillance, Epidemiology, and End Results) data. Various techniques with linear regression, decision trees, Gradient Boosting Machines (GBM), SVM, and tradition group setups were employed. These methods utilized tumor ranking, tumor sex (male or female), age, measurements, and the count of primary tumors as significant data aspects for comparison and prediction evaluation. Rather than categorizing, the forecast was treated as a long-term objective to enhance longevity. The findings indicated that projected values, primarily numerical, aligned well with actual values during low to medium survival intervals. Among the methods used, group through 15.05 Root Mean Square Error (RMSE) demonstrated effectiveness. GBM was identified as the key model within the custom ensemble. Decision trees had insufficient distinct outcomes to draw conclusions. Among the five models, GBM exhibited the highest accuracy with a 15.32 RMSE. For statistical analysis, SVM did not prove to be the optimal choice. Recommendations were made to consider the traditional Cox proportional hazards model as a guiding approach. Ultimately, SEER was shown to be a valuable resource for estimating patient survival time by applying certain machine-learning techniques to lung cancer data. Its performance was comparable to more conventional methods involving specific data collection methodologies.

Table 2. 1 Review of The State of The Art

Sr.No.	Author	Algorithm	Results in Accuracy
1	Shanthi et al. [7]	GLCM and Gabor filter	84%
2	Senthil Kumar et al. [43]	Image Segmentation Algorithms, Guaranteed Convergence Particle Swarm Optimization (GCPSO)	92%
3	Askary et al. [44]	Lung Image Database Consortium (LIDC), KNN, Decision Tree (DT), or SVM	90.67%
4	Prabu Kumar et al. [40]	Fuzzy C-Means, CS optimization, SVM	89%
5	Narayanan et al. [45]	Computed Tomography (CT) and Chest Radiography (CR)	77% in CR and 50% in CT
6	Ankita et al. [42]	Fuzzy C Means Clustering, GLCM, SVM	96.7%
7	Baboo et al. [46]	hybrid of PeSOA, FA, AIS, and BAT	90%
8	Faisal et al. [47]	SVM, Decision Tree, MLP, NN, and NB	90%
9	Firouzabadi et al. [41]	PSO, SVM and Decision Tree	93.04%
10	Yamuna Devi et al. [48]	Search (CS) Optimization Algorithm	91%

Gopi et al. [72] developed a methodology involving noise reduction to eliminate undesired regions from images. The process begins by converting the grayscale image into a normalized version using k-means clustering and an SVM classifier. Feature extraction is achieved using the GLCM approach, and tumor classification is conducted through an SVM classifier. Additionally, the proposed methodology incorporates a deep learning algorithm.

Sarker et al. [73] utilized the k-means algorithm for the segmentation and classification of

lung tumors in 3-dimensional CT images. The initial steps involve lung image segmentation using thresholding and morphological techniques. The segmented area is then identified. Classification is based on tumor nodule metastasis. In the proposed approach, the t-SNE and PCA algorithms are employed for attribute identification, facilitating the identification of lung cancer cells.

Katre et al. [74] investigated lung cancer stages, focusing on noise reduction through the use of a median filter. Image enhancement was performed using a high-boost operator, complemented by image processing and data categorization techniques. Classification tasks were carried out utilizing an SVM classifier, contributing to lung cancer diagnosis through CAD (Computer-Aided Diagnosis) systems.

Huang et al. [76] utilized the maximum information coefficient approach for feature selection. An SVM classifier was employed for classification purposes. Nodule separation was detected through a three-dimensional CNN (Convolutional Neural Network) that utilized a local geometric model-based filter. Subsequently, the isolated nodules were classified using a three-dimensional CNN, with the overall strategy involving PCA (Principal Component Analysis). The effectiveness of lung cancer detection, including the use of weighted features, was demonstrated.

Paing et al. [77] demonstrated the classification of three-dimensional lung nodule border characteristics. Attributes such as diameter, area, major and minor axis length, aspect ratio, perimeter, circularity, and margin features were considered. The classification task was executed through the application of an SVM classifier.

Xiao et al. [78] explored lung cancer detection and appearance aspects, utilizing spherical harmonics and shape properties for lung nodule identification. Classification tasks were achieved through an auto-encoder classifier. A stick filter derivative was constructed using the rotating kernel transformation for fissure segmentation. Lung cancer detection and classification were executed using k-means clustering. Prior to clustering, lung images underwent median and Wiener filtering for artifact reduction. Fuzzy k-means were employed for clustering, with statistics and GLCM used for feature extraction.

Monkam et al. [79] prepared raw lung images for analysis by applying smoothing and filtering. Traditional methods were used for data segmentation, leading to the generation of a new set of nodule candidates. CNN was used for feature extraction and selection. Commonly, machine learning classifiers, such as the Naive Bayes classifier, were utilized

for detection and classification purposes. CAD and CNN architectures were employed, considering data properties and transfer learning. Feature hierarchies were extracted, and classification was performed using a CNN architecture.

Wajid et al. [81] outlined the lung detection process. Shape histogram features based on local energy were extracted. Image enhancement was conducted using the CLACHE technique. Classifiers like SVM and echo state networks were employed for data classification. The detection of major lung illnesses was automated using chest radiography and feed-forward Artificial Neural Networks (ANN). Histogram equalization was utilized for preprocessing. Statistical and geometrical traits were used to extract information, followed by data classification using ANNs.

Leong et al. [82] proposed a Gene Expression Programming (GEP) model for lung cancer prediction based on microarray data. Researchers constructed multiple GEP prediction models using two distinct gene selection strategies. The authors utilized real microarray lung cancer datasets to test and compare their generalized estimating equation (GEE) models. Rigorous tests ensured the reliability of the cross-data validation process, ultimately concluding that the GEP model was the most effective for predicting lung cancer.

Rendon-Gonzalez et al. [83] employed Support Vector Machines (SVM) for the automatic classification and segmentation of lung nodules. Morphological operations and thresholding techniques were initially used to calculate masks before any post-processing. Shape and textural attributes derived from CT lung nodule images were employed to extract information, with classification conducted using SVM classifiers.

Bhuvaneshwari et al. [84] introduced genetic K-Nearest Neighbors (KNN) as a nonparametric approach for early cancer detection, specifically in CT lung scans. Genetic KNN aimed to reduce the manual interpretation time for CT scans. This approach is expected to efficiently and rapidly classify scans. MATLAB-based image processing tools were used to process CT lung scan images. The researchers extensively examined performance indicators such as classification rate and false positive rate. In genetic KNN, K samples were chosen per iteration, with a classification accuracy of 90% considered a suitable fitness level. The approach consistently achieved high accuracy rates.

A specific algorithmic-based classifier, along with distinct enhancement channels, demonstrated automatic lung tumor localization in small region CT images. The researchers incorporated a diverse range of sizes, shapes, and samples in their database. This Computer-

Aided Diagnosis (CAD) approach encompassed five steps: lung segmentation, specific tumor enhancement, tumor starting position determination, feature extraction, and arrangement. The method utilized a unique tumor enhancement channel, achieving a low classification level for tumor cells and a short processing time, rendering it suitable for its intended application.

Arun et al. [85] delved into the realm of perilous tumor growth types that pose a global concern. The spread of diseases through cell proliferation in the lungs can have worldwide implications. Early detection and treatment of these diseases can thwart tumor growth in affected patients. The study focuses on examining various components of information mining systems to forecast the progression of lung tumors in patients. Leveraging information mining concepts enhances the characterization of lung tumors. In a similar manner, the author streamlined the information mining process by drawing parallels with components of an insect group. In this data mining methodologies were employed to ascertain the appropriate reference age and sequence of infection for lung tumor growth stages. The fundamental structure of lung tumor growth, embedded in medical analysis, stands to benefit from these findings. The proposed model uses deep learning techniques to predict lung cancer. PCA and t-SNE are harnessed for feature extraction in lung tumor detection within the proposed approach.

Diciotti et al. [86] used a computer algorithm to categorize images into clusters based on their level of similarity. The Histogram Equalization technique utilizes pre-processed images to extract features, which are then evaluated by neural network classifiers to ascertain whether a patient's initial disease condition is typical or unusual. To predict a patient's survival rate, specific features are examined, and neural network techniques are favored due to their ability to enhance system performance. However, the proposed method excludes images of poor quality as it is unable to forecast tumor cells accurately. Various data mining techniques can enhance the analysis of lung tumor progression. Among them, Naive Bayes, decision trees, and neural networks stand out as promising models for predicting lung disease patients using extensive remote sensing data. In proposed methods for predicting lung tumor progression have the possibility for development.

Bhattacharjee et al. [87], studied decision trees are a widely utilized tool for problem detection. However, due to the inclusion of an extensive range of characteristics, retaining the information can become challenging. In their learning, an SVM was employed to recognize quality components in lung tissue microarrays. The proposed approach overcomes

the limitations of existing models, enabling the identification of lung cancer tumor cells more effectively.

Kim et al. [88] conducted a study implementing Decision Tree methodologies for lung cancer prediction. Utilizing the OSHRI (Occupational Safety and Health Research Institute) dataset spanning from 1992 to 2007, with 153 instances, features like histological type, age, industry type, smoking pack-years, salaried duration, dormancy, and workplace substance exposure were considered and extracted for model training. The CART (Classification and Regression Test) model was applied for lung cancer prediction, with a particular focus on the relationship between lung cancer and employment type.

Petousis et al. [92] developed and estimated a set of Dynamic Bayesian Networks (DBNs) using longitudinal data. The study utilized five DBNs on the LDCT (Low-Dose Computed Tomography) arm of the NLST (National Lung Screening Trial) dataset. The models were constructed and explored using a forward design methodology, leveraging learning structure solutions. The features used in this study included smoking status, cancer history, demographics, exposure to risk factors, and LDCT screening outcome data. The proposed DBN-based model exhibited superior performance compared to more traditional models like naive Bayes and logistic regression. It concludes that the proposed model has considerable promise for first recognition of cancer through using SA-SLnO algorithm.

Table 2. 2 Comparative Analysis of Research Work

Author	Accuracy	Specificity	Sensitivity	Methodologies	Features	Challenges
Pradhan et. al. [2023] [49]	93.15	93.15	93.14	BF-SSA-HR-DEL	A greater comprehension analysis data reduces likelihood of creating a bad decision, leading to more accurate results.	Security of data and longitudinal explosion.
Faruqui et al 2021. [50]	91.1	90	90	Lung Net	High precision and little complexity	Do not use an algorithm for remote learning
Xiao et	92	90	92	WGAN	resolved online	The data

al.2021 [51]					categorization issue with data imbalance.	distribution was not accounted for in forecast model.
Naik et. Al. 2021 [52]	85	80	81	3D-CNN	improvement of categorization system's precision in real-world applications.	early failure in the detection of cancerous lesions.
Wang et. al.2021 [53]	87.3	86	86.2	3D-CNN and RNN	effective recall performance for pulmonary nodule detection.	-Multiple counts of the same nodule provide difficulties. Considers only a single nodule with higher malignancy ratings
Tian et al.2021 [54]	91.6	90	91.4	SAR optimization algorithm	Get under control of difficulties with local optima near the major centroids.	It is necessary to assess using additional performance metrics.
Suresh et. al.2020 [55]	87.15	84.98	89.22	CNN	a high capacity for gathering important relevant information with few stacked structures.	to look at many datasets and deeply complex structures.

2.3 Research Gap

After understanding the literature review, found majorly the three issues in the lung cancer prediction system, which are listed below

1. Inadequate Decision-making Algorithms

2. Incorrect classification of the disease in its early stages
3. Lack of pattern recognition in the various images such as X-rays, CT scans and PET scan

Along with the major issues listed above, found the following shortcomings or research gap

1. Accurate and proper identification of Lung Cancer from 2D images was not provided by many of the existing methods. This is majorly due to the inconsistency in analyzing pulmonary nodules.
2. The literature review highlights the need for Lung Cancer prediction models that integrate image feature extraction for high accuracy
3. The study brings out the requirement of novel deep learning models that integrate hybrid inputs such as Image features, Risk factors, and Symptoms.
4. The existing research work underperforms in differentiating the nature of pulmonary nodules located in CT scans
5. Most of the work done on the development of Lung Cancer detection/prediction models has performed well in reducing the time complexity but while doing this, they failed to determine the severity level and the growth rate of lung cancer.
6. The developed system reduced the computational complexity without degrading the system's performance, but it failed to determine the severity level and the growth rate of lung cancer.
7. The main drawback of the cellular automata method is that it was unable to reduce the rate of error and did not increase the system's dependability.

Chapter Summary

In this chapter, a review of the literature based on studies on the detection of lung cancer using various algorithms has been published. The chosen papers try to find to provide a summary of research based on various models and techniques used for lung cancer detection, datasets used, and performance measures taken into consideration. It has been determined, after a thorough examination of numerous subtopics, that the presented literature survey play a significant role in identifying future research opportunities, research gaps, and the limitations of previous works.

Development of Feature Selection and Classification Model

There are various motivations for employing either supervised or unsupervised learning for data manipulation, with common goals such as data compression and visualization. Both PCA and t-SNE are techniques utilized in machine learning and data analysis to reduce data dimensionality while retaining its fundamental characteristics. These techniques are frequently applied in complex datasets for data visualization and exploratory analysis, aiding in a deeper understanding of patterns and relationships based on the specific objectives of the analysis. t-SNE is great for visualizing complex relationships and clusters within data and PCA effectively reduces dimensionality while preserving as much variation as possible, which makes it useful for noise reduction and feature selection.

Feature extraction, a dimension-reduction technique, involves partitioning vast amounts of raw data into smaller groups. Given the high dimensionality of datasets, significant computational resources are necessary for their processing. Feature extraction involves methods that choose and combine variables into features. These features substantially reduce the data processing load while effectively and comprehensively representing the entirety of the original dataset. It proves beneficial when processing demands fewer resources while retaining essential data and this technique aids analysis by minimizing redundant information.

3.1 Principal Component Analysis

It is a statistical technique that aims to transform high-dimensional data into a lower-dimensional form while retaining most of the original data's variability. The process identifies principal components, linear combinations of initial features, that effectively capture the most significant patterns in the data. Real-world datasets exhibit linear correlations between variables, which are compatible with widely used examination methods. The PCA data mining system receives all extracted features and employs them to enhance the sorting process. PCA is a linear technique, which means that it may not capture complex non-linear relationships in the data [56].

A sample covariance matrix is used to find an eigen function, which is used to compute the coefficients based on the principal components. It holds coefficients for their most important elements in a matrix structure. The modification of the principal constituents stored in all diagonal elements and data with the highest variance yields the most valuable information

for classification.

Scree Plot: A Scree Plot is a graphical representation which has eigenvalues or factors or Principal Component Analysis for a particular analysis. This graph normally starts high on the left side, then quickly falls and towards the end flattens out. By analyzing the nature of this plot, it can be identified that the first component usually accounts for the most variable value and subsequently next few components reach out to be a moderate count. Figure 3.1 shows the Scree Plot of PCA components which converts the feature space PCA1 and PCA2 with their respective numbers. In Figure 3.2, the scree plot shows the significance of the principal count numbers in PCA analysis. The Scree plot help to decide the reliability of PCA functions. The components are organized based on how widely they cover a wide range of possibilities. PCA 1 is the most diverse, followed by PCA 2 and so on.

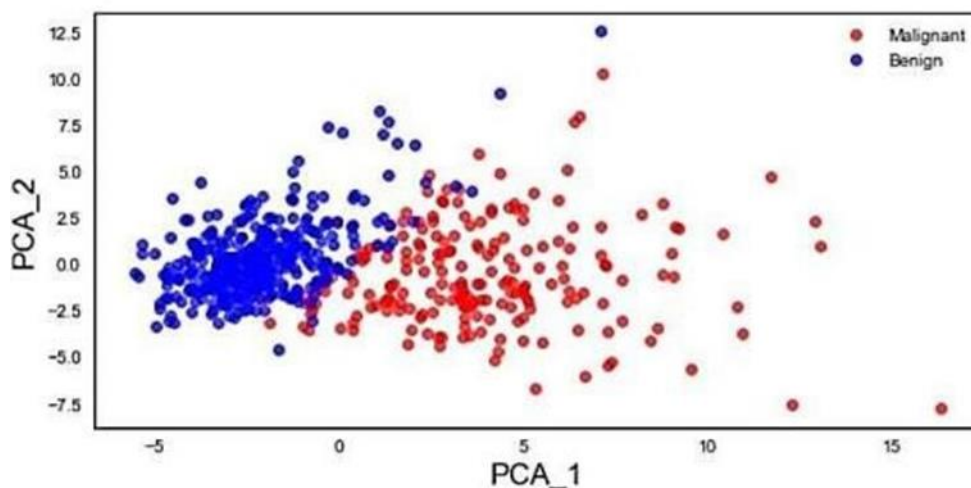


Figure 3. 1 PCA feature space

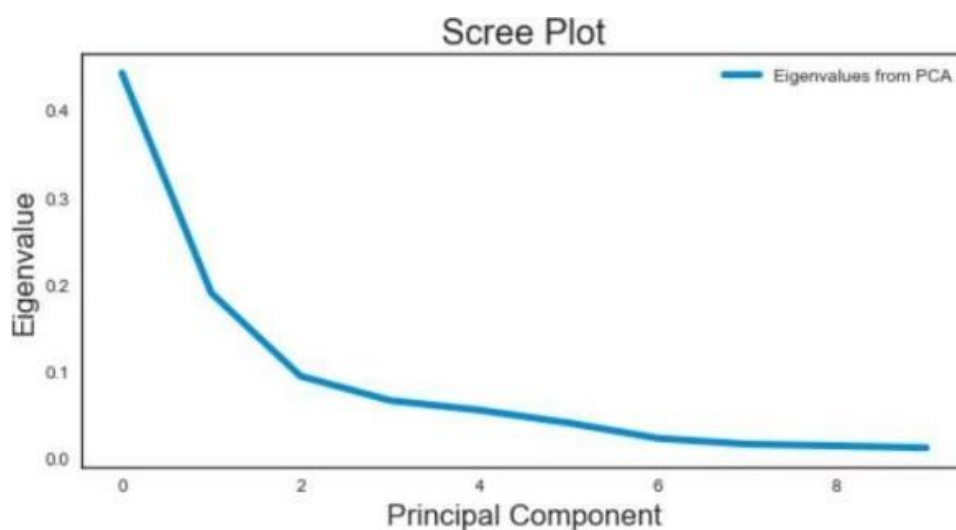


Figure 3. 2 Scree Plot

3.2 t- Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a method employed to achieve nonlinear dimensionality reduction and visualize data patterns. Its objective is to transform high-dimensional data points into a space with fewer dimensions, preserving relationships and similarities between the data points. Unlike PCA, t-SNE focuses on preserving the pairwise similarities of data points rather than the total variance. PCA and LDA (Linear Discriminant Analysis) are both used for visualization and dimensionality reduction, but t-SNE is only used for visualization. It is suitable for visualization high dimensional datasets

t-SNE generates a probability distribution that measures pairwise similarities in the original high-dimensional space and another distribution in the lower-dimensional space. It then tries to minimize the difference between these distributions and effectively group similar data points into a lower dimension. t-SNE is often used to explore clusters or groupings of data and reveal hidden structures that may not be readily visible in a higher-dimensional space. It is particularly useful for visualizing complex and non-linear relationships between data points. Because the scale of t-SNE is four times the number of 'N' objects, that is why its use is limited to datasets with only a ten thousand input object and also learning is too slow to be practical and the memory requirements are too large [57].

3.3 Feature Extraction Model

It is the initial stage in developed lung cancer analysis model. The t-SNE [56] and PCA [55] methods are employed to extract characteristics from the input data. This procedure is crafted to reduce attribute duplication by generating pertinent attributes from pre-existing data entities. Thus, novel group of features can summarize most information included in the original set of features. The techniques such as t-SNE and PCA used for the feature extraction process are explained below.

PCA: The goal of this unsupervised learning technique is to reduce dimensionality of big datasets. It uses mathematical concepts to reduce parameters dataset's vector space which are "various variables are probable correlated into smaller diversity parameters known as essential elements". The PCA's dimensionality reduction is given in Eq. (7).

$$PCA = E'M \quad (7)$$

Here, the data matrix and the eigenvector matrix are represented as M and E' , respectively. The eigenvector matrix is computed using the covariance matrix termed that is given in Eq. (8).

$$E' = E_V \cdot D_i^{-\frac{1}{2}} \quad (8)$$

In Eq. (8), the terms denote matrix of eigenvectors of C_V and a diagonal matrix of eigenvalues of C_V . Assume V as the $G \times H$ matrix with the q^{th} column is $M_z - \mu$, which is represented in Eq. (9).

$$V = [M_1 - \mu, \dots, M_Z - \mu] \quad (9)$$

Here, the mean vector is represented as μ that is computed as $\mu = \left(\frac{1}{Z}\right)(M_1 + \dots + M_Z)$. The covariance matrix with the size $S_z \times S_z$ is given in Eq. (10).

$$C_V = \frac{1}{Z-1} V V^T \quad (10)$$

Finally, the obtained principles are given as $PCA(M, P_c)$, where the number of preserved principal components is given as P_c . Hence, the PCA-extracted features are denoted as in^{PCA}

t-SNE: It is method of reducing dimensions while preserving essential information and internal structures within high-dimensional data. It is used to identify the prevailing error type through dimensionality reduction and the subsequent revelation of error characteristics.

Assume a DS -dimensional dataset $L = \{l_1, l_2, \dots, l_{ND}\}$, which is used for representing the DS -dimensional dataset $l \ll L$. The t-SNE is employed for retaining the relationship of the neighbor and picturing the structure of complex datasets by adopting a Gaussian. To decrease the variance among two deliveries, through high similarity the association across input. To provide a clear comparison, t-SNE is utilized to transform the complex Euclidean distances between objects l_j into uncertain possibilities $\rho_{j|i}$. The joint probabilities $\rho_{j|i}$ are described to compute “the pairwise similarity between l_i and l_j by summarizing as given in Eq. (11) and Eq. (12).

$$\rho_{j|i} = \frac{\exp\left(-\frac{\|l_i - l_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|l_i - l_k\|^2}{2\sigma_i^2}\right)} \rho_{i|i} = 0 \quad (11)$$

$$\rho_{ij} = \frac{\rho_{j|i} + \rho_{i|j}}{2ND} \quad (12)$$

Here, the term σ_i indicates the inconsistency bound of the Gaussian function that is relatively related to the number of adjacent neighbors. There is a conditional distribution ρ_i for each σ_i . The pairwise similarity \wp_{ij} is measured among o_i and o_j in low-dimensional space as represented in Eq. (13).

$$\wp_{ij} = \frac{\left(1 + \|o_i - o_j\|^2\right)^{-1}}{\sum_{k \neq f} \left(1 + \|o_k - o_f\|^2\right)^{-1}} \wp_{ji} = 0 \quad (13)$$

The Kullback-Leibler difference among the joint distributions ρ and a normal measure of the authenticity \wp at the modelling process is validated. The solution process is improved by adding what is given in Eq. (14).

$$\delta^{(it)} = \delta^{(it-1)} + \eta \frac{\partial Cb}{\partial \delta} + \beta(it) (\delta^{(it-1)} - \delta^{(it-2)}) \quad (14)$$

Here, $\beta(it)$ denotes the momentum at iteration it , the learning rate is fixed and the cost function is mentioned as Cb that is given in Eq. (15).

$$Cb(O) = KL(\rho \parallel \wp) = \sum_i \sum_j \rho_{ij} \log \frac{\rho_{ij}}{\wp_{ij}} \quad (15)$$

The gradient $Cb(O)$ is obtained by employing the gradient decent approach to the cost function, where the optimal matrix O is given in Eq. (16).

$$\frac{\partial Cb}{\partial o_i} = 4 \sum_{j \neq i} (\rho_{ij} - \wp_{ij}) \wp_{ij} Nz(o_i - o_j) \quad (16)$$

Here, the normalization term is formulated in Eq. (17).

$$Nz = \sum_{k \neq f} \left(1 + \|o_k - o_f\|^2\right)^{-1} \quad (17)$$

Thus, the t-SNE-extracted characteristics are represented by in^{PCA} . The joint features of PCA and t-SNE obtained throughout the feature extraction process are represented $in_i = \{in^{PCA}, in^{t-SNE}\}$, by a newly developed attribute correlation-based optimized weighted feature extraction procedure.

3.3.1 Innovative Weighted Feature Extraction Model Based on Attribute Correlation

This cancer detection application of innovative attribute correlation-based optimized weighted feature extraction is its main contribution. The dataset is divided into two sections based on class 1 (normal class) and class 2 (abnormal class) for this novel feature extraction model. Let the matrix of the attribute i_y agreeing to class 1 be denoted as $MT_{i_y}^1$ and the matrix attribute i_y agreeing to class 2 be denoted as $MT_{i_y}^2$, where $i_y = 1, 2, \dots, NY$ and the number of attributes NY accessible in the dataset. The association between the characteristics of class 1 and class 2 should be determined. Eq. (18) displays the computational method for the connection between class 1 and class 2 corresponding to i_y^{th} the attribute. Here, $Cov(MT_{i_y}^1, MT_{i_y}^2)$ indicates computed based on Eq. (19), and $\sigma_{MT_{i_y}^1}$ and $\sigma_{MT_{i_y}^2}$ represent the standard deviation of matrix $MT_{i_y}^1$ and $MT_{i_y}^2$, separately.

$$CR_{i_y} = [MT_{i_y}^1 \text{I} MT_{i_y}^2] = \frac{Cov(MT_{i_y}^1, MT_{i_y}^2)}{\sigma_{MT_{i_y}^1} \sigma_{MT_{i_y}^2}} \quad (18)$$

$$Cov(MT_{i_y}^1, MT_{i_y}^2) = E\left[(MT_{i_y}^1 - \mu_{MT_{i_y}^1})(MT_{i_y}^2 - \mu_{MT_{i_y}^2})\right] \quad (19)$$

The characteristic connection is used to limit the feature vector. Eq. (20) affects how the feature is formulated. By defining the weight to be multiplied for each feature based on connection value, the suggested correlation-based optimized weighted feature extraction is accepted in an improved method. To analyze the unvarying variance over big datasets, this method is used. The formula of weighted feature extraction is decided using Eq. (19).

$$in_i^{new} = \begin{cases} (1 - Wg_{i_y}) * CR_{i_y} * in_i & \text{if } (CR_{i_y} > 0.5) \\ Wg_{i_y} * CR_{i_y} * in_i & \text{else} \end{cases} \quad (20)$$

Here, the terms Wg_{i_y} in_i represent the nominated weight of the attribute and extracted combined feature vector, respectively. The SA-SLnO algorithm that has been used to optimize the weight of attributes. As a result, the format of Eq. (21) used to build the new feature vector.

$$in_i^{new} = \{in_1^{new}, in_2^{new}, \dots, in_{NF}^{new}\} \quad (21)$$

Here, NF is the whole quantity of extracted features.

3.4 Dataset

Datasets 1 and 2 are obtained from open source Kaggle. These datasets are of the integer type. Dataset 1 comprises a total of 55,395 patient records and has a size of 1.88 MB. A healthy group of 27915 people aged 18 to 87 years and a cancer patients group of 27480 people with lung cancer were the two sample of this study. Patient information is included in the dataset, including details on smoking, yellow fingers, anxiety, weariness, allergies, wheezing, alcohol use, coughing, chest pain, etc. The collected patient data consisted of 27686 males and 27808 females.

Dataset 2 consists of a total of 401949 patients data between the ages of 14 and 73 years and has a size of 51 MB. It collects data on 51,237 healthy and 350,712 unhealthy people. The database includes factors such as alcohol use, smoking, obesity, chest pain, fatigue etc. including patient information. Table 3.1 provides some of the dataset's characteristics, with 1 indicating "YES" and 2 indicating "NO". The attribute information of the datasets are as follows:

Table 3. 1 Attributes of Dataset

Attributes	Values ('YES' as 2 and 'No' as 1)
Gender	Male/Female
Age	14 – 87
Smoking	Yes / No
Yellow Fingers	Yes / No
Fatigue	Yes / No
Allergy	Yes / No
Anxiety	Yes / No
Wheezing	Yes / No
Alcohol	Yes / No
Coughing	Yes / No
Chest Pain	Yes / No

Chapter Summary

In this chapter covered feature selection methods such as PCA and t-SNE, PCA is a technique used to reduce the dimensionality of data while preserving its invariance, and t-SNE helps visualize complex data patterns. The chapter introduces novel approaches that Weighted Feature Extraction Model based on Attribute Correlation. The description of relevant mathematical equations within the context of feature selection and the weighted feature extraction model and also introduces the datasets 1 and 2 of lung cancer patients used for analysis, including several characteristics related to the patient's health.

This chapter emphasizes the importance of thoroughly exploring and analyzing two innovative methodologies: the SA-SL_nO-RNN algorithm for predicting and diagnosing lung cancer, and the optimization of the RNN model through configuring hidden layers. These methodologies have the potential to revolutionize the field of lung cancer prediction by significantly enhancing accuracy and efficiency in diagnostic procedures. The SA-SL_nO-RNN algorithm represents a novel approach to predicting lung cancer.

4.1 Proposed SA-SL_nO-RNN Deep Learning Model for Prediction of Lung Cancer

4.1.1 System Architecture

Lung cancer requires early detection and prediction due to its destructive nature and potential future impact on the human body. Extensive utilization of deep learning and advanced artificial intelligence technology has assisted medical practitioners in early predicting such perilous illnesses. These sophisticated techniques can detect and unveil hidden patterns in lung cancer tumor sizes that may not be identifiable by medical surgeons or equipment.

Traditional methods of machine learning like Naive Bayes, Support Vector Machine, and Random Forest are typically used for the prediction of Lung Cancer in recent studies. However, there are certain limitations and disadvantages of these methods like less accurate prediction and time complexity which need to be improved for producing a more dependable and consistent result with higher accuracy. Hence in this research, proposing an advanced SA-SL_nO-RNN method with a weighted average feature selection and optimization technique. Figure 4.1 represents the system architecture diagram.

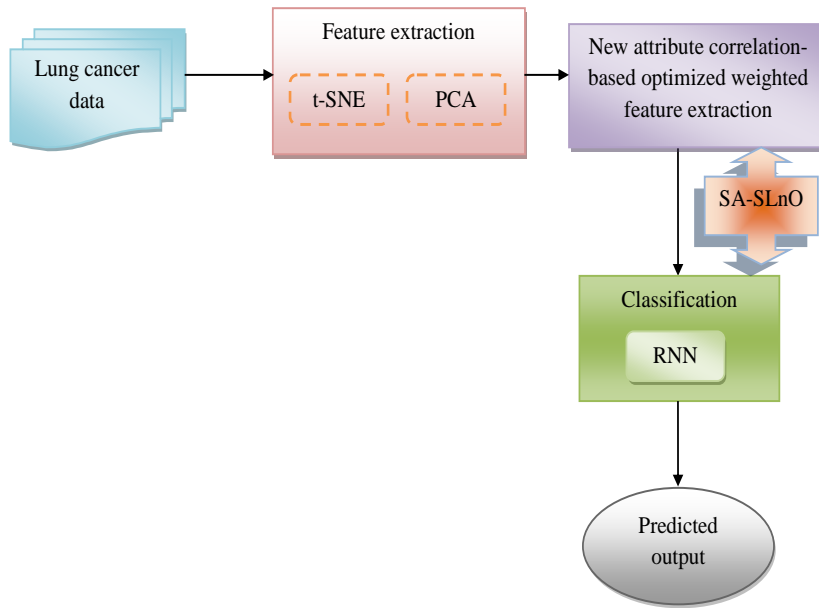


Figure 4. 1 Proposed Lung Cancer Detection Architecture

The projected lung cancer diagnosis model takes its input from the patient's medical record. "a) Data Gathering, (b) Feature extraction, (c) New attribute correlation-based optimal weighted feature extraction, and (d) Prediction" are some of the different phases of this model's development. In the beginning, the data collection is carried out to acquire various benchmark datasets from Kaggle datasets, which include attribute information from different patients in the form of health records. The t-SNE and PCA techniques are used to extract features from the input data during the feature extraction step. The feature extraction step in the diagnosis model is crucial for reducing data dimensionality by eliminating redundant information. This feature enhancement contributes to improved inference and training speed. The new optimized correlation-based weighted feature extraction stage receives the derived features and uses the suggested SA-SLnO algorithm to optimize weights. Weighted feature extraction is developed based on the correlation between characteristics of different classes. This is done to ensure uniform scaling of all the data, thereby reducing complexity. SA-SLnO optimizes the hidden neurons of the RNN. The standard SLnO method is transformed into the SA-SLnO algorithm by creating a new fitness-based solution for the random vector. In conclusion, the proposed lung cancer diagnosis model efficiently detects output from optimized RNN classifier. The main goal of this proposed lung cancer diagnosis model is to minimize the MSE among the actual and predicted output.

4.1.2 Objective Model for Proposed Lung Cancer Diagnosis

In this section calculating fitness values of patients using fitness function. The concept of a fitness function, which plays a pivotal role in quantifying the fitness levels of patients in the context of lung cancer prediction. This function evaluates the relevance and contribution of various attributes, allowing the model to assign appropriate weights to different features. As a result, fitness function, as stated in Eq 1., is used reduced of MSE among actual and anticipated outcomes in model for lung cancer diagnosis.

$$FF = \arg \min_{\{Wg_i, HN_{RNN}\}} (MSE) \quad (1)$$

In Eq. (1), FF as a fitness function, word HN_{RNN} denotes the secret neurons of RNN and Wg_i expressions of weights of attributes that need to be optimized, wherever $i = 1, 2, \dots, NF$ and total list of properties in the dataset is denoted as NF . The MSE is computed by Eq. (2).

$$MSE = \frac{1}{tn} \sum_{w=1}^{tn} (AO_w - PO_w)^2 \quad (2)$$

Here, term tn denotes the number of data points, actual and predicted output is indicated as AO_w and PO_w , respectively.

4.1.3 Optimized RNN

The data are classified into normal and abnormal using an optimized RNN in proposed lung cancer diagnostic model. The classification method utilizes the correlation-based weighted feature extraction stage's extracted features as an input. The sum of veiled neurons in the RNN is optimized for SA-SLnO algorithm to improve classification accuracy. The feature of recurrent neural network (RNN) is used to characterize the relationship between the current and previous output of the sequence in the order of the sequence growth model. RNN's performance is constrained by the time and effort required to train it. The optimized RNN is illustrated in Figure 4.2.

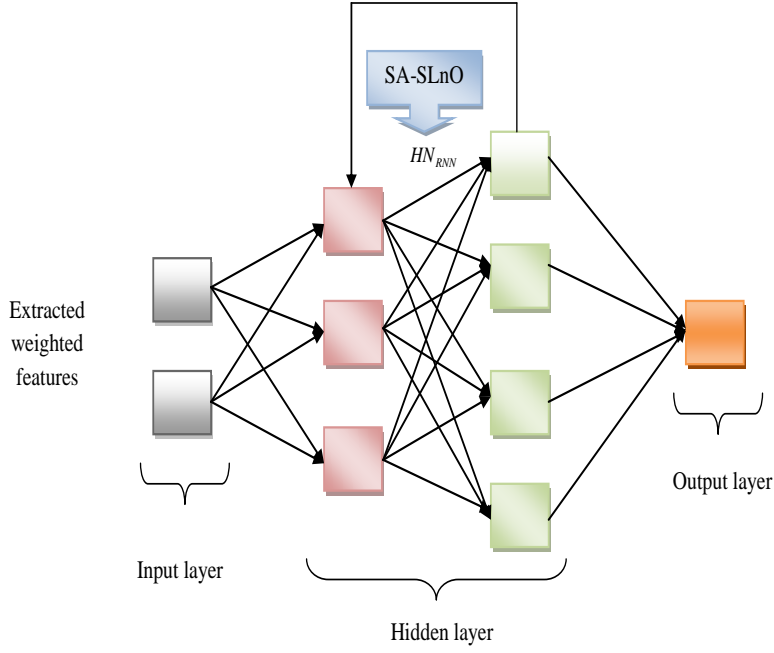


Figure 4. 2 The proposed optimized RNN

The input is considered as (in_i, h_{i-1}) , where the input features are considered for i^{th} picture cut and the earlier secret state. In Eq. (3) and Eq. (4), the apprise gate and reset gate are calculated.

$$u_i = \delta(Wt^{inu} in_i + Wt^{hu} h_{i-1}) \quad (3)$$

$$rg_i = \delta(Wt^{inrg} in_i + Wt^{hrg} h_{i-1}) \quad (4)$$

Here, a logistic sigmoid function is represented δ and the weight matrices are termed as Wt^{inu} , Wt^{hu} , Wt^{inrg} and Wt^{hrg} . In Eq. (5),

$$\tilde{h}_i = \tan(Wt^{inh} in_i + Wt^{hh}(h_{i-1} \otimes rg_i)) \quad (5)$$

Here, an element-wise multiplication is denoted as \otimes , the previous state h_{i-1} and the candidate state \tilde{h}_i is computed in i^{th} hidden activation state h_i of GRU, given in Eq. (6).

$$\begin{aligned} h_i &= (1 - u_i) \otimes \tilde{h}_i + u_i \otimes h_{i-1} \\ &= (1 - u_i) \otimes \tilde{h}_i + u_i \otimes h_{i-1} \end{aligned} \quad (6)$$

As an amendment to this conventional RNN, manipulative hide neurons using algorithm for improving accuracy of diagnosis through reducing the error between measured and actual value.

4.1.4 Proposed SA-SLnO Algorithm and Model

The objective of the proposed SA-SLnO algorithm is to optimize weight parameters during the new attribute correlation-based weighted feature extraction phase and to optimize hidden neurons of RNN during the classification stage. This enhances the development of lung cancer diagnosis performance. The self-adaptive approach has been created to address a differnt of optimization concerns in the experimental results that have show significant advantages in resolving optimization issues and it is also used in significant feature selection problems. The SA-SLnO method has been suggested to improve the performance of the optimized weighted feature extraction procedure and diagnosis stage. Self-Adaptive algorithms are created by either modifying adaptive operators or adaptive functions. There are several approaches to implement adaptation in conventional optimization concepts. Performance has improved as a result of choosing an appropriate adaptive model in the current optimization procedures. The concept is derived from the hunting techniques of sea lions, characterized by their superior hunting abilities, lucid vision, and quick movement. The algorithm has four crucial stages:

Tracking of prey: The tracking behavior of SLnO is formulated in Eq. (7), for best solution

$$\vec{D} = \left| 2\vec{R} \cdot \vec{S}(it) - \vec{T}p(it) \right| \quad (7)$$

Here, the term it represents current iteration, the distance \vec{D} is estimated between the target prey and sea lion, and terms $\vec{S}(it)$ $\vec{T}p(it)$ target prey respectively. The random vector, \vec{R} in boundary of $[0, 1]$ for discovering optimal or close to the optimal solution. In Eq. (8), the behavior of sea lion at next iteration is modelled.

$$\vec{S}(it + 1) = \vec{R}(it) - \vec{D} \cdot \vec{Q} \quad (8)$$

In Eq. (8), the term $(it + 1)$ denotes the next iteration and the random value \vec{Q} is linearly reduced from 2 to 0.

Vocalization phase: Sea lions communicate through a variety of vocalizations, particularly during hunting and chasing behaviors. Sea lions exhibit the capability to detect sounds both above and underwater, with sound frequencies below 0.25 Hz being optimal for hunting, thus leading to the formulation of the remaining equation in Eq. (9) following prey identification.

$$\bar{L}S_{leader} = \left| \frac{\bar{W}_1(1 + \bar{W}_2)}{\bar{W}_2} \right| \quad (9)$$

$$\bar{W}_1 = \sin \theta \quad (10)$$

$$\bar{W}_2 = \sin \phi \quad (11)$$

Here, the speed of sounds in water and air is termed as \bar{W}_1 and \bar{W}_2 , respectively and the speed of sound of the sea lion leader is represented as $\bar{L}S_{leader}$.

Exploitation or attacking phase: Initially, the prey is observed by the sea lion's master who advises another lion about it. The prey is encircled by all the sea lions. Find the best solution for target hunt. The hunting behavior is mathematically formulated by proposing two techniques circle update position and dwindling position. In dwindling method works created on random value \bar{Q} , which is decreased from 2 to 0 which enables the sea lions to move near prey and enclose it. Consequently, the next location of a search agent can be placed somewhere among agent leading position and the current best agent position. The sea lions track fish and start hunting from the edges in the circle updating position as formulated in Eq. (12).

$$\vec{S}(it + 1) = \left| \vec{R}(it) - \vec{S}(it) \right| \cdot \cos(2\pi a) + \vec{R}(it) \quad (12)$$

Here, a random number is given as a in a range of $[-1, 1]$, absolute value is termed as $|\cdot|$ and distance among the target prey, $|\vec{R}(it) - \vec{S}(it)|$ denotes the search agent. The hunting procedure is performed in a circular motion using $\cos(2\pi a)$ Eq. (12).

Exploration phase or search for prey: Generally, the search process is performed randomly. Therefore, the random value \bar{R} is utilized. If $\bar{R} \in [-1, 1]$, then sea lions are forced to move from sea lion's leader and target prey. Here, sea lion updated best position, Global search agent is responsible for locating the optimal solution at a global level when $\bar{R} > 1$ it is modelled in Eq. (13).

$$\bar{D} = \left| 2\bar{R} \cdot \vec{S}_{rand}(it) - \vec{S}(it) \right| \quad (13)$$

$$\vec{S}(it + 1) = \vec{S}_{rand}(it) - \bar{D} \cdot \bar{Q} \quad (14)$$

Here, random sea lion chosen as of recent population is denoted $\vec{S}_{rand}(i)$. The performance of SLnO for solving the diverse test functions is improved than the conventional

optimization algorithms. SLnO demonstrates superior searching behavior, effectively avoiding locally optimal solutions and accelerating the speed of convergence. When compared to other meta-heuristic algorithms, the SLnO method performed better and demonstrated competitive performance. Based on the effective features of the SLnO, it can be implemented with self-adaptive behavior. A new algorithm termed SA-SLnO is proposed using the advantages of self-adaptive nature and the efficiency of adapting these functions or operators has proven its importance in literature. It is specifically suggested for improving model execution. The SA-SLnO proposed through defining new formula for random value \bar{R} in Eq. (15), which is purely depended on the fitness function of previous and current iterations. The proposed algorithm works based on the improvement by using the adaptively updated random value \bar{R} that depends on the fitness function. Therefore, this algorithm is called a self-adaptive algorithm.

$$\vec{R} = Tau \quad (15)$$

$$Tau = \frac{f(it-1) - f(it)}{f(it-1)} \quad (16)$$

Now, i as iteration, and fitness function solution iteration of previous and current state specified as $f(it-1)$ and $f(it)$, respectively. In Algorithm 1, pseudo-code of proposed SA-SLnO algorithm is given.

Algorithm 1: Proposed SA-SLnO [27]
Initialize population
Select \vec{S}_{rand}
Compute fitness for each search agent
if ($i < \max \text{ iterations}$)
Determine $\vec{L}S_{leader}$ using Eq. (9)
if $\vec{L}S_{leader} < 0.25$
if $Q < 1$
Find \bar{R} by self-adaptive concept using Eq. (15)

Alter the state of the current search agent with the use of Eq. (8).

Else

Choose \vec{s}_{rand}

And use to it update location of recent search agents with the use of Eq. (14).

end if

Else

Again, use to it update the location of recent search agents with the use of Eq. (12).

end if

In case the search agent goes beyond $\vec{L}S_{leader}$

Moves to the first condition of if

Else

Calculate fitness for every search agent

Amend \vec{s} based on a greater solution

Return \vec{s} based on the bad solution

End if

end if

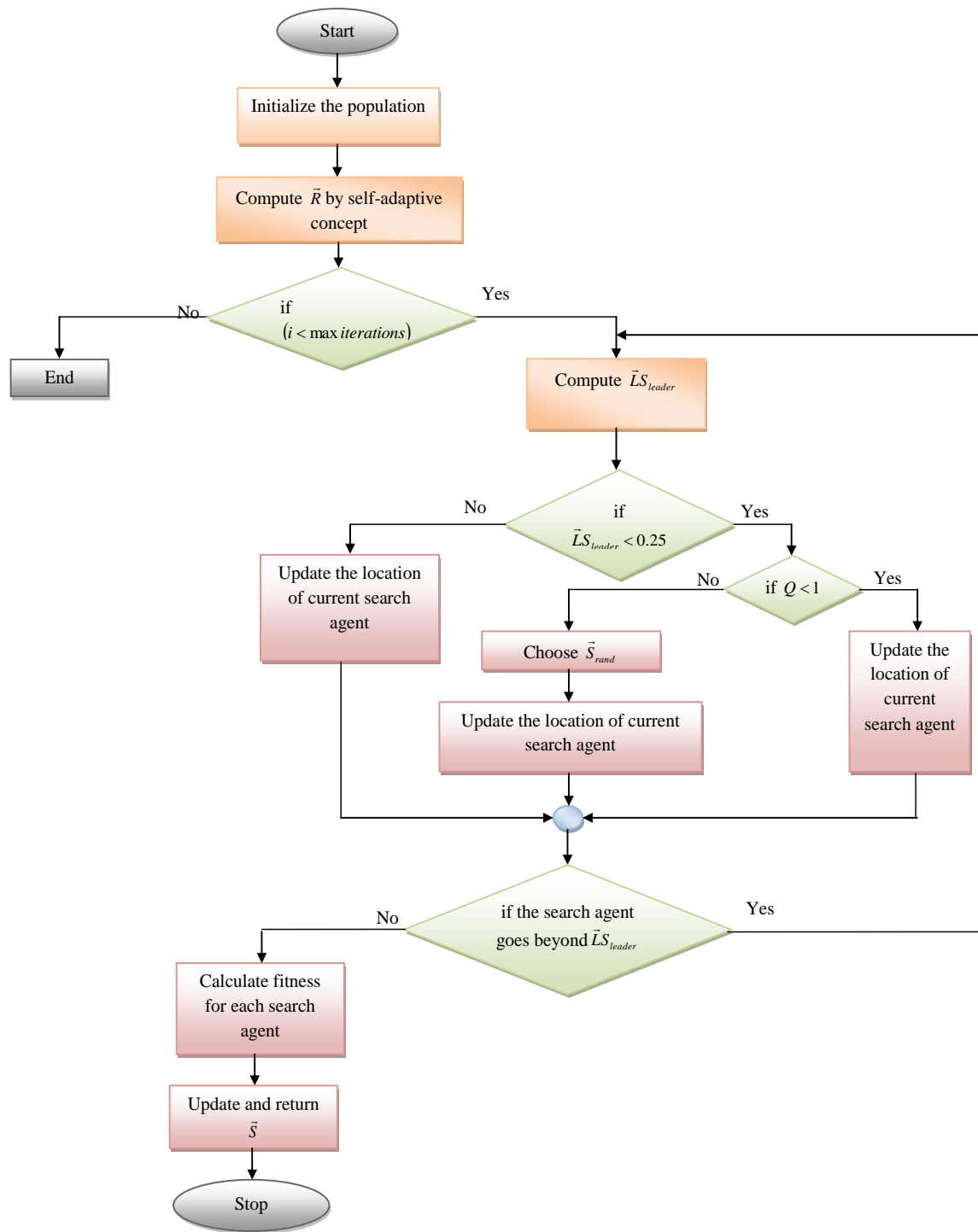


Figure 4. 3 Flowchart of the proposed SA-SLnO algorithm

Chapter Summary

This chapter summarizes the utilization of new deep-learning techniques for lung cancer estimate. It introduces enhanced weighted feature extraction methods rooted in critical attributes for the proposed deep learning model (SA-SLnO-RNN). The primary contribution lies in the advancement of weighted feature extraction techniques, which are guided by feature correlations to refine lung cancer prediction. The propose model for lung cancer estimate, on Recurrent Neural Network, amplifies the significance of feature weights to optimize the performance of RNN's feature correlations and the activation of neurons within its layers. The chapter also includes a detailed flowchart depicting the structure and functioning of proposed SA-SLnO-RNN model.

Verification and Validation of Proposed Model

This chapter discusses comparative analysis of the SA-SL_nO-RNN methodology with the existing models. The dataset of proposed model taken from Kaggle website and consists of nine features as described in Section 3.4. The execution of proposed model compared with conventional optimization models like PSO (Partical Swarm Optimization), GWO (Gray Wolf Optimization), GSO (Glowworm Swarm Optimization) and SL_nO (Sea Lion Optimization) and conventional model like DT, SVM, NN (Neural Network), DNN (Deep Neural Network) and RNN. For prediction of lung cancers, proposed model is compared with other models on basis of various performance parameters like accuracy, F1-score, precision and recall.

5.1 Work Environment

For developing lung cancer prediction model's experimentation MATLAB tool is used as it is very powerful for mathematical calculation also provides matrix and array language orientation, which makes it simple. It employs statistical analysis and deep learning modeling approaches which helps to model learning and carry out classification tasks without delay from text dataset. The planned lung cancer diagnosis is developed in MATLAB 2019a. The experimentation is executed on a machine running Windows 11 with 8GB RAM and an Intel i5 processor.

MATLAB is a language used for numerical computation, data analysis, and algorithm development. MATLAB is also known as Matrix Laboratory. It is mostly used in engineering, physics, mathematics, and finance field. It has intuitive syntax and extensive fitted functions. Deep Learning with MATLAB is used as follows:

1. MATLAB is designed for scientists and engineers to analyze and enhance systems and products quickly and efficaciously. MATLAB Language is a matrix language that expresses the maximum number of common mathematical expressions. MATLAB is identified as one of the best programming languages that can handle deep learning in an easy and straightforward way.
2. Labeling is one of the most challenging tasks in deep learning, and MATLAB is one of the best practices to implement it. MATLAB gives interactive deep learning apps, including signal data, audio data, images, and video.

3. MATLAB can help generate artificial data if there aren't enough records to guarantee accuracy. This is a huge advantage because deep learning relies on huge amounts of data.
4. MATLAB has many approaches to interaction and transforming the data of the deep learning frameworks.
5. MATLAB additionally supports Python interaction.

5.2 Performance Measures

The performance metrics used for estimation of proposed model are described below. (a) Accuracy: “proportion opinion of closely projected to whole remarks”.

$$Arc = \frac{(tr^{po} + tr^{ng})}{(tr^{po} + tr^{ng} + fa^{po} + fa^{ng})} \quad (32)$$

(b) Sensitivity: “numeration the true positive values”.

$$Sty = \frac{tr^{po}}{tr^{po} + fa^{ng}} \quad (33)$$

(c) Specificity: “counting the true negatives, which are resolute indeed”.

$$Spy = \frac{tr^{ng}}{fa^{ng}} \quad (34)$$

(d) Precision: “average of positive observations that are exactly calculated to total number of correct systems shaping the conditions”

$$Pcs = \frac{tr^{po}}{tr^{po} + tr^{po}} \quad (35)$$

(e) FPR: “the ratio of the count of high false prediction to the entire count of negative prophecies”

$$FPR = \frac{fa^{po}}{fa^{po} + tr^{ng}} \quad (36)$$

(f) FNR: “number of positives which vintage negative test products with test”.

$$FNR = \frac{fa^{ng}}{tr^{ng} + tr^{po}} \quad (37)$$

(g) NPV: “possibility that subjects with a negative showing test truly don't have the disease”.

$$NPV = \frac{fa^{ng}}{fa^{ng} + tr^{ng}} \quad (38)$$

(h) FDR: “number of false positives in all excluded hypotheses”.

$$FDR = \frac{fa^{po}}{fa^{po} + tr^{po}} \quad (39)$$

(i) F1 score: “sympathetic mean between precision and recall. Also used for geometric amount to rate performance”.

$$F1-score = \frac{Sty \cdot Pcs}{Pcs + Sty} \quad (40)$$

(j) MCC: “correlation coefficient computed by four values”.

$$MCC = \frac{tr^{po} \times tr^{ng} - fa^{po} \times fa^{ng}}{\sqrt{(tr^{po} + fa^{po})(tr^{po} + fa^{ng})(tr^{ng} + fa^{po})(tr^{po} + fa^{ng})}} \quad (41)$$

Here, the terms tr^{po} , tr^{ng} , fa^{po} , fa^{ng} referred to as “the true positives, true negatives, false positives, and false negatives,” respectively.

5.3 Performance Analysis on Optimization Algorithms for Dataset 1 using SA-SL_nO Algorithm

The expectation of lung cancer analysis model is evaluated through multiple conventional optimization algorithms for dataset 1, which is given in Figures 5.1,5.2,5.3,5.4, and 5.5. The accuracy of established SA-SL_nO-based RNN classifier improves than other algorithms in terms of learning percentages.

- It demonstrates 1.2% improvement compared to PSO-RNN and GWO-RNN when study ratio 85.
- The SA-SL_nO-RNN shows a 12% precision improvement over PSO-RNN, a 1.6% improvement over GWO-RNN, a 16% improvement over GSO-RNN, and a 23% improvement over SL_nO-RNN, all while considering a learning percentage of 45.
- The SA-SL_nO-RNN achieves a 50% lower FPR compared to PSO-RNN, a 12.5% lower FPR compared to GWO-RNN, a 66% lower FPR compared to GSO-RNN, and a 74% lower FPR compared to SL_nO-RNN, all while in view of study ratio of 45.
- FNR of SA-SL_nO-RNN is significantly reduced, showing a 97% decrease compared to PSO-RNN, a 91% decrease compared to GWO-RNN, a 94% decrease compared to GSO-RNN, and a 94.4% decrease compared to SL_nO-RNN, all at a learning percentage of 45.
- The SA-SL_nO-RNN achieved a 35% higher F1-score compared to PSO-RNN, an 8% lead over GWO-RNN, a 17% advantage over GSO-RNN, and a 25% improvement over SL_nO-RNN, all while sustaining a 45% learning percentage.

Consequently, the expand optimized RNN model demonstrates superior performance across a range of measures, particularly at the initial percentage for dataset 1.

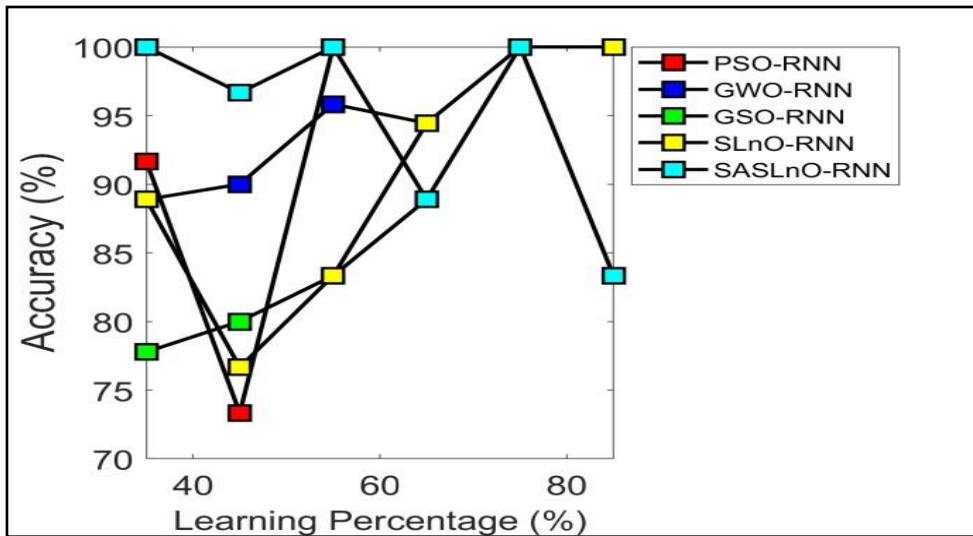


Figure 5.1 Comparative Analysis of Accuracy

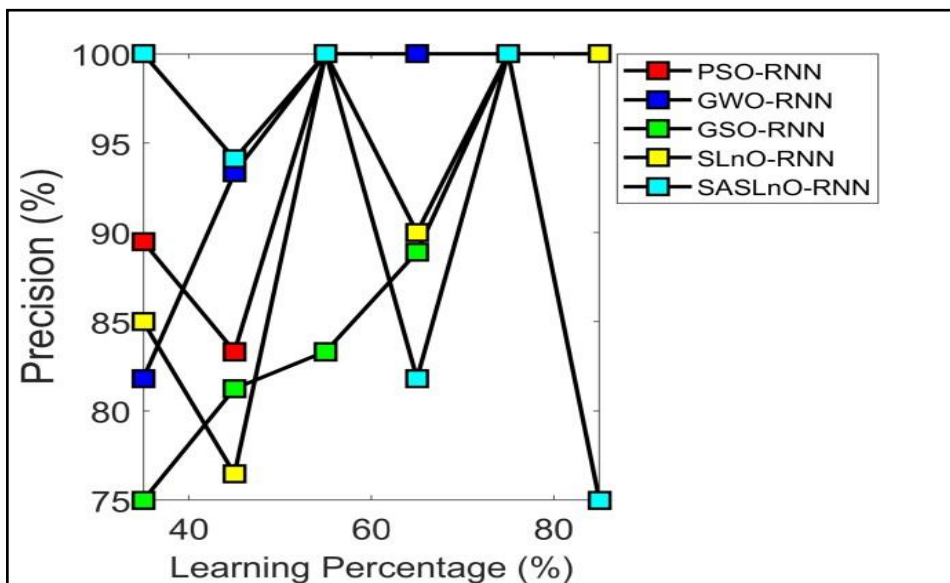


Figure 5.2 Comparative Analysis of Precision

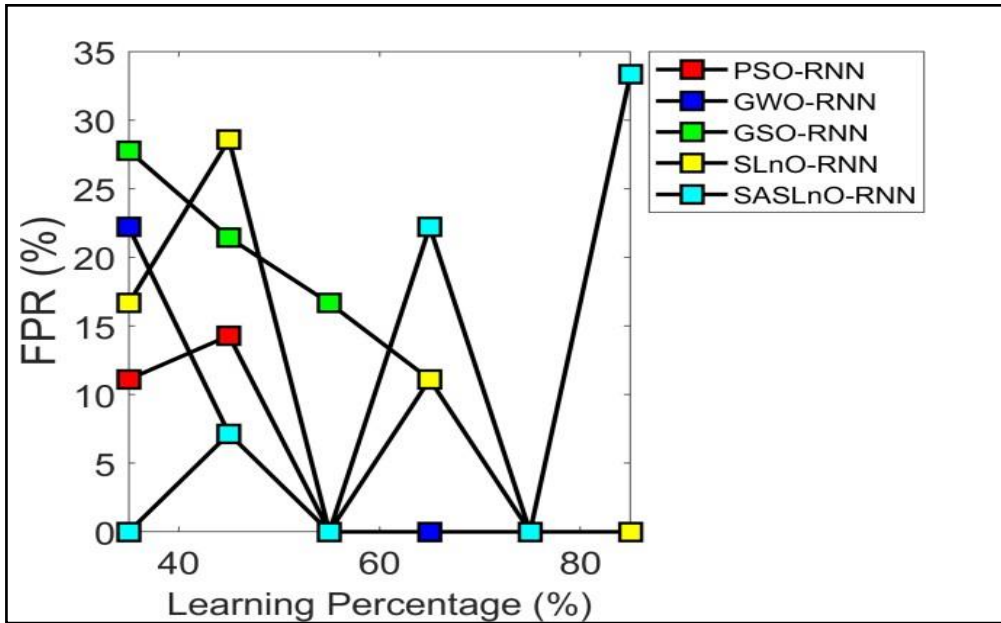


Figure 5.3 Comparative Analysis of FPR

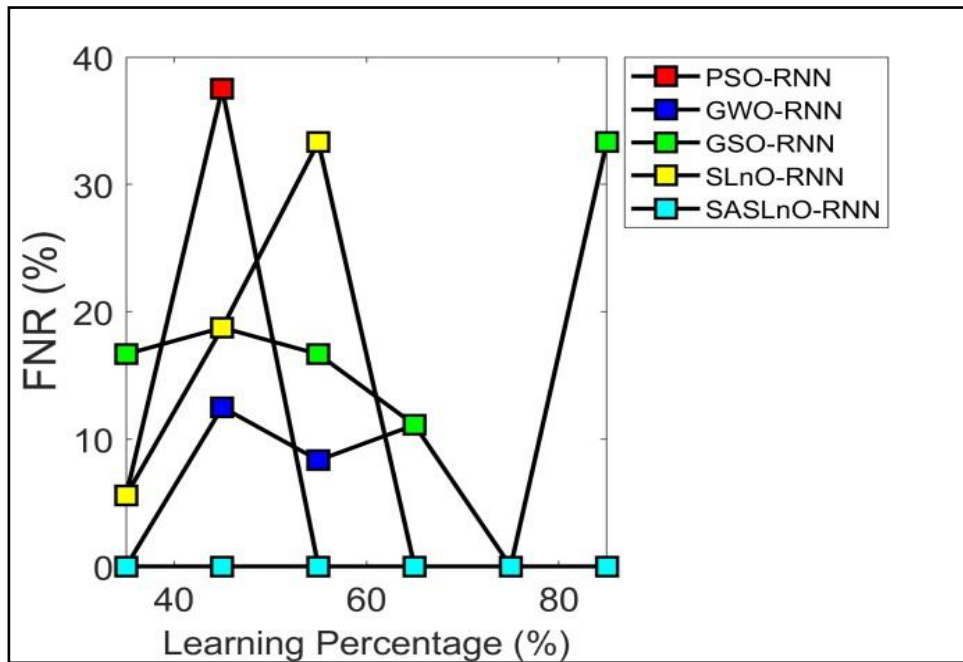


Figure 5.4 Comparative Analysis of FNR

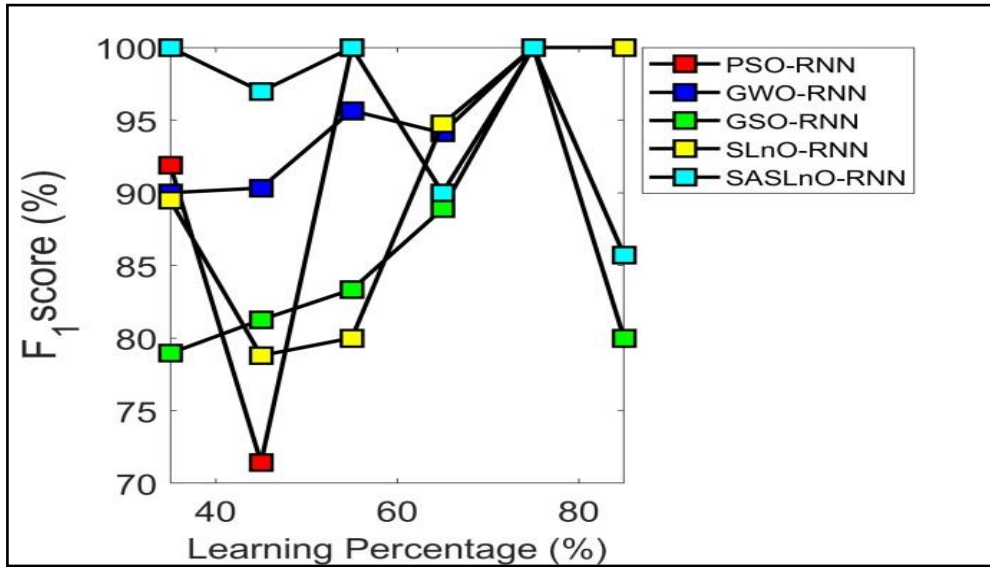


Figure 5. 5 F1-score Comparative Analysis

5.4 Comparative analysis of Model Performance

The performance evaluation of the SA-SLnO-based optimized RNN model is illustrated in figures 5.6 to 5.10, in comparison with alternative machine learning models.

- Figure 5.6 presents a comparison of accuracy among all models, including the proposed SA-SLnO RNN model. Especially, the proposed SA-SLnO-RNN model shows a 3% higher accuracy than SVM, an 8% accuracy gain over traditional RNN, and a 16% accuracy improvement over the DT algorithm.
- Figure 5.7 provides a comparative analysis of precision between our proposed methodology and other machine learning models. The proposed model shows a 6.7% higher precision compared to SVM and an 8.5% precision advantage over both RNN and DNN.
- Figure 5.8 provide the FPR of the predictable SA-SLnO-RNN, at a learning percentage of 45, demonstrating a 57% reduction when compared to SVM, NN, and DNN. It also boasts a 14% reduction over RNN. The FNR of the SA-SLnO-RNN is impressively lowered by 94% compared to DNN and 92% compared to RNN.
- Lastly, from Figure 5.10, it becomes apparent that F1-score of proposed SA-SLnO-RNN model, at a learning percentage of 45, is superior to other models. Specifically, it showcases a 2.6% improvement over SVM, a 16% improvement over NN, a 14% improvement over DNN, and an 8.3% improvement over RNN.

Hence, it can be concluded that in every parameter analysis, our proposed model

outperforms other machine learning models used in recent studies for the prediction of lung cancer on dataset 1.

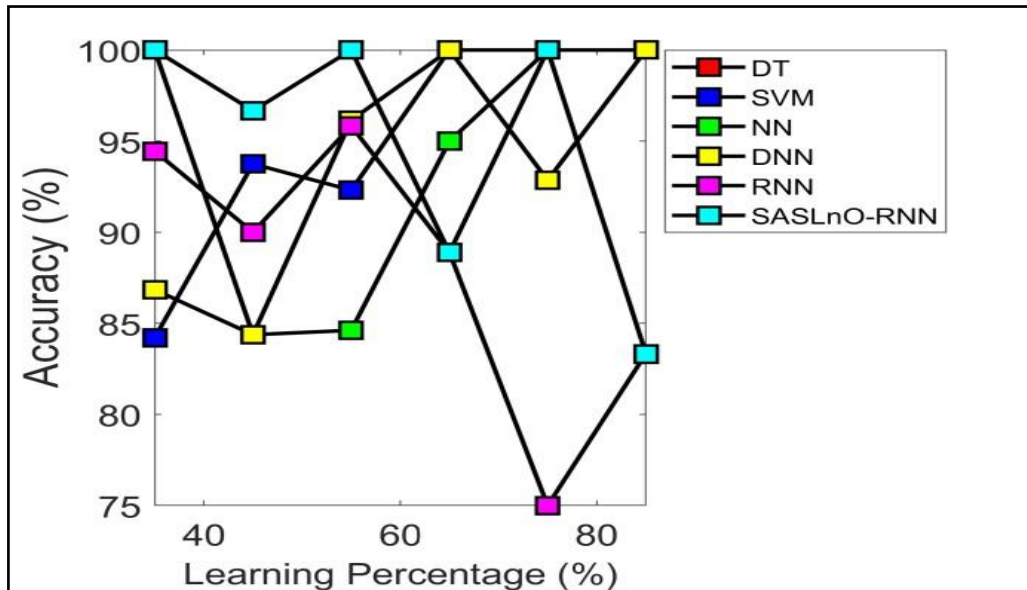


Figure 5. 6 Analysis for Accuracy using Machine Learning Algorithm

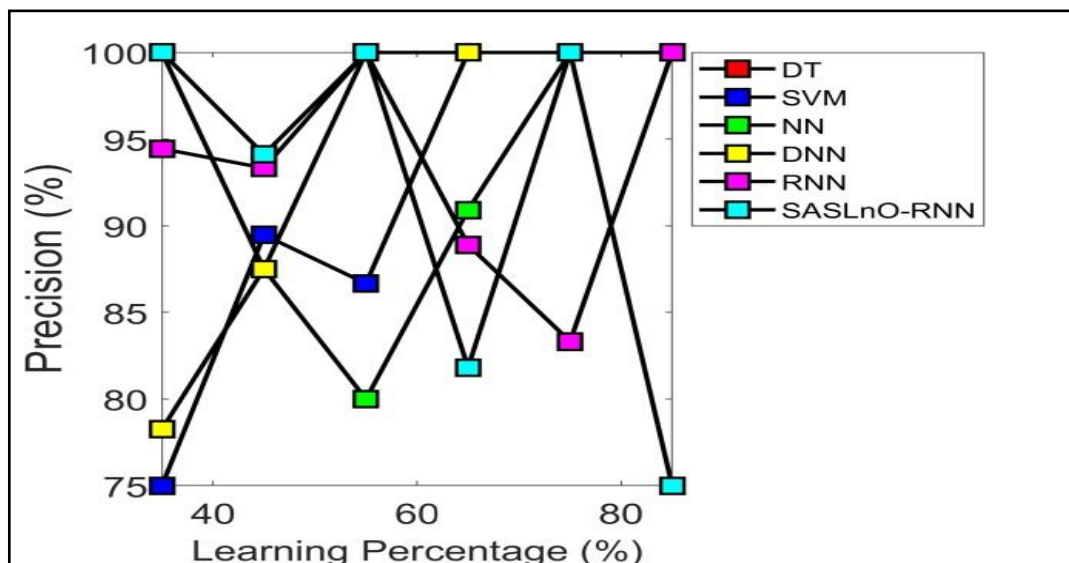


Figure 5. 7 Analysis for Precision using Machine Learning Algorithm

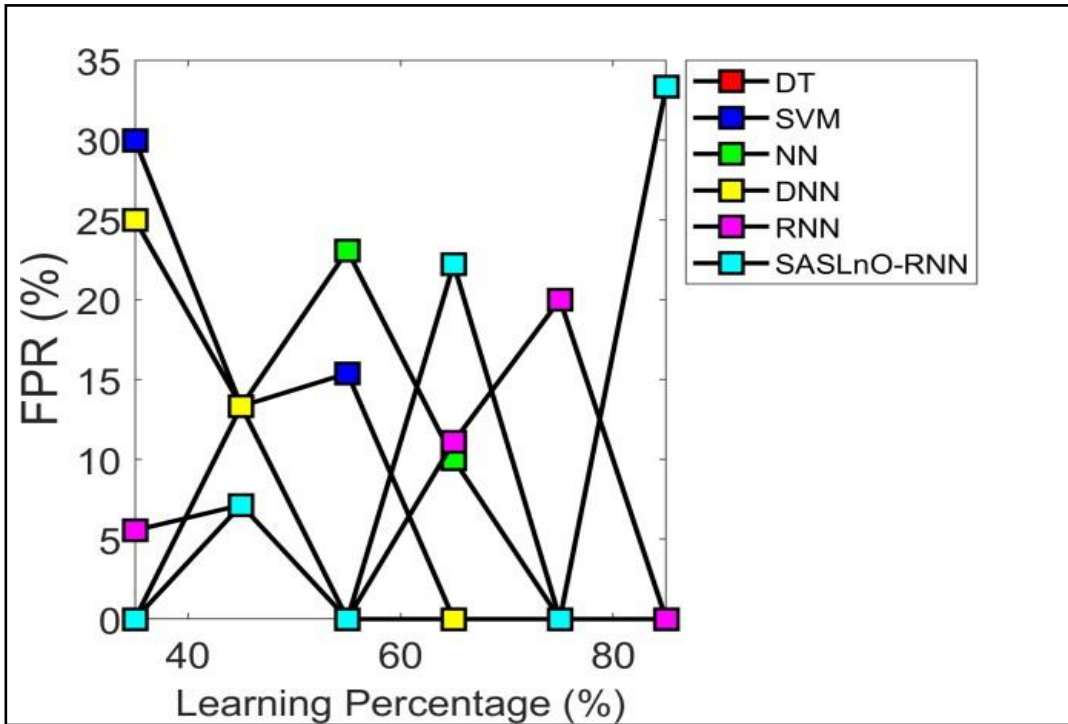


Figure 5.8 Analysis for FPR using Machine Learning Algorithm

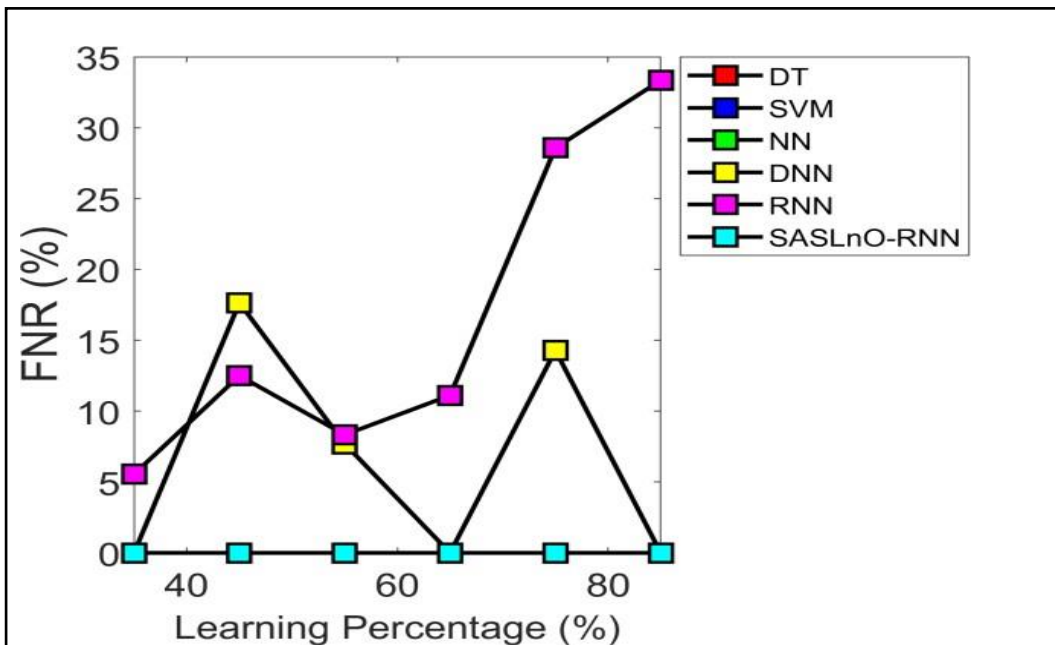


Figure 5.9 Analysis for FNR using Machine Learning Algorithm

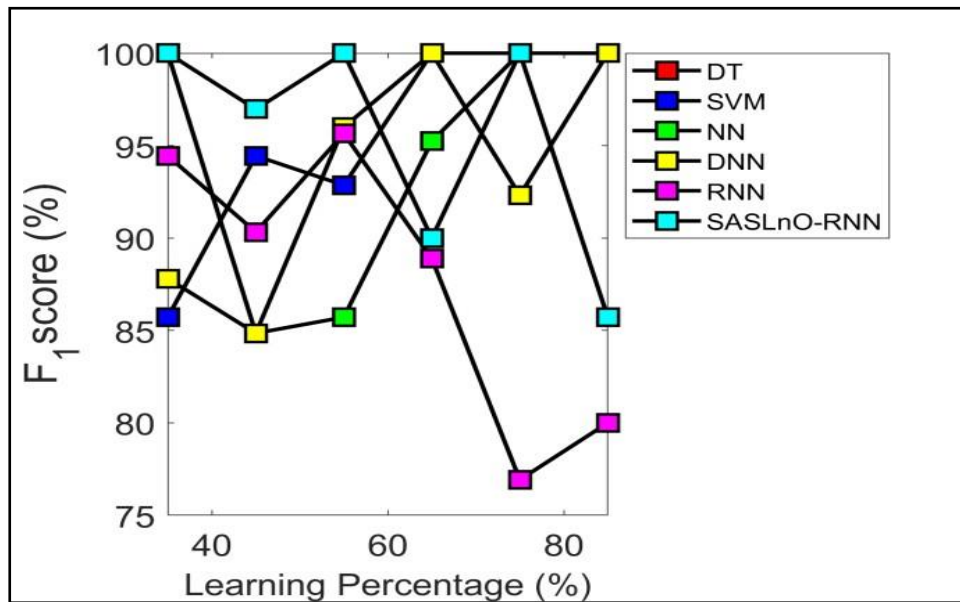


Figure 5.10 Analysis for F1-Score using Machine Learning Algorithm

5.5 Optimized Feature Extraction with Weighted Attributes Based on Correlation.

The proposed model for lung cancer diagnosis introduces a novel optimized weighted feature extraction approach to enhance its efficiency. This method's performance is evaluated in comparison with the standard feature extraction process, define in Figures 5.11 to 5.15.

- The accuracy of the optimized weighted feature extraction process at a learning percentage of 55 is greater than the accuracy of the standard feature extraction process by 25%.
- The precision of the optimized weighted feature extraction process demonstrates a 24% improvement over the standard feature extraction process, specifically when considering a knowledge percentage of 45.
- The FPR of the optimized weighted feature extraction process showcases an 82% reduction compared to the standard feature extraction process at a study ratio 45.
- The F1-score optimized weighted feature extraction process outperforms the standard feature extraction process by 30% at same study ratio.

Consequently, implementation proposed lung cancer analysis model utilizing the optimized weighted feature extraction process yields significantly enhanced outcomes.

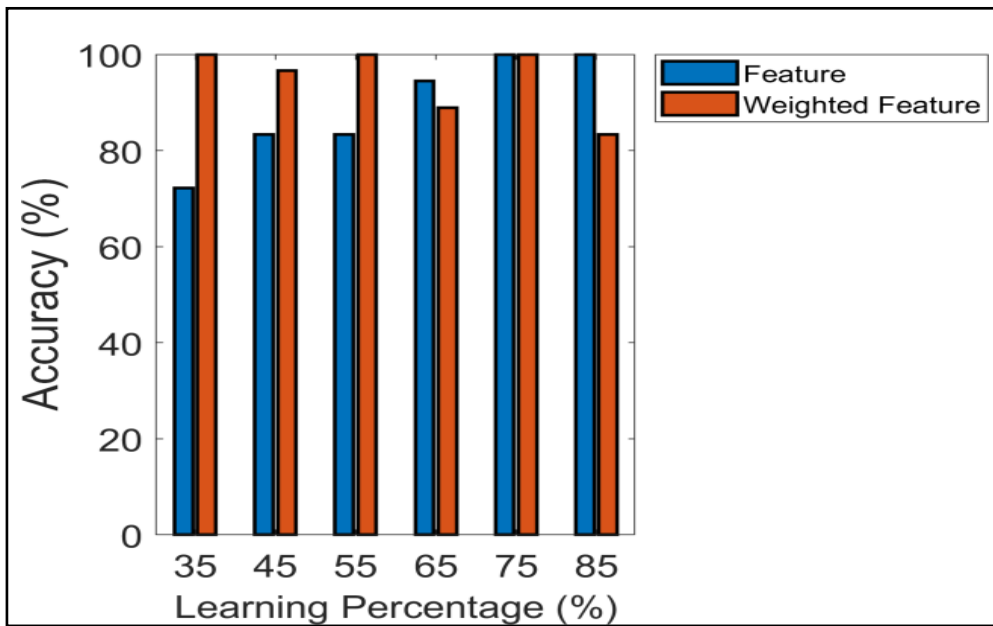


Figure 5. 11 Impact of a lung cancer diagnosis on accuracy using attribute correlation-based optimized weighted feature extraction.

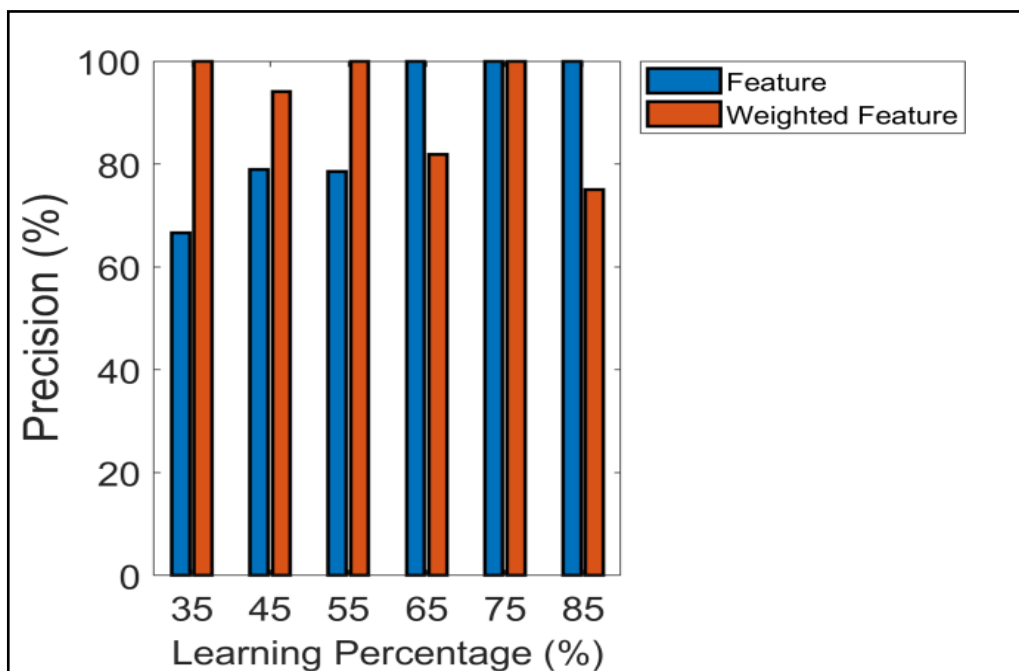


Figure 5. 62 Impact lung cancer diagnosis on attribute correlation-based optimized weighted feature extraction for rescission.

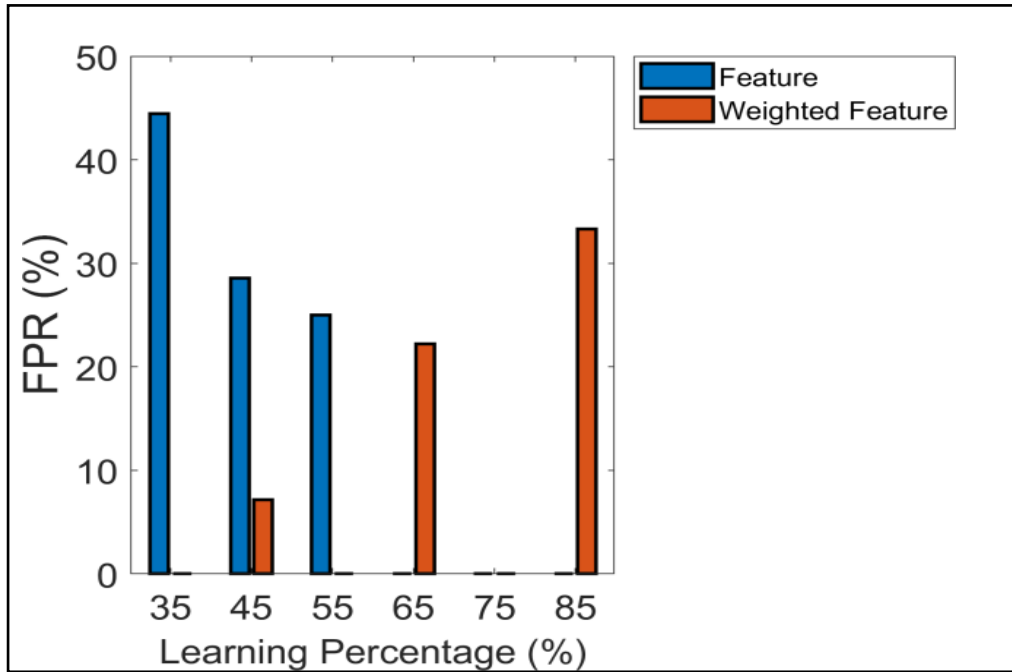


Figure 5.13 Impact lung cancer diagnosis on false positive rate (FPR) using attribute correlation-based optimized weighted features.

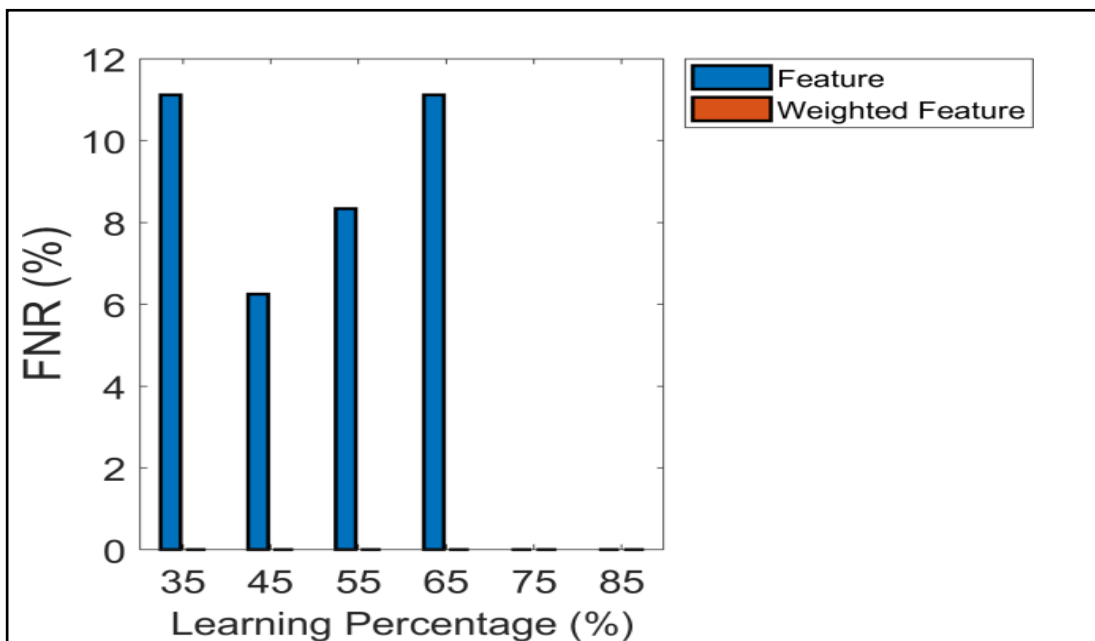


Figure 5.14 Impact lung cancer diagnosis for false negative rate (FNR) using optimized weighted features based on attribute correlation.

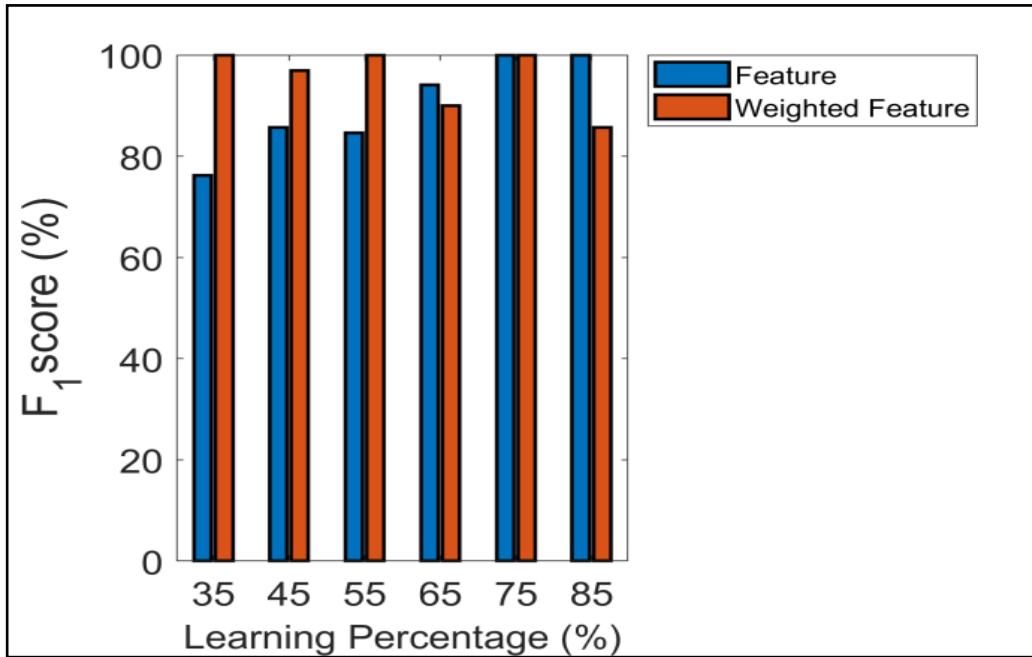


Figure 5. 15 Impact diagnosis of lung cancer on F1 score concerning attribute correlation-based optimized weighted feature extraction.

5.6 Performance Evaluation of Existing Techniques

The valuation of proposed lung cancer analysis model's performance is shown in Figures 5.16 and 5.17, where it is compared against an existing lung cancer diagnosis model across different learning percentages for two datasets.

- For dataset 1, the SA-SL_nO-RNN achieves significantly higher accuracy rates than Adaboost, DT, DBN, and Boosted SVM, with enhancements of 5.5%, 3.2%, 4.3%, and 1% respectively, at a learning percentage of 75.
- For dataset 2, SA-SL_nO-RNN shows a substantially greater accuracy rate than Adaboost, DT, DBN, and Boosted SVM by margins of 7.9%, 3.2%, 5.5%, and 6.7% respectively, at a learning percentage of 85%.

As a result, the designed lung cancer diagnosis model using the SA-SL_nO-RNN approach showcases remarkable performance advantages when compared to conventional methods.

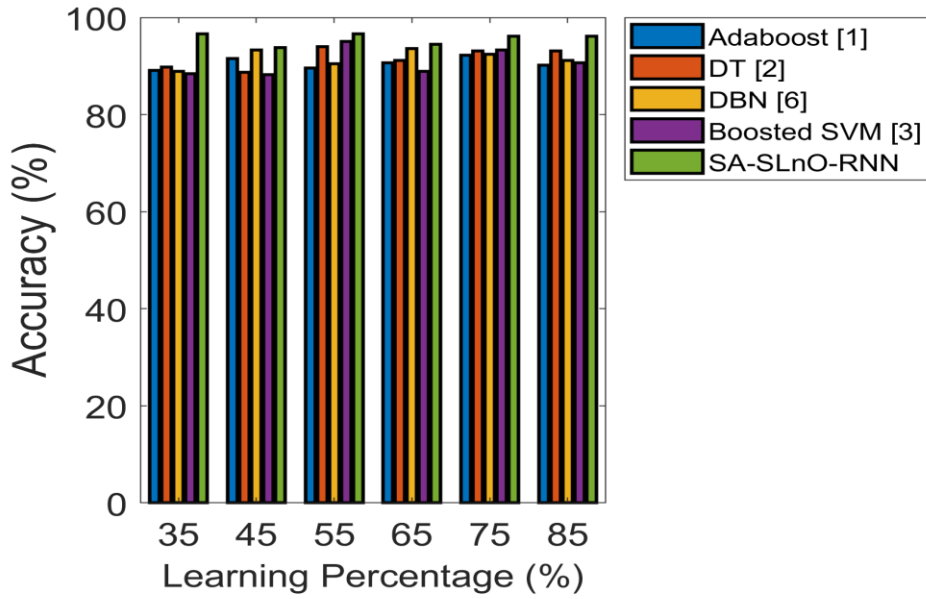


Figure 5. 16 Performance evaluation of the created lung cancer diagnosis using the currently used traditional methods for (a) dataset 1

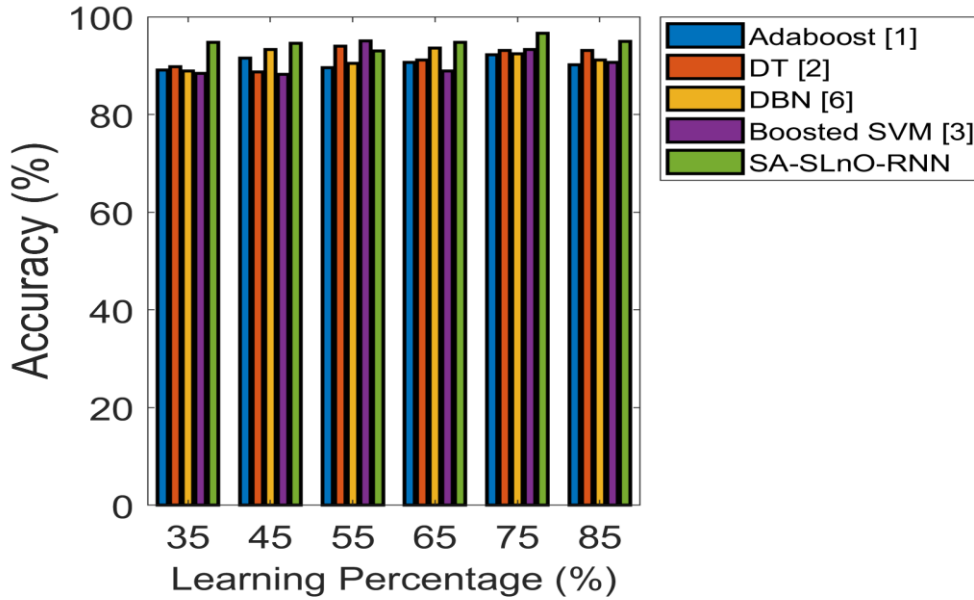


Figure 5. 17 Performance evaluation of the created lung cancer diagnosis using the currently used traditional methods for dataset 1(a)

5.7 Performance Evaluation for Dataset 2

The evaluation of progress lung cancer identification model is presented for dataset 2, along with existing algorithms and classifiers, as shown in Figures 5.18 and 5.19.

- The accuracy achieved by the proposed SLnO-RNN for lung cancer detection shows illustrious superiority, superior PSO-RNN, GWO-RNN, GSO-RNN, and even its previous iteration of SLnO-RNN by margins of 1.5%, 4.3%, 3.1%, and 5.6% respectively, at a study ratio 75.
- The execution of lung cancer diagnosis, measured in terms of accuracy, using the proposed SLnO-RNN method, outperforms DT, SVM, NN, DNN, and RNN by enhancements of 1.5%, 4.3%, 3.1%, and 5.6% respectively, again at a study ratio of 75.

These results highlight model's remarkable accuracy-driven performance, superior alternative algorithms, and classifiers, particularly in the context of dataset 2.

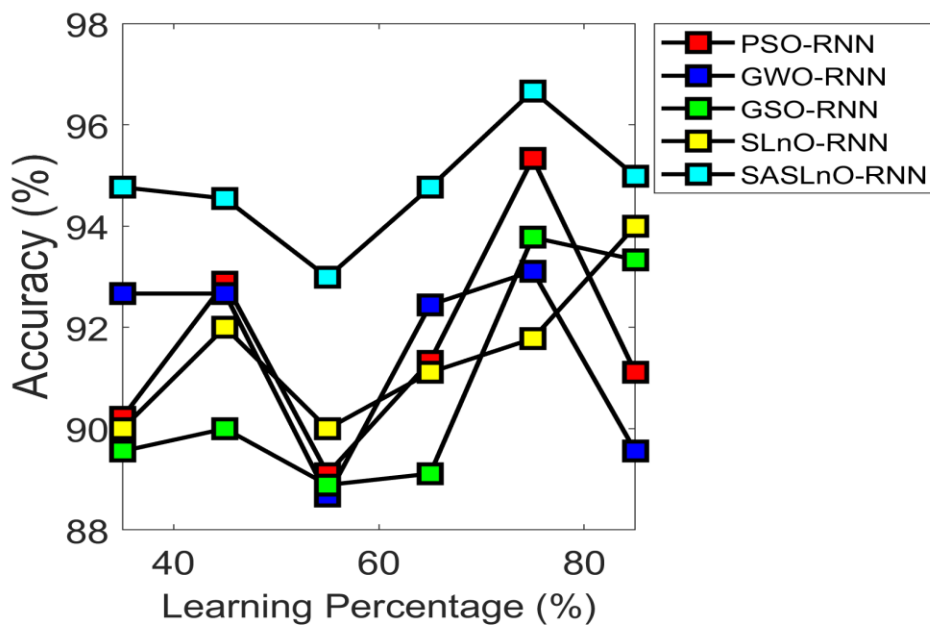


Figure 5. 18 Performance evaluation of the created lung cancer diagnosis for dataset 2 using existing conventional techniques and comparing with conventional algorithms

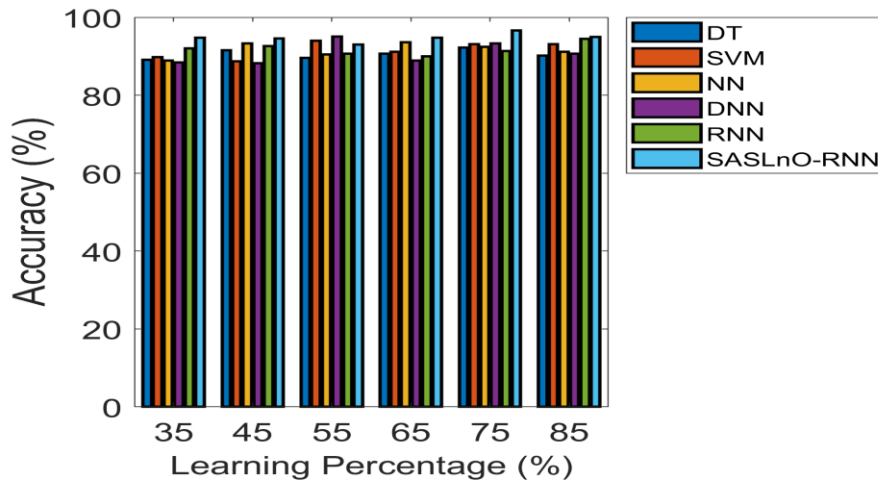


Figure 5. 19 Performance evaluation of the created lung cancer diagnosis for dataset 2 using existing conventional techniques and comparing with existing classifiers

5.8 Overall Performance Evaluation on Existing Models

Table 5.1 and Table 5.2 provide a comprehensive valuation of actual execution of current lung cancer analysis model in comparison to traditional optimization methods such as PSO-RNN, GWO-RNN, GSO-RNN and classifiers across both datasets.

- For dataset 1, the SA-SLnO-RNN approach determines a remarkable accuracy improvement of 31%, 7.4%, 20%, and 26% over traditional optimization method and its previous SLnO-RNN iteration, respectively.
- Furthermore, the sensitivity of the implemented SA-SLnO-RNN model beats that of traditional optimization methods, and SLnO-RNN by borders of 60%, 14%, 23%, and 74%, respectively, for dataset 1.
- Additionally, the SA-SLnO-FDR RNN model shows reductions of 64%, 11%, 68%, and 74% compared to traditional optimization methods, and SLnO-RNN, subsequently, for dataset 1.
- Similarly, for dataset 2, the SA-SLnO-based RNN achieves enhanced accuracy rates, surpassing traditional optimization methods, and SLnO-RNN by 1.7%, 2%, 5%, and 2.7% respectively.
- Additionally, the performance of the SA-SLnO-based RNN is evaluated alongside traditional classifiers. Notably, for dataset 1, the accuracy of the SA-SLnO-RNN method surpasses DT and SVM by 3.1%, while also outperforming NN, DNN, and RNN by 14.5%, 7.4%, and 3.1%, respectively.
- Moreover, the precision of the designed SA-SLnO-RNN model showcases

impressive improvements, with margins of 31.8%, 31.4%, 24.4%, 39%, and 47.3% over DT, SVM, NN, DNN, and RNN, respectively.

Collectively, these findings underscore the superior performance of current lung cancer analysis model utilizing optimized weighted feature extraction based on the optimized RNN approach, outperforming other algorithms in terms of accuracy and precision.

Table 5. 1 Overall Performance Analysis of The Developed and Traditional Meta-Heuristic-Based RNN [27]

Measures	PSO-RNN	GWO-RNN	GSO-RNN	SLnO-RNN	SA-SLnO-RNN
Dataset 1					
Precision	0.83333	0.93333	0.8125	0.76471	0.94118
Sensitivity	0.625	0.875	0.8125	0.8125	1
Accuracy	0.73333	0.9	0.8	0.76667	0.96667
Specificity	0.85714	0.92857	0.78571	0.71429	0.92857
NPV	0.85714	0.92857	0.78571	0.71429	0.92857
MCC	0.49099	0.80178	0.59821	0.53035	0.93485
FPR	0.14286	0.071429	0.21429	0.28571	0.071429
F1-score	0.71429	0.90323	0.8125	0.78788	0.9697
FNR	0.375	0.125	0.1875	0.1875	0
FDR	0.16667	0.066667	0.1875	0.23529	0.058824
Dataset 2					
Measures	PSO-RNN	GWO-RNN	GSO-RNN	SLnO-RNN	SA-SLnO-RNN
Accuracy	0.92889	0.92667	0.9	0.92	0.94544
Sensitivity	0.56667	0.48571	0.40816	0.52778	0.64
Specificity	0.95476	0.96386	0.9601	0.95411	0.98353
Precision	0.47222	0.53125	0.55556	0.5	0.69565
FPR	0.045238	0.036145	0.0399	0.045894	0.016471
FNR	0.43333	0.51429	0.59184	0.47222	0.36

NPV	0.95476	0.96386	0.9601	0.95411	0.98353
FDR	0.52778	0.46875	0.44444	0.5	0.30435
F1-score	0.51515	0.50746	0.47059	0.51351	0.66667
MCC	0.47943	0.46848	0.42284	0.47017	0.64855

Table 5. 2 Overall Performance Analysis of the Developed and Traditional Classifiers [27]

Measures	DT	SVM	NN	DNN	RNN	SA-SL _n O-RNN
Dataset 1						
Precision	0.89474	0.89474	0.875	0.875	0.93333	0.94118
Sensitivity	1	1	0.82353	0.82353	0.875	1
Accuracy	0.9375	0.9375	0.84375	0.84375	0.9	0.96667
Specificity	0.86667	0.86667	0.86667	0.86667	0.92857	0.92857
NPV	0.86667	0.86667	0.86667	0.86667	0.92857	0.92857
MCC	0.88059	0.88059	0.68885	0.68885	0.80178	0.93485
FPR	0.13333	0.13333	0.13333	0.13333	0.071429	0.071429
F1-score	0.94444	0.94444	0.84848	0.84848	0.90323	0.9697
FNR	0	0	0.17647	0.17647	0.125	0
FDR	0.10526	0.10526	0.125	0.125	0.066667	0.058824
Dataset 2						
Accuracy	0.91556	0.88667	0.93333	0.88222	0.92667	0.94544
Sensitivity	0.475	0.33962	0.55882	0.28302	0.54839	0.64
Specificity	0.95854	0.9597	0.96394	0.96222	0.95465	0.98353
Precision	0.52778	0.52941	0.55882	0.5	0.47222	0.69565
FPR	0.041463	0.040302	0.036058	0.037783	0.045346	0.016471
FNR	0.525	0.66038	0.44118	0.71698	0.45161	0.36

NPV	0.95854	0.9597	0.96394	0.96222	0.95465	0.98353
FDR	0.47222	0.47059	0.44118	0.5	0.52778	0.30435
F1-score	0.5	0.41379	0.55882	0.36145	0.50746	0.66667
MCC	0.45477	0.36508	0.52277	0.31691	0.46961	0.64855

The following figures 5.20 to 5.39 give complete performance comparison of the developed and traditional meta-heuristic-based RNN:

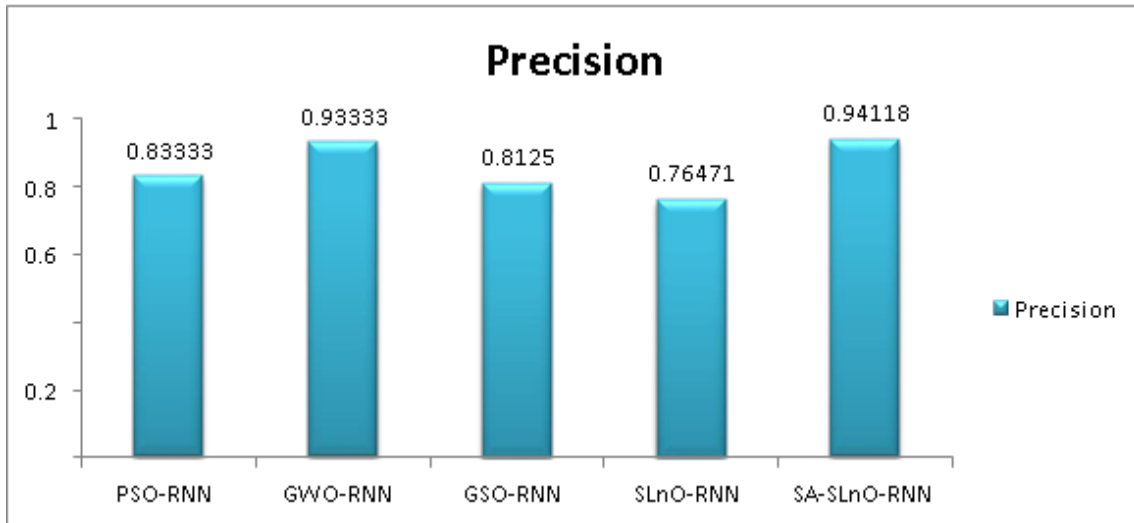


Figure 5. 20 Comparison of the Precision of a Meta-Heuristic-Based RNN to Traditional Methods

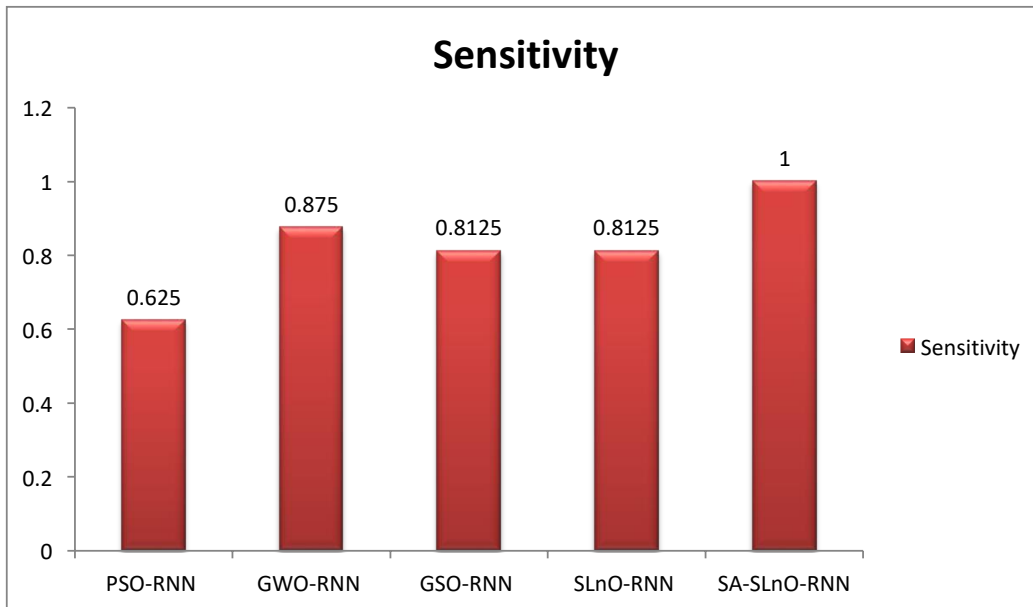


Figure 5. 21 Comparison of the Sensitivity of a Meta-Heuristic-Based RNN to Traditional Methods

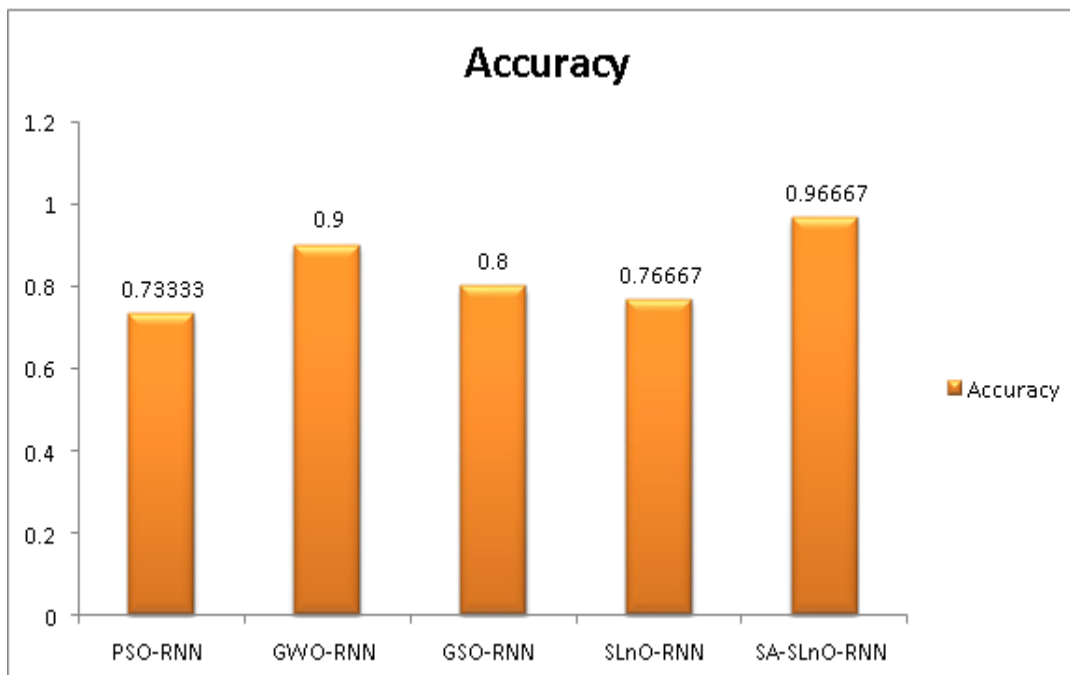


Figure 5. 22 Comparison of the Accuracy of a Meta-Heuristic-Based RNN to Traditional Methods

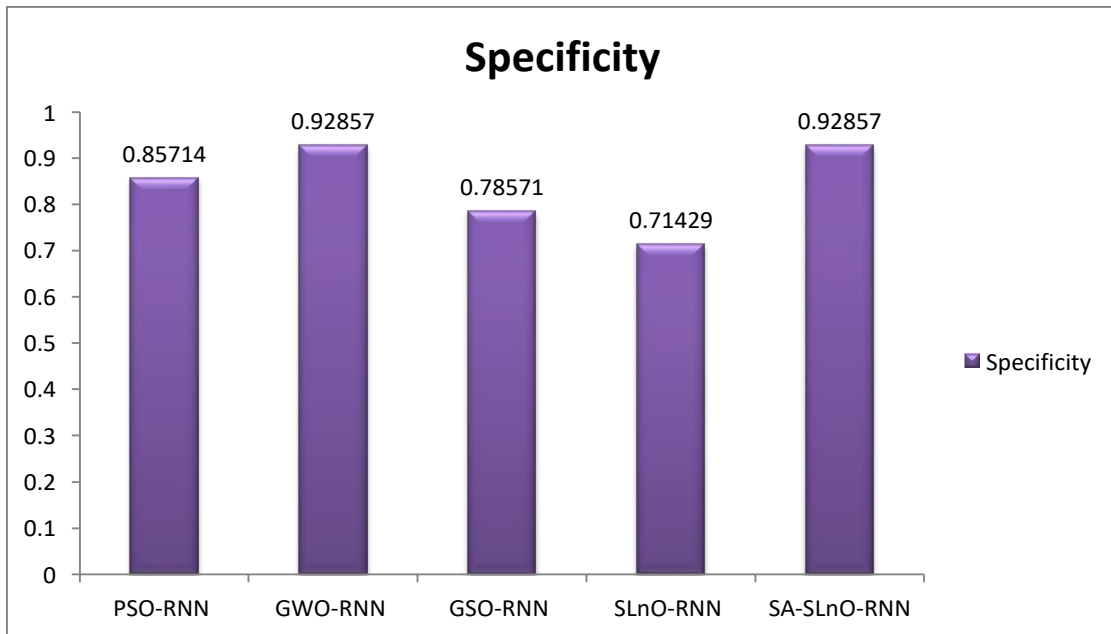


Figure 5. 23 Comparison of the Specificity of a Meta-Heuristic-Based RNN to Traditional Methods

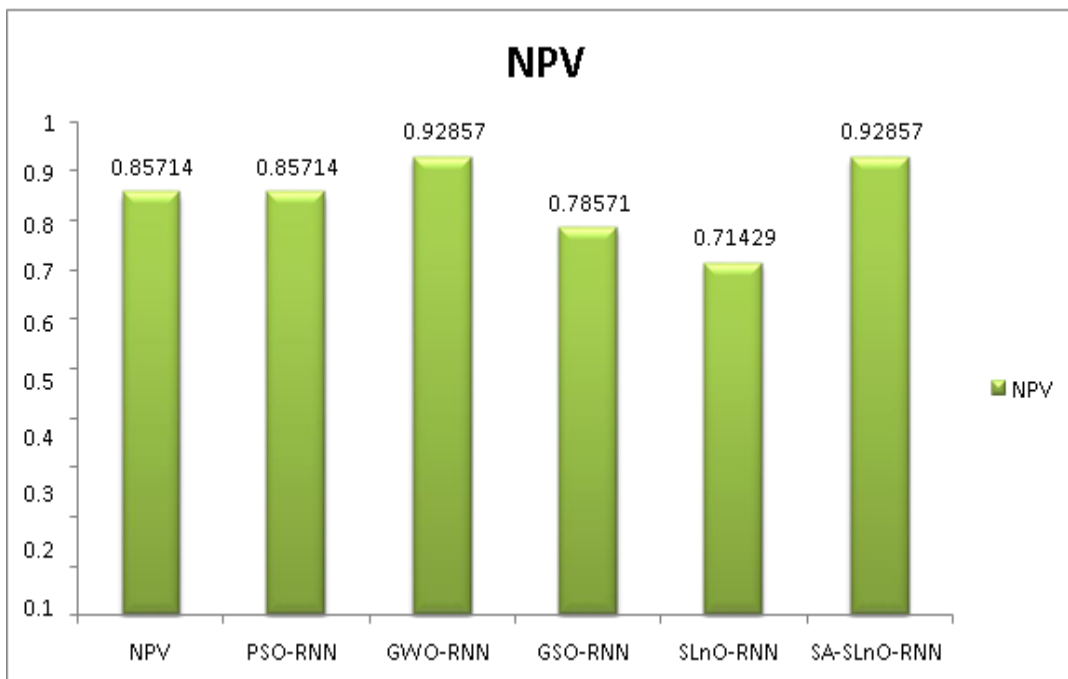


Figure 5. 24 Comparison NVP of a Meta-Heuristic-Based RNN to Traditional Methods

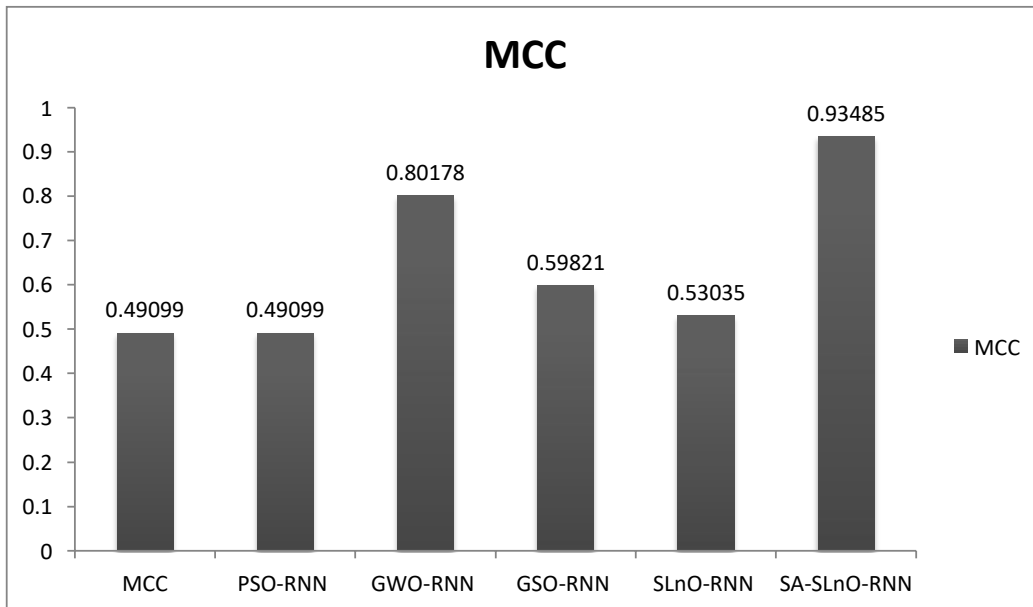


Figure 5. 25 Comparison MCC of a Meta-Heuristic-Based RNN to Traditional Methods

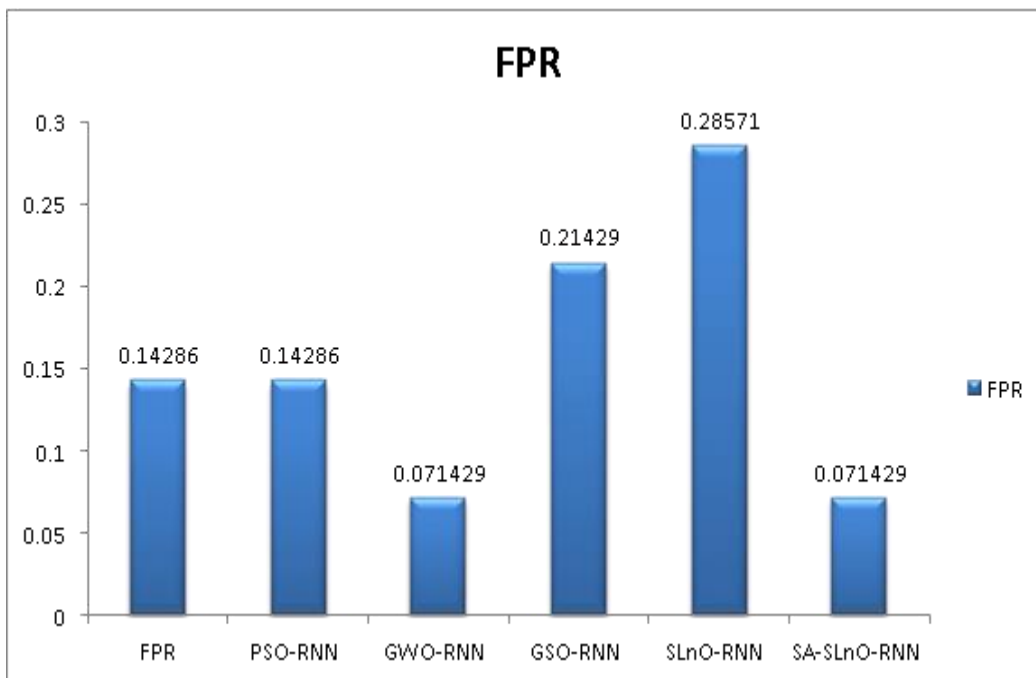


Figure 5. 26 Comparison of FPR of a Meta-Heuristic-Based RNN to Traditional Methods

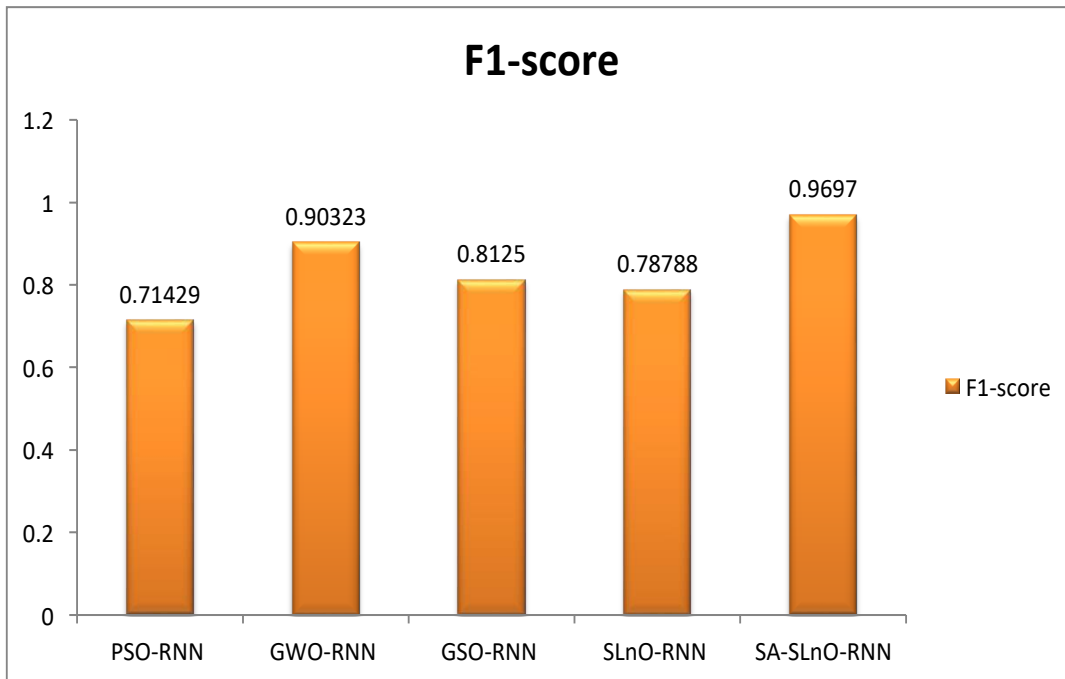


Figure 5. 27 Comparison F1-score of a Meta-Heuristic-Based RNN to Traditional Methods

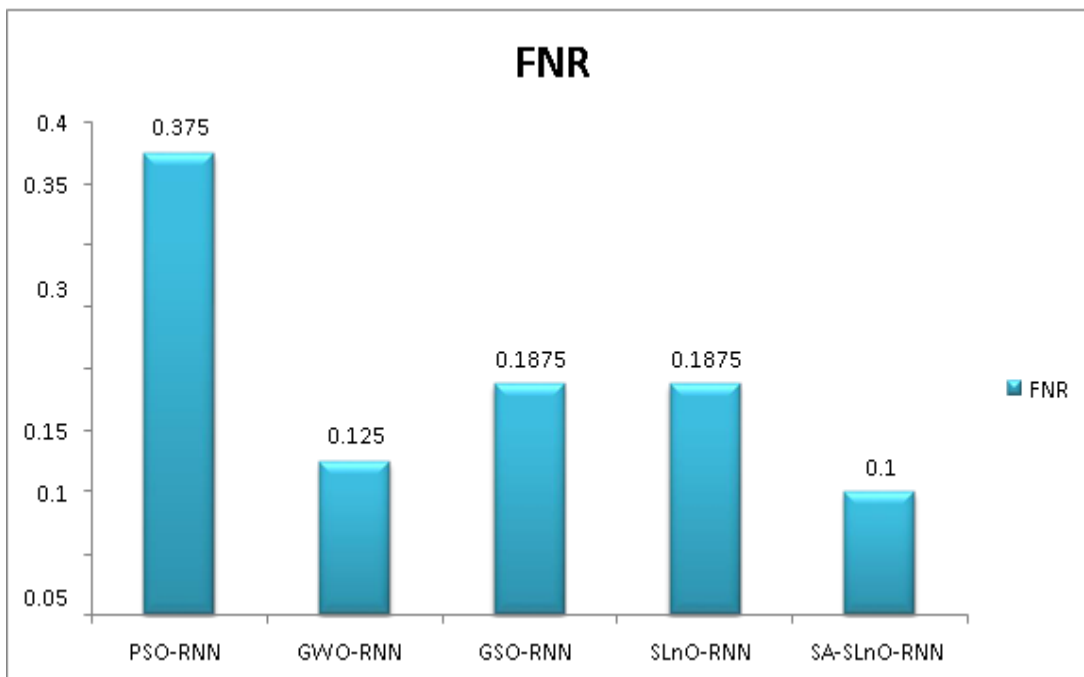


Figure 5. 28 Comparison FNR of a Meta-Heuristic-Based RNN to Traditional Methods

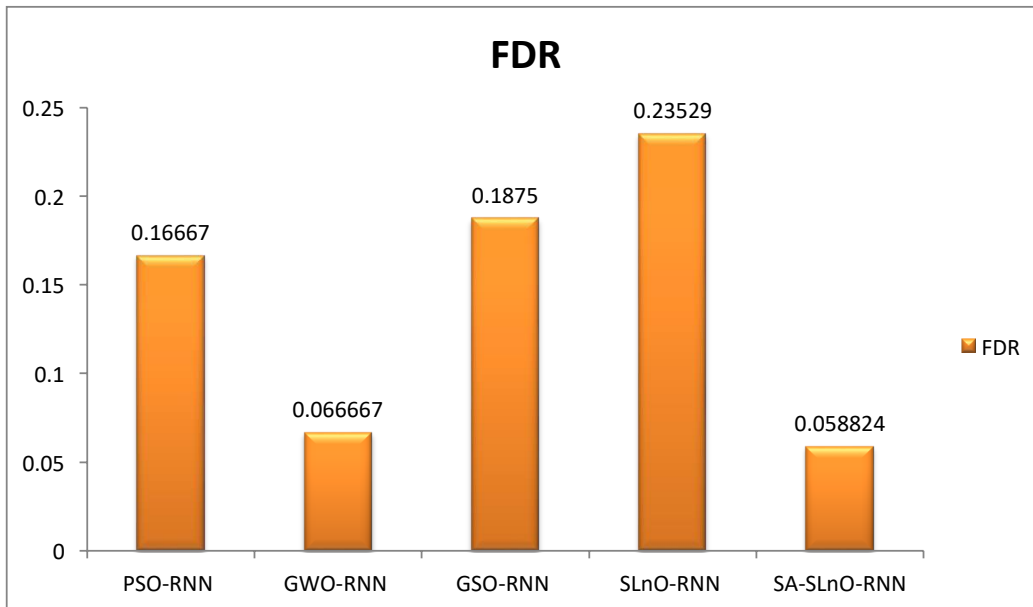


Figure 5. 29 Comparison FDR of a Meta-Heuristic-Based RNN to Traditional Methods

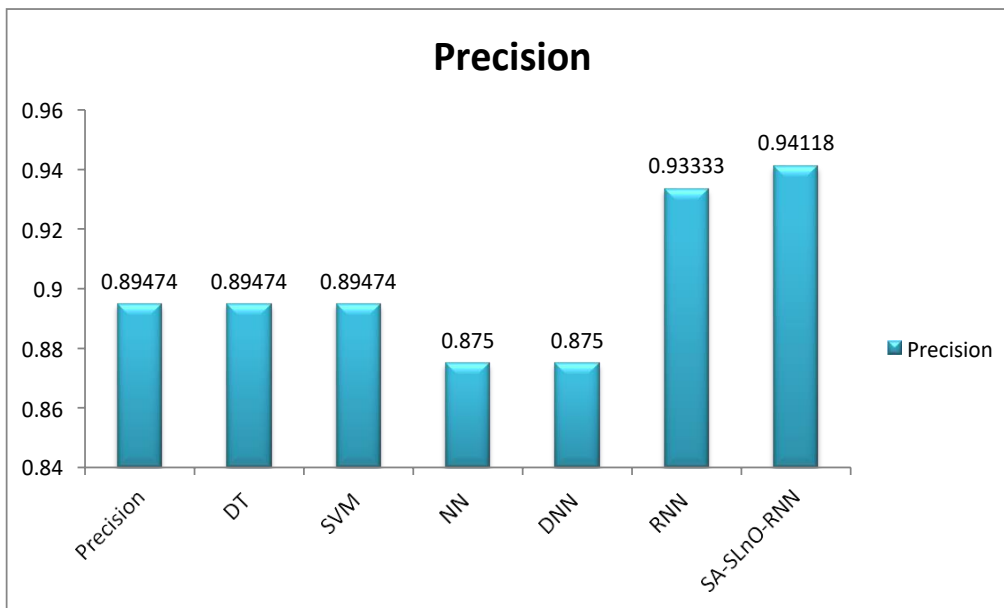


Figure 5. 30 Precision Comparison of the developed and traditional classifiers

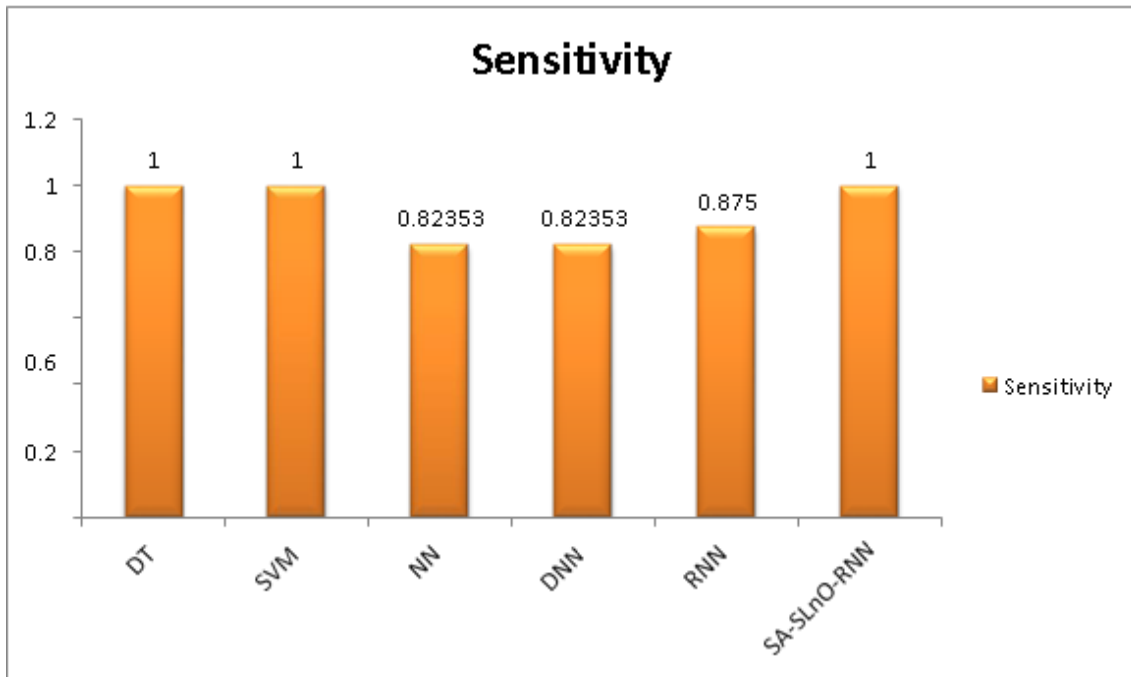


Figure 5. 31 Sensitivity Comparison of the developed and traditional classifiers

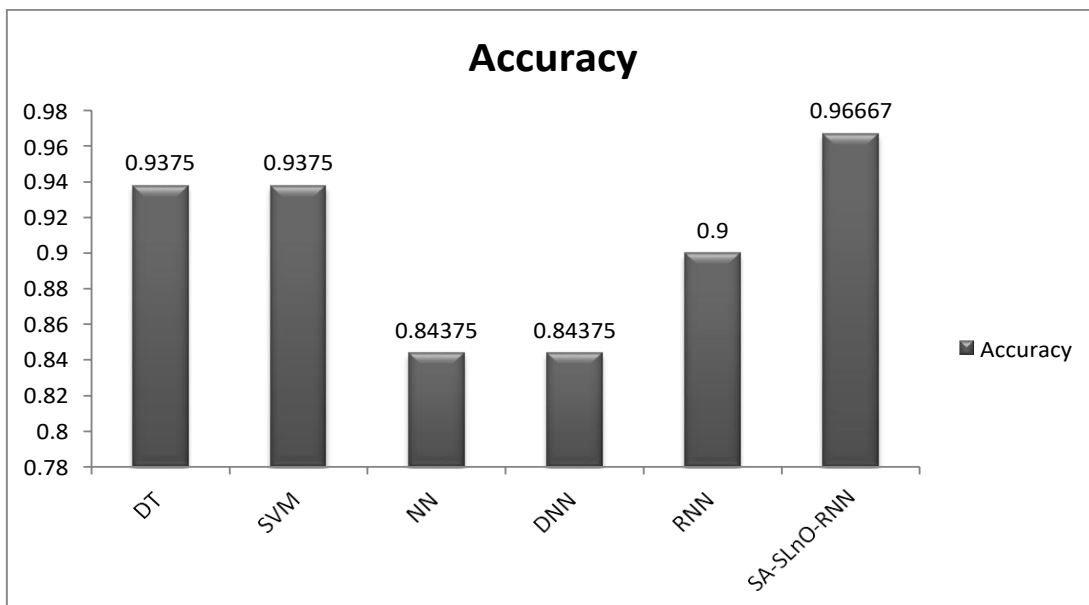


Figure 5. 32 Accuracy Comparison of the Developed and Traditional Classifiers

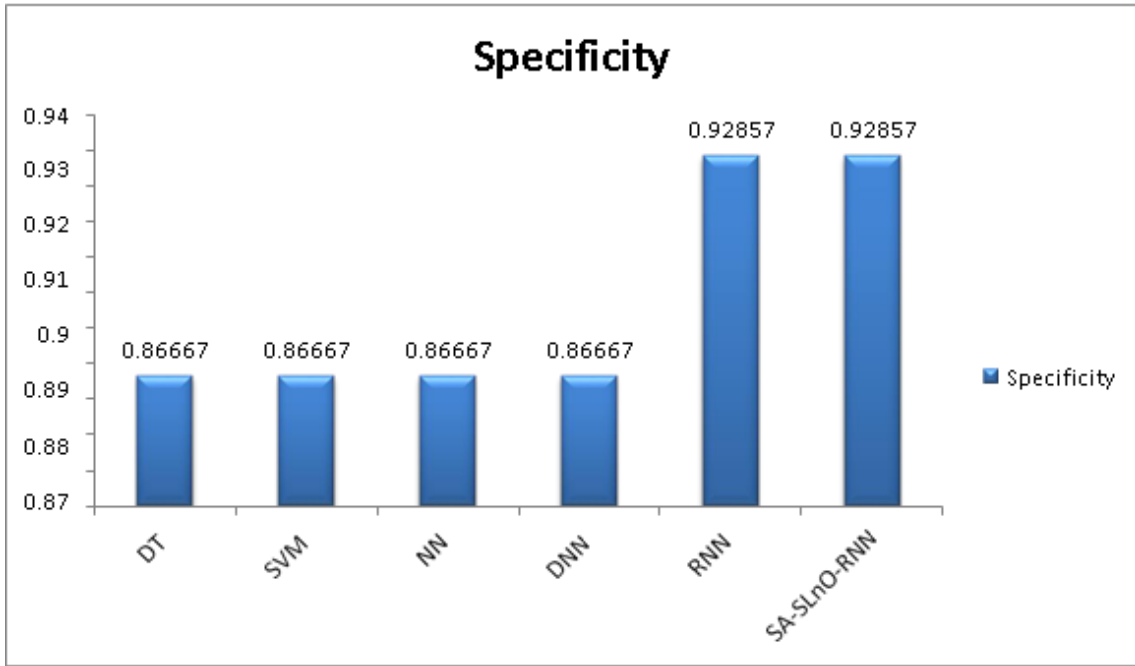


Figure 5. 33 Specificity Comparison of the developed and traditional classifiers

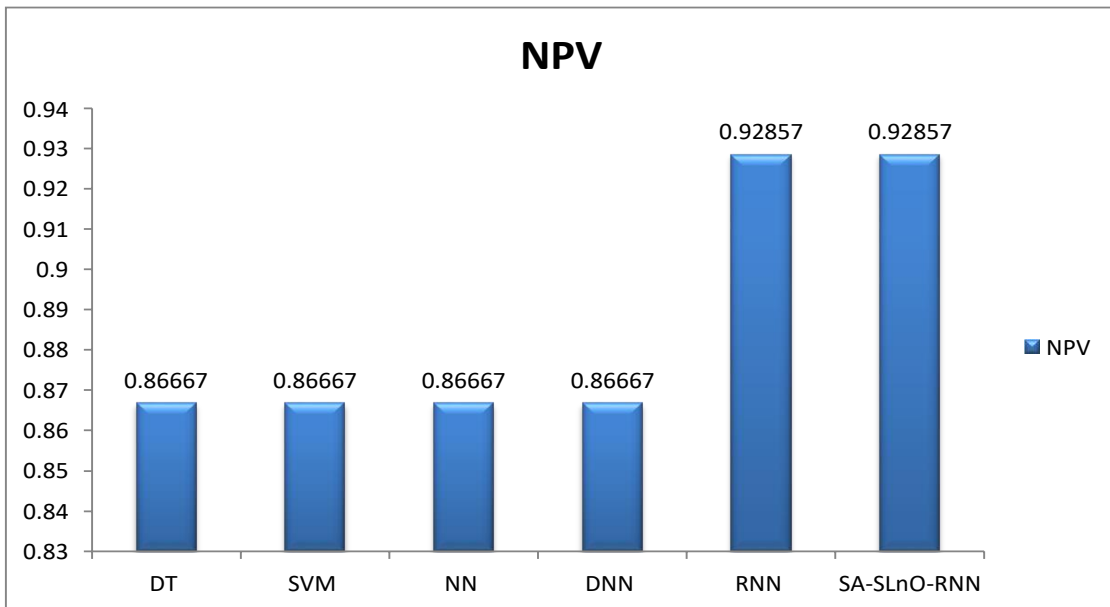


Figure 5. 34 NPV Comparison of the developed and traditional classifiers

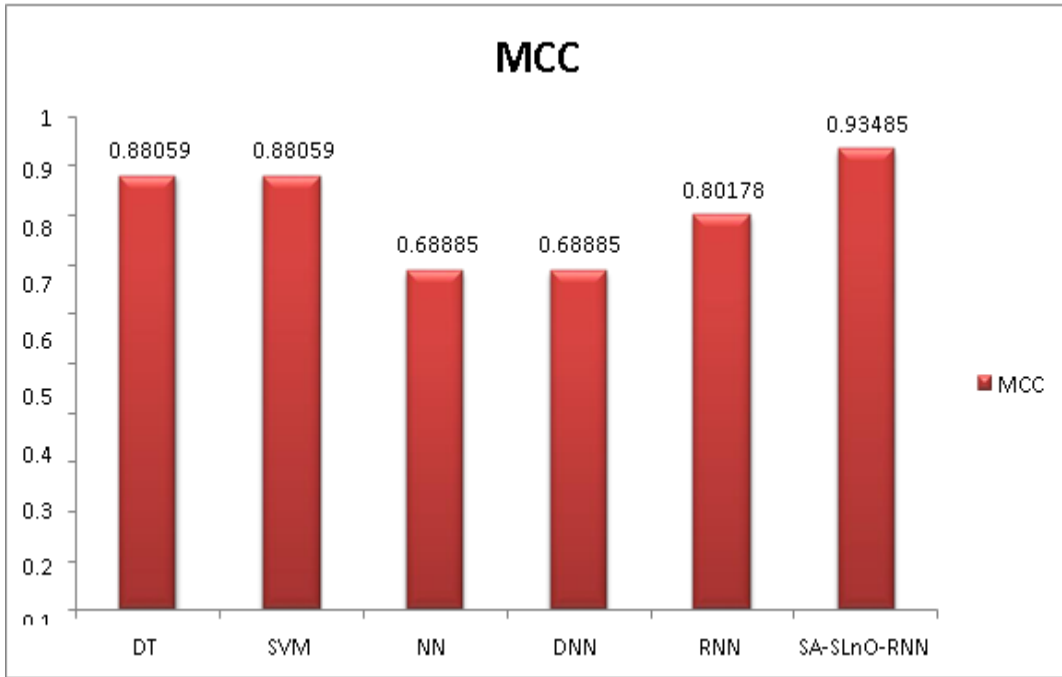


Figure 5. 35 MCC Comparison of the developed and traditional classifiers

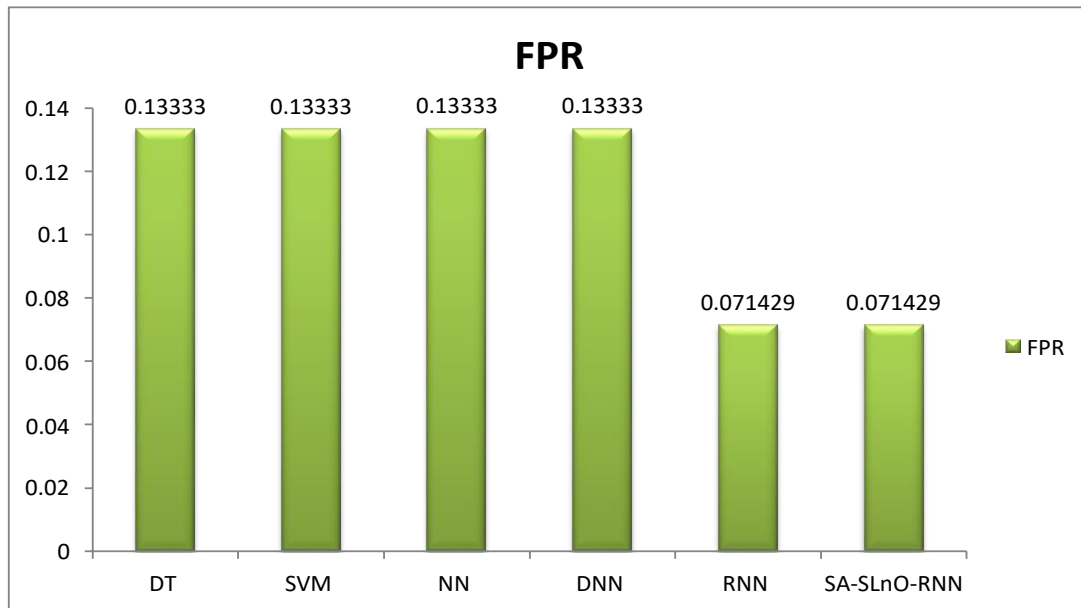


Figure 5. 36 FPR Comparison of the developed and traditional classifiers

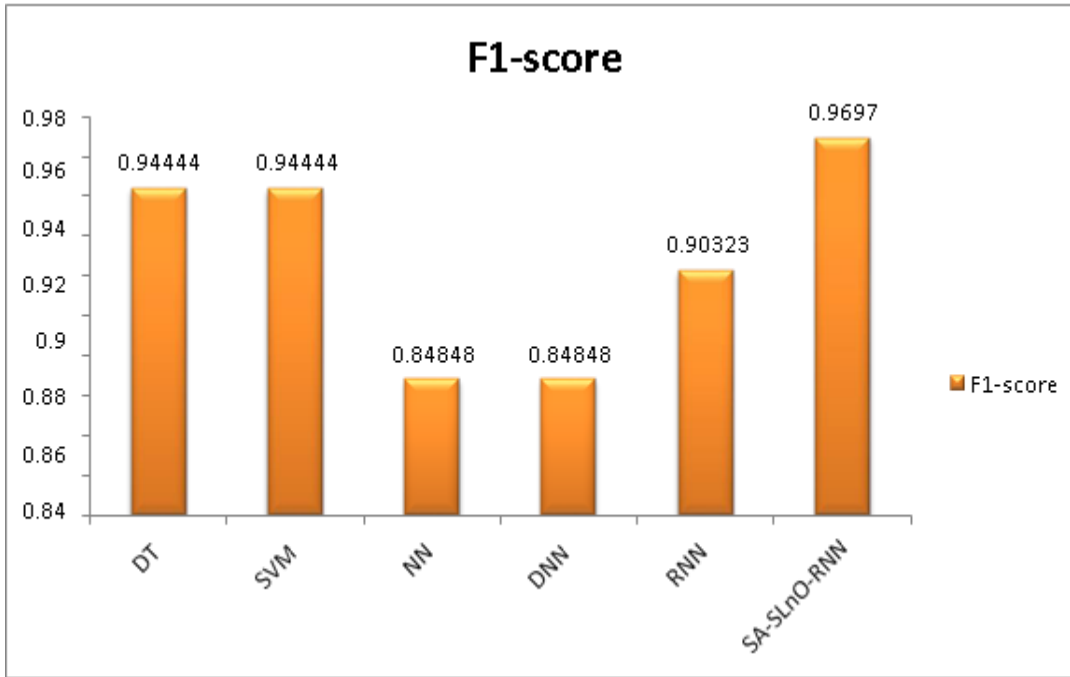


Figure 5. 37 F1-Score Comparison of the developed and traditional classifiers

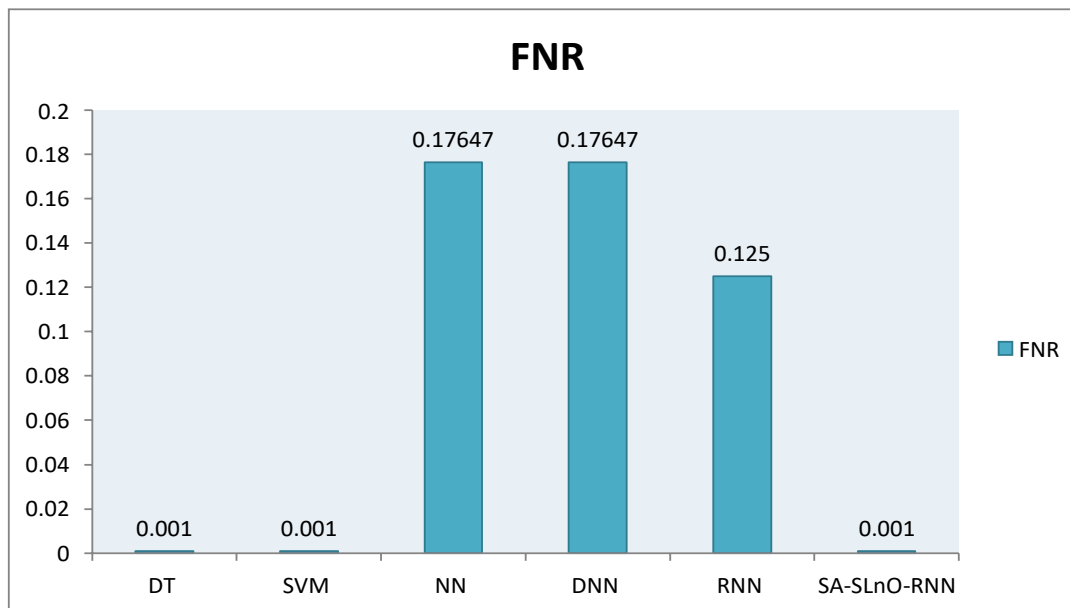


Figure 5. 38 FNR Comparison of the developed and traditional classifiers

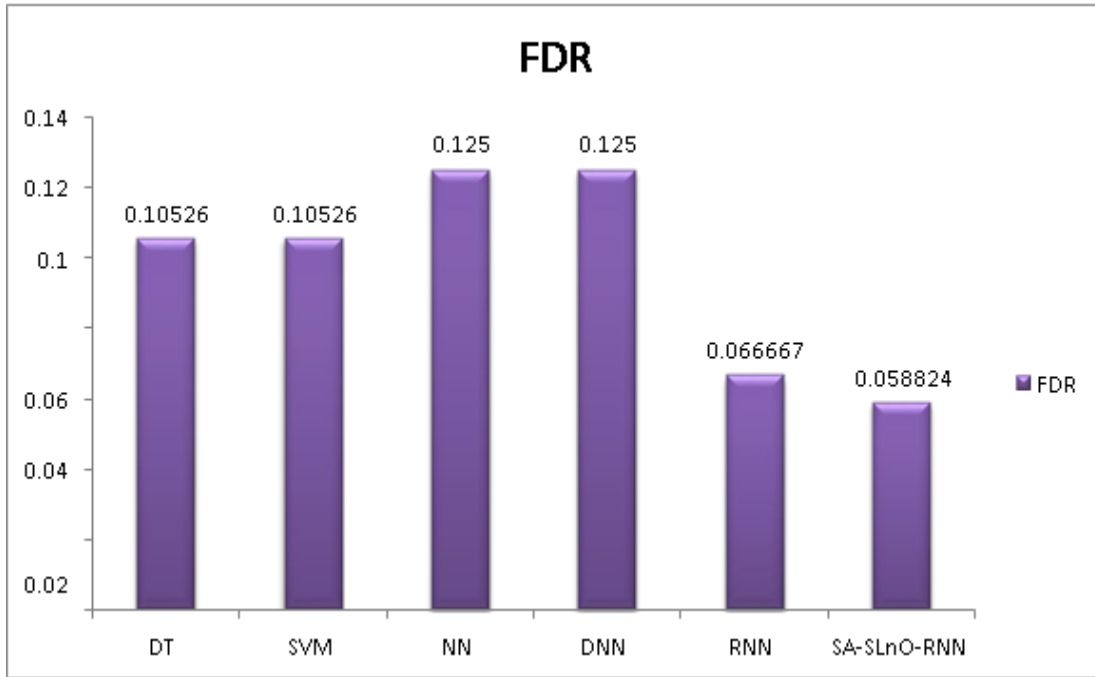


Figure 5. 39 FDR Comparison of the developed and traditional classifiers

5.9 Performance Analysis on the Time Complexity of Proposed Model.

Early prediction lung cancer disease imperative criteria are selecting the best-performing model along with other performance parameters like accuracy, precision and F1-Score. Proposed model not only performs better in terms of the most accurate results of prediction but also the early prediction in terms of time was achieved. Tables 5.3,5.4 and 5.5 show the comparative analysis time complexity of each model for lung cancer prediction.

- Table 5.3, observed that the proposed SA-SLnO-RNN prediction time was just 0.24679 seconds compared to 2.1381 for GSO-RNN, 0.5652 for GWO-RNN, and 11.786 for PSO-RNN on various optimization techniques used with Neural Network.
- When evaluating time complexity on different classifiers, then found that the duration of the SA-SLnO-RNN is 94%, 45%, 9.8%, 82%, and 28.6% minimized than SVM, NN, DNN, and RNN, respectively.
- The SA-SLnO-RNN's time is also 95.6%, 88%, 48.5%, and 9.9% faster than that of Adaboost, DT, DBN, and Boosted SVM, respectively.

Hence, the recommended model perceives reduced time when compared with the other approaches.

Table 5. 3 Performance evaluation of the lung cancer diagnostic model using several optimization-based techniques regarding complexity [27]

Measures	PSO-RNN	GWO-RNN	GSO-RNN	SLnO-RNN	SA-SLnO-RNN
Time (sec)	11.786	0.5652	2.1381	0.28134	0.24679

Table 5. 4 Performance evaluation of the built lung cancer diagnostic model using several classifiers in terms of time complexity

Measures	DT	SVM	NN	DNN	RNN	SA-SLnO-RNN
Time (sec)	4.184	0.45136	0.2737	1.3894	0.34601	0.24679

Table 5. 5 Comparing the lung cancer diagnosis model's overall performance in terms of time complexity to several traditional models

Measures	Adaboost	DT	DBN	Boosted SVM	SA-SLnO-RNN
Time (sec)	5.6683	2.1382	0.47967	0.274	0.24679

Chapter Summary

This chapter explores the results of the models proposed and used in this work to develop a deep learning-based lung cancer detection system. Calculated the performance metrics that were used for the optimized weighted feature extraction with attribute correlation, the machine learning algorithm, the overall performance evaluation, and the estimation of the proposed model. A novel lung cancer detection model utilizing an enhanced deep learning technique and attribute correlation-based enhanced weighted feature extraction. The stages of this suggested diagnosis paradigm were data mining, attribute extraction, strategy collecting, and description. The feature extraction process was used to analyze the acquired input data using t-SNE and PCA algorithms. The features were initially extracted through the attribute correlation-based optimized weighted attribute mining process, an optimization using SA-SLnO recommended method.

This research proposed a novel lung cancer diagnostic model using correlation-based optimal weighted feature extraction and an optimal deep learning approach. The proposed diagnosis model consists of stages such as data gathering, feature extraction, measurement strategy collection, and prediction. The collected input data were entered into the feature extraction process using t-SNE and PCA methods. The extracted features are processed using an algorithm called SA-SL_nO in a correlation-based optimized feature extraction process. Data features are classified as normal and abnormal using the optimized RNN classifier. The optimization of the hidden neuron is done using the proposed SA-SL_nO algorithm. Observational analysis shows that the performance of SA-SL_nO-RNN is 3.1% better than DT and SVM, 14.5% better than NN and DNN, and 7.4% better than RNN. Therefore, the proposed model produces better performance than other optimization and machine learning methods.

As a result, the proposed model provides better performance than optimization and machine learning techniques. The challenge for the proposed SA-SL_nO-RNN model is that it does not solve the combinatorial optimization problem. It does not effectively balance the use of algorithms and search capabilities. The main problem is the lack of data, latitudinal explosion and low local search capabilities. Therefore, this problem can be solved by using advanced smart techniques to detect lung cancer in the future. Since medical data is more sensitive than other data, new privacy protection algorithms may also be implemented in the future to protect data from unauthorized hackers.

References

- [1] M. S. P. & Z. Rahman, "New Method for Lung Nodule Detection Using Deep Neural Networks for CT Images", *International Conference on Electrical, Computer and Communication Engineering*, pp. 1-6, 2019.
- [2] A. K. F. S. A. L. N. D. N. W. B. E. A. G. G. & E.-B. A. Soliman, "Image-based CAD system for accurate identification of lung injury," *International Conference on Image Processing*, pp. 121-125, 2016.
- [3] M. N. a. A. B. A. A. Raweh, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," *IEEE Access*, Vols. vol. 6,, pp. 15212-15223, 2018.
- [4] V. & G. H. Mekali, "Solitary pulmonary nodules classification based on tumor size and volume of nodules," *International Conference on Applied and Theoretical Computing and Communication Technology*, pp. 394-398, 2018.
- [5] S. R. R. & A. Potghan, "Multi-Layer Perceptron Based Lung Tumor Classification," *International Conference on Electronics, Communication and Aerospace Technology*, pp. 499-502, 2018.
- [6] A. T. W. A. S. N. K. Y. H. A. O. & N. K. Haga, "Classification of early-stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variab," *Radiation Oncology*, vol. 13, pp. 27-35, 2018.
- [7] S. & R. N. Shanthi, "Non-small-cell lung cancer prediction using radiomic features and machine learning methods," *International Journal of Computers and Applications*, pp. 1-9, 2019.
- [8] W. J. M. D. J. C. P. & Y. Sun, "Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis," *Radiation Oncology*, vol. 11, pp. 1-8, 2018.
- [9] Y. N. H. W. S. W. B. & L. Riti, "Feature extraction for lesion margin characteristic classification from CT Scan lungs image", *International Conference on Information Technology*, *Information Systems and Electrical Engineering*, pp. 54-58, 2018.
- [10] R. G. S. M. P. & J. Janudhivya, "A new approach for lung cancer cell detection using Mumford-shah algorithm," *International Conference on Communication and Signal Processing*, pp. 1041-1044, 2016.
- [11] P. & G. S. Sangamithraa, "Lung tumour detection and classification using EK-Mean clustering," *International Conference on Wireless Communications, Signal Processing and Networking*, pp. 2201-2206, 2016.

- [12] F. T. G. M. S. Kanavati, "Weakly-supervised learning for lung carcinoma classification using deep learning," *Sci Rep* 10, 2020.
- [13] M. B. S. K. Z. & K. F. Faisal, ""An Evaluation of Machine Learning Classifiers and Ensembles for Early-Stage Prediction of Lung Cancer," *3rd International Conference on Emerging Trends in Engineering, Sciences and Technology IEEE*, 2018.
- [14] K. & N. N. Pradeep, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4 5 and Naive Bayes Algorithms for healthcare analytics," *Procedia computer science*, vol. vol. 132, pp. 412-420, 2018.
- [15] Samaya Madhavan, M. Tim Jones, "Deep Learning Architecture," IBM Developer, 7 9 2017. [Online]. Available: <https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/>.
- [16] S. & A. B. Senthil, "Lung cancer prediction using feed forward back propagation neural networks with optimal features," *International Journal of Applied Engineering Research*, Vols. vol. 13, no. 1, pp. 318-325, 2018.
- [17] S. H. Y. & H. N. Kido, "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)," *International Workshop on Advanced Image Technology*, pp. 1-4, 2018.
- [18] A. Sherstinsky, ""Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network Physica D: Nonlinear Phenomena," 2020.
- [19] S. S. M. M. S. A. S. S. S. S. Javaid Ahmad Wani, "Machine Learning and Deep Learning Based Computational Techniques in Automatic Agricultural Diseases Detection: Methodologies Applications, and Challenges. Archives of Computational Methods in Engineering," 2021.
- [20] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali and W. Cao, "A Survey on Theories and Applications for Self-Driving Cars Based on Deep Learning Methods," *Applied Sciences*, 2020.
- [21] W. V. R. Victor Takashi Hayashi, ""Hand-Free authentication for Virtual Assistants with Trusted IoT Device and Machine Learning," 2022.
- [22] D. T. A. C. S. C. John E. Ball, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, 2017.
- [23] D. H. R. M. D. D. D. a. A. G. P. R. Kshirsagar, "A Review on Application of Deep Learning in Natural Language Processing," *5th International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 1834-1840, 2022.
- [24] C.-H. Y.-S. S. C.-J. H. K.-C. L. a. P.-H. H. Chou, "Design of Desktop Audiovisual

- Entertainment System with Deep Learning and Haptic Sensations," *Symmetry* 12, 2020.
- [25] J. W. S. X. J. Y. C. Z. Q. Z. S. H. X. L. A. Xiaoxiao Ma, "A Comprehensive Survey on Graph Anomaly Detection with Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [26] R. L. M. K. D. a. J. A. Siegel, "Cancer statistics," *CA: A Cancer Journal for Clinicians*,, 2020.
- [27] K. C. P. & R. S. Pradhan, "A deep learning-based approach for detection of lung cancer using self adaptive sea lion optimization algorithm (SA-SL_{NO}).," *J Ambient Intell Human Comput*, pp. 12933-12942, 2022.
- [28] C. C. a. P. P. D. Patel, "Predicting Mutation Status and Recurrence Free Survival in Non-Small Cell Lung Cancer: A Hierarchical ct Radiomics – Deep Learning Approach," *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.
- [29] D. a. K. V. Palani, "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification," *Journal of medical systems* 43, pp. 1-12, 2019.
- [30] W. Abdul, "An Automatic Lung Cancer Detection and Classification (ALCDC) System Using Convolutional Neural Network," *13th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 443-446, 2020.
- [31] ". C. D. D. U. M. L. A. I. C. o. I. S. S. (. S. Mukherjee and S. U. Bohra, "Lung Cancer Disease Diagnosis Using Machine Learning Approach," *3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 207-211, 2020.
- [32] T. G. a. N. P. S. R. Jena, "Texture Analysis Based Feature Extraction and Classification of Lung Cancer," *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1-5, 2019.
- [33] A. T. W. A. S. N. K. Y. H. A. O. & N. K. Haga, "Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis," *Radiological physics and technology*, Vols. vol. 11, no. 1, pp. 27-35, 2018.
- [34] H. L. H. S. E. M. G. X. X. J. R. L. T. & H. C. Cao, "'Multi-Branch Ensemble Learning Architecture Based on 3D CNN for False Positive Reduction in Lung Nodule Detection," *IEEE Access* , vol. Vol. 7, pp. 67380-67391, 2019.
- [35] Ö. G. M. & Ö. Günaydin, "'Comparison of Lung Cancer Detection Algorithms," *Electrical-Electronics & Biomedical Engineering and Computer Science*, pp. 1-4, 2019.

- [36] L. W. Y. Y. L. L. & W. B. Li, "A New Strategy to Detect Lung Cancer on CT Images," *International Conference on Image, Vision and Computing*, pp. 716-722, 2018.
- [37] P. a. G. S. Lobo, "Classification and Segmentation Techniques for Detection of Lung Cancer from CT Images," *International Conference on Inventive Research in Computing Applications*, pp. 1014-1019, 2018.
- [38] C. K. v. R. S. S. A. G. P. J. C. S. E. S.-P. C. W. M. M. A. P. U. P. M. v. G. B. Ciompi F, "Corrigendum: Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Sci Rep. 2017 Apr 19;7:46479 - PubMed*, 2017.
- [39] S. a. Yasar, "Lung cancer detection system using lung CT image processing," *International Conference on Computing, Communication, Control and Automation*, pp. 1-5, 2017.
- [40] M. A. L. & G. K. Prabukumar, "An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier," *Journal of Ambient Intelligence and Humanized Computing*, Vols. vol. 10, no. 1, pp. 267-293, 2019.
- [41] F. V. A. S. M. & M. M. Firouzabadi, "An optimized framework for cancer prediction using immunosignature," *Journal of medical signals and sensors*, vol. 8, 2018.
- [42] A. B. F. J. C. A. B. J. B. R. v. B. V. a. F. H. Lynch CM, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *Int. J. Med. Inf. 108*, pp. 1-8, 2017.
- [43] K. V. K. & K. K. Senthil Kumar, "Lung cancer detection using image segmentation by means of various evolutionary Algorithms," *Computational and mathematical methods in medicine*, 2019.
- [44] X. S. J. & V. Huang, "Lung nodule detection in CT using 3D convolutional neural networks," *International Symposium on Biomedical Imaging*, pp. 379-383, 2017.
- [45] B. H. R. & T. Narayanan, "Analysis of various classification techniques for computer aided detection system of pulmonary nodules in CT," *National Aerospace and Electronics Conference and Ohio Innovation Summit*, pp. 88-93, 2016.
- [46] S. & E. Baboo, "A classification and analysis of pulmonary nodules in CT images using random forest," *International Conference on Inventive Systems and Contro*, pp. 1226-1232, 2018.
- [47] M. B. S. K. Z. & K. Faisal, "An Evaluation of Machine Learning Classifiers and Ensembles for Early-Stage Prediction of Lung Cancer," *3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST) IEEE*, pp. 1-4, 2018.

- [48] K. & N. Yamunadevi, "An Optimized Classification of Human Cancer Disease for Gene Expression Data," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, pp. 8-15, 2018.
- [49] K. S. C. P. & T. R. Pradhan, "HRDEL: High ranking deep ensemble learning-based lung cancer diagnosis model. Expert Systems with Applications,," *Journal of Ambient Intelligence and Humanized Computing*, vol. 213, p. 118956, 2023.
- [50] M. A. Y. W. A. A. B. a. M. A. M. Nuruzzaman Faruqui, "LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data," *Computers in Biology and Medicine*, 2021.
- [51] X. Y. B. C. J. Z. M. L. H. Z. M. H. Y. C. Y. Z. B. X. C. H. R. L. K. a. C. B. Mingxiang Feng, "Detection of circulating genetically abnormal cells using 4-color fluorescence in situ hybridization for the early detection of lung cancer," *Journal of Cancer Research and Clinical Oncology*, vol. 147, p. 2397–2405, 2021.
- [52] A. N. a. D. R. Edla, "Lung Nodule Classification on Computed Tomography Images Using Deep Learning," *Wireless Personal Communications*, vol. 116, pp. 655–690,, 2021.
- [53] W. W. a. G. Charkborty, "Automatic prognosis of lung cancer using heterogeneous deep learning models for nodule detection and eliciting its morphological features," *Applied Intelligence*, vol. 51, p. 2471–2484, 2021.
- [54] J. W. a. Z. L. Yawen Xiao, "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data," *Computers in Biology and Medicine*, vol. 135, p. 104540, 2021.
- [55] S. S. a. S. Mohan, "ROI-based feature learning for efficient true positive prediction using convolutional neural network for lung cancer diagnosis," *Neural Computing and Applications*, vol. 32, p. pp. 15989–16009, 2020.
- [56] N. Z. A. M. R. N. H. M. & U. A. Zainal, "Glowworm Swarm Optimization (GSO) Algorithm for Optimization Problems: A State-of-the-Art Review," *Applied Mechanics and Materials*, vol. 421, pp. 507-511, 2013.
- [57] L. v. d. Maaten, "Accelerating t-SNE using Tree-Based Algorithms," *Journal of Machine Learning Research*, pp. 1-21, 2014.
- [58] K. V. K. & K. K. Senthil Kumar, "Lung cancer detection using image segmentation by means of various evolutionary Algorithms," *Computational and mathematical methods in medicine*, pp. 1-18, 2019.
- [59] M. d. C. L. C. A. J. M. Cazzolato, "Oliveira, LFM, dos Santos Kaster, D, JuniBREATH: Heat Maps Assisting the Detection of Abnormal Lung Regions in CT Scans," *International Symposium on Computer-Based Medical Systems*, pp. 248-253, 2017.

- [60] Y. Z. L. C. H. & Y. N. Li, "Lung Nodule Detection With Deep Learning in 3D Thoracic MR Images," *IEEE Access*, vol. 7, no. 7, pp. 37822-37832, 2019.
- [61] R. & S. Manickavasagam, "GACM based segmentation method for Lung nodule detection and classification of stages using CT image," *International Conference on Innovations in Information and Communication Technology*, pp. 1-5, 2019.
- [62] A. & R. R. Sharma, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," *Computer methods and programs in biomedicine*, vol. 178, pp. 219-235, 2019.
- [63] L. T. G. Z. L. W. G. L. Z. Y. J. & C. Q. Yu, "Prediction of pathologic stage in non-small cell lung cancer using machine learning Algorithm based on CT image feature analysis," *BMC cancer*, vol. 19, p. 464, 2019.
- [64] J. L. C. R. S. M. T. M. S. A. M. & L. H. Wu, "Treatment outcome prediction for cancer patients based on radiomics and belief function theory," *IEEE Transactions on Radiation and Plasma Medical Sciences*, Vols. vol. 3, no. 2, pp. 216-224, 2018.
- [65] D. S. E. T. M. A. Tsoukalas, G. Tsilimidos, D. Calina and A. Tsatsakis, "Metabolic Fingerprint of Chronic Obstructive Lung Diseases: A New Diagnostic Perspective," *Metabolites*, 9, p. 290, 2019.
- [66] B. Z. W. P. J. H. D. J. Y. & Z. Z. He, "A biomarker basing on radiomics for the prediction of overall survival in non-small cell lung cancer patients," *Respiratory research*, Vols. vol. 19, no. , pp. 1-8, 2018.
- [67] W. J. M. D. J. C. P. & Y. F. 2. Sun, "Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis," *Radiation Oncology*, Vols. vol. 13, no. 1, pp. 1-8, 2018.
- [68] R. & R. K. Tekade, "Lung Cancer Detection and Classification Using Deep Learning," *International Conference on Computing Communication Control and Automation*, pp. 1-5, 2018.
- [69] S. S. J. M. A. & F. S. A. Raoof, "Lung Cancer prediction using machine learning: A comprehensive approach. In 2020 2nd International conference on innovative mechanisms for industry applications," *IEEE*, pp. 108-115, 2020.
- [71] R. S. S. R. J. M. A. a. G. C. Shah, "Results of surgical treatment of stage i and ii lung cancer," *The Journal of cardiovascular surgery*, pp. 169-172, 1996.
- [72] J. C. P. J. J. B. W. G. L. R. J. A. a. M. C. F. Nesbitt, "Survival in early-stage non-small cell lung cancer," *The Annals of thoracic surgery*, 1995.
- [73] S. S. a. F. K. M. Sakamoto, "Flat-types of Electrodes to Estimate Lung Cancer Position Using Resonant Circuit Technology," *IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, pp. 61-63, 2020.

- [74] A. & R. R. Sharma, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," *Computer methods and programs in biomedicine* , vol. vol. 178, pp. 219-235, 2019.
- [75] F. N. a. I. K. O. Abiezer., "Identification of Lung Cancer in Smoker Person Using Ensemble Methods Based on Gene Expression Data," *2022 5th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 89-93, 2022.
- [76] S. & R. N. Shanthi, "Non-small-cell lung cancer prediction using radiomic features and machine learning methods," *International Journal of Computers and Applications*, pp. 1-9, 2019.
- [77] M. S. P. & Z. Rahman, "A New Method for Lung Nodule Detection Using Deep Neural Networks for CT Images," *International Conference on Electrical, Computer and Communication Engineering*, pp. 1-6, 2019.
- [78] W. Z. F. L. Z. & W. L. Zuo, "Multi-Resolution CNN and Knowledge Transfer for Candidate Classification in Lung Nodule Detection," *IEEE Access* , vol. vol. 7, pp. 32510-32521, 2019.
- [79] J. W. J. W. Y. L. H. N. T. P. J. & Q. D. Wang, "Pulmonary Nodule Detection in Volumetric Chest CT Scans Using CNNs-Based Nodule-Size-Adaptive Detection and Classification," *IEEE Access*, vol. vol. 7, pp. 46033-46044 , 2019.
- [80] L. T. G. Z. L. W. G. L. Z. Y. J. & C. Q. “. Yu, "Prediction of pathologic stage in non-small cell lung cancer using machine learning Algorithm based on CT image feature analysis," *BMC cancer* , Vols. vol. 19, no. 1, p. 464, 2019.
- [81] M. & Y. A. Saritas, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, Vols. vol. 7, no. 2, pp. 88-91, 2019.
- [82] N. S. M. & R. El-Askary, "Feature Extraction and Analysis for Lung Nodule Classification using Random Forest," *8th International Conference on Software and Information Engineering*, pp. 248-252, 2019.
- [83] C. H. J. J. Y. & G. F. Zhao, "Lung Nodule Detection via 3D U-Net and Contextual Convolutional Neural Network," *International Conference on Networking and Network Applications*, pp. 356-361, 2018.
- [84] R. & S. Manickavasagam, "GACM based segmentation method for Lung nodule detection and classification of stages using CT image," *International Conference on Innovations in Information and Communication Technology*, pp. 1-5, 2019.
- [85] H. K. D. & X. X. Tang, "Automated pulmonary nodule detection using 3D deep convolutional neural networks," *International Symposium on Biomedical Imaging*, pp. 523-526, 2018.

- [86] W. L. C. F. W. & X. X. Zhu, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," *Winter Conference on Applications of Computer Vision*, pp. 673-681, 2018.
- [87] D. P. K. H. G. & S. D. Thomas, "Lung Consolidation Detection through Analysis of Vocal Resonance Signals," *International Engineering in Medicine and Biology Society*, pp. 957-960., 2018.
- [88] S. S. A. a. M. W. K. Fotouhi, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *Journal of biomedical informatics 90* , 2019.
- [89] J. A. S. & A. Alam, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier," *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering*, pp. 1-4, 2018.
- [90] M. & B. P. Mukherjee, "Segmentation of lungs nodules by iterative thresholding method and classification with Reduced Features," *International Conference on Inventive Communication and Computational Technologies*, pp. 450-455, 2018.
- [91] A. C. S. B. N. Z. S. & R. Rao, "Improved detection of lung fluid with standardized acoustic stimulation of the chest," *IEEE journal of translational engineering in health and medicine*, vol. Vol. 6, pp. 1-7, 2018.
- [92] B. N. N. E.-N. S. E. Y. A. N. M. T. & A. R. A. El-Sherbiny, "Blb-brain/lung cancer detection and segmentation and breast dense calculation," *International Workshop on Deep and Representation Learning*, pp. 41-47, 2018.
- [93] H. K. H. J. S. Y. K. & L. J. Chung, "Automatic lung segmentation with juxta-pleural nodule identification using active contour model and bayesian approach," *IEEE journal of translational engineering in health and medicine*, vol. Vol. 6, pp. 1-13, 2018.
- [94] H. K. H. J. S. Y. K. & L. J. Chung, "Automatic lung segmentation with juxta-pleural nodule identification using active contour model and bayesian approach," *IEEE journal of translational engineering in health and medicine*, vol. Vol. 6, pp. 1-13, 2018.
- [95] S. & V. T. Perumal, "Lung cancer detection and classification on CT scan images using enhanced artificial bee colony optimization," *International Journal of Engineering & Technology*, Vols. vol. 7, no. 2.26, pp. 74-79, 2018.
- [96] J. & L. S. Lyu, "Using multi-level convolutional neural network for classification of lung nodules on CT images," *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 686-689, 2018.
- [97] D. C. E. A. G. S. G. & M. NarainPonraj, "Analysis of LBP and LOOP Based Textural Feature Extraction for the Classification of CT Lung Images," *International Conference on Devices, Circuits and Systems*, pp. 309-312, 2018.

- [98] S. H. Y. & H. N. Kido, "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)," *International Workshop on Advanced Image Technology*, pp. 1- 4, 2018.
- [99] M. & S. Paing, "Improved Random Forest (RF) Classifier for Imbalanced Classification of Lung Nodules," *International Conference on Engineering, Applied Sciences, and Technology*, pp. 1-4, 2018.
- [100] S. D. J. S. A. d. P. A. N. R. & M. Netto, "Modified Quality Threshold Clustering for Temporal Analysis and Classification of Lung Lesions," *IEEE Transactions on Image Processing*, Vols. vol. 28, no. 4, pp. 1813-1823, 2018.
- [101] S. R. R. & A. Potghan, "Multi-Layer Perceptron Based Lung Tumor Classification," *International Conference on Electronics, Communication and Aerospace Technology*, pp. 499-502, 2018.
- [102] K. & N. N. Pradeep, "Lung cancer survivability prediction based on performance using classification techniques of support vector machines, C4". 5 and Naive Bayes Algorithms for healthcare analytics.," *Procedia computer science*, , vol. vol. 132, 2018.

List Of Publications

1. Pradhan, Kanchan; Chawla, Priyanka (2020). *Medical Internet of Things using machine learning algorithms for lung cancer detection. Journal of Management Analytics*, (), 1–33. doi:10.1080/23270012.2020.1811789.
2. Kanchan S. Pradhan; Priyanka Chawla, Sanyog Rawat (2022). *A deep learning- ---based approach for detection of lung cancer using self-adaptive sea lion optimization algorithm (SA-SL_nO)* Journal of Ambient Intelligence and Humanized Computing <https://doi.org/10.1007/s12652-022-04118-y>
3. Kanchan S. Pradhan; Priyanka Chawla (2019). *Lung Cancer Prediction Using Machine Learning Algorithm. Think India Journal*. ISSN: 0971-1260 Vol -22- NO-30 Special Issue--December-2019
4. Kanchan Pradhan; Priyanka Chawla; Lung Cancer Detection using CT Images, *International Conference on Advances in Signal Processing, VLSI, Communication and Embedded System (ICSVCE),2022*.
5. Kanchan Pradhan; Priyanka Chawla; Lung Cancer Detection using Deep LearningAlgorithm, *International Conference on Advances in Signal Processing, VLSI, Communication and Embedded System (ICSVCE),2022*.
6. Kanchan Pradhan; Priyanka Chawla; Rajiv Tiwari; HRDEL: High Ranking Deep Ensemble Learning based Lung Cancer Diagnosis Model, *Expert Systems with Applications,2023*.