VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgavi-590 018, Karnataka, India



### An Internship Report On

Lung Cancer Prediction

Submitted in Partial Fulfillment of the requirement for the award of the degree of

**BACHELOR OF ENGINEERING**

IN

#### COMPUTER SCIENCE AND ENGINEERING

**Submitted By**

**Nagashree C R 1SJ20CS092**

#### Carried out at

INFLOW TECHNOLOGIES

Bangalore, Karnataka Under the guidance of

Internal Guide External Guide

#### Ashok K N Paras

**Assistant Professor Master Trainee**

#### Dept. Of CSE, SJCIT Inflow Technologies Pvt Ltd



**S J C INSTITUTE OF TECHNOLOGY DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

CHIKKABALLAPUR-562101 2023-2024

#### ||Jai Sri Gurudev||

**Sri Adichunchanagiri Shikshana Trust®**

**S.J.C INSTITUTE OF TECHNOLOGY, Chickballapur - 562101 Department of Computer Science and Engineering**

**CERTIFICATE**

This is to certify that the Internship work entitled **“LUNG CANCER PREDICTION”** carried out by **NAGASHREE C R** bearing **1SJ20CS092** a bonafide student of Sri Jagadguru Chandrashekaranatha Institute of Technology in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of **Visvesvaraya Technological University, Belgaum** during the year **2023-24**. It is certificated that all corrections / suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The Internship report has been approved as it satisfies the academic requirements in respect of Internship work prescribed for the said Degree.

Signature of Coordinator **Prof. Swetha T** Assistant Professor Dept. of CSE,SJCIT

Signature of Guide

**Prof. ASHOK K N**

Assistant Professor Dept. of CSE,SJCIT

Signature of HOD

**Dr. Manjunath Kumar B H** Professor & HOD, Dept. of CSE,SJCIT

Signature of Principal **Dr. G T Raju** Principal, SJCIT, Chickballapur



#### External Examiners:

**Name of the Examiners Signature with Date**

1.

2.

# COMPANY CERTIFICATE



**DECLARATION**

I, **NAGASHREE C R** student of VIII semester B.E in Computer science & Engineering at S J C Institute of Technology, Chickballapur, hereby declare that the Internship work entitled” Lung Cancer Prediction” has been independently carried out by me under the supervision of **Ashok K N** Assistant Professor and the coordinator **Swetha T** Assistant Professor, submitted in partial fulfillment of the course requirement for the award of degree in **Bachelor of Engineering** in **Computer Science & Engineering** of **Visveswaraya Technological University, Belgavi** during the year 2023-2024. I further declare that the report has not been submitted to any other University for the award of any other degree.

#### PLACE:CHIKKABALAPUR NAGASHREE C R

**Date: USN: 1SJ20CS92**

# ABSTRACT

Lung cancer is a potentially lethal illness. Cancer detection continues to be a challenge for medical professionals. The true cause of cancer and its complete treatment have still not been discovered. Cancer that is caught early enough can be treated. Image processing methods such as noise reduction, feature extraction, identification of damaged regions, and maybe a comparison with data on the medical history of lung cancer are used to locate portions of the lung that have been impacted by cancer. This research shows an accurate classification and prediction of lung cancer using technology that is enabled by machine learning and image processing. To begin, photos need to be gathered. In the experimental investigation, 83 CT scans from 70 distinct patients were utilized as the dataset. The geometric mean filter is used during picture preprocessing. As a consequence, image quality is enhanced. The *K*-means technique is then used to segment the images. The part of the image may be found using this segmentation. Then, classification methods using machine learning are used. For the classification, ANN, KNN, and RF are some of the machine learning techniques that were used. It is found that the ANN model is producing more accurate results for predicting lung cancer.

# ACKNOWLEDGEMENT

With reverential pranam, we express my sincere gratitude and salutations to the feet of his holiness **Byravaikya Padmabhushana Sri Sri Sri Dr. Balagangadharanatha Maha Swamiji,** & his holiness **Jagadguru Sri Sri Sri Dr. Nirmalanandanatha Swamiji** of Sri Adichunchanagiri Mutt for their unlimited blessings. First and foremost we wish to express my deep sincere feelings of gratitude to our institution, **Sri Jagadguru Chandrashekaranatha Swamiji Institute of Technology.** For providing me an opportunities for completing my internship work successfully.

I extend deep sense of sincere gratitude to **Dr. G T Raju**, **Principal, S J C Institute of Technology, Chickballapur**, for providing an opportunity to complete the Internship Work.I extend special in-depth, heartfelt, sincere gratitude to our HOD **Dr. Manjunatha Kumar B H, Professor and Head of the Department, Computer Science and Engineering, S J C Institute of Technology, Chickballapur,** for her constant support and valuable guidance of the Internship Work.

I convey our sincere thanks to Internship Internal Guide **Ashok K N, Assistant Professor, Department of Computer Science and Engineering, S J C Institute of Technology,** for his/her constant support, valuable guidance and suggestions of the Internship Work.

I am thankful to Internship External Guide **Paras, Master Trainee,Inflow Technologies Pvt Ltd, Bangalore** for providing valuable guidance and encouragement of the Internship Work.

I also feel immense pleasure to express deep and profound gratitude to our Internship Coordinator **Prof. Swetha T, Assistant Professor, Department of Computer Science and Engineering, S J C Institute of Technology,** for his guidance and suggestions of the Internship Work.

Finally, I would like to thank all faculty members of Department of Computer Science and Engineering, S J C Institute of Technology, Chickballapur for their support.

I also thank all those who extended their support and co-operation while bringing out this Internship Report.

**NAGASHREE C R (1SJ20CS92)**

|  |  |  |
| --- | --- | --- |
| **Chapter No** | **Chapter Title** | **Page No** |
| **1** | **COMPANY PROFILE** | **1-5** |

|  |  |  |
| --- | --- | --- |
| **CONTENTS**  Declaration | | i |
| Abstract | | ii |
| Acknowledgement | | iii |
| Contents | | iv |
| List of Figures | | vii |
| List of Tables | | viii |
|  | |  |
| 1.1 History of the Organization | | 1 |
| 1.1.1 Objectives | | 1 |
| 1.1.2 Operations of the Organization | | 2 |
| 1.2 Major Milestones | | 3 |
| 1.3 Structure of the Organization | | 4 |
| 1.4 Services Offered | | 5 |
| **2** | **ABOUT THE DEPARTMENT** | **6-10** |
| 2.1 Specific Functionalities of the Department | | 6 |
| 2.2 Process Adopted | | 6 |
| 2.3 Testing | | 7 |
| 2.4 Structure of the Department | | 8 |
| 2.5 Roles and Responsibilities of Individuals | | 10 |
| **3** | **TASK PERFORMED** | **11-15** |
|  | |  |
| **4** | **REFLECTION NOTES** | **16-30** |
| 4.1 Experience | | 16 |
| 4.2 Technical Outcomes | | 16 |
| 4.2.1 System Requirement Specification | | 17 |
| 4.3 System Analysis and Design | | 17 |

* + 1. [Existing System 17](#_TOC_250011)
    2. [Disadvantages of the Existing System 18](#_TOC_250010)
    3. [Proposed System 19](#_TOC_250009)
    4. [Advantages of the Proposed System 20](#_TOC_250008)
  1. [System Architecture 21](#_TOC_250007)
     1. [Data Flow Diagram 22](#_TOC_250006)
     2. UML Diagrams 23
        1. [USE CASE Diagram 23](#_TOC_250005)
     3. [Sequence Diagram 24](#_TOC_250004)
     4. [Activity Diagram 25](#_TOC_250003)
  2. [Implementation 25](#_TOC_250002)
     1. [Modules 26](#_TOC_250001)
  3. [Screen Shots 27](#_TOC_250000)

|  |  |  |
| --- | --- | --- |
| **5** | **CONCLUSION** | 31 |

**BIBLIOGRAPHY**

# LIST OF FIGURE

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Name of the Figure** | **Page No.** |
| Figure 2.2 | Process adopted | 7 |
| Figure 2.3 | Department Structure | 8 |
| Figure 4.4 | System Architecture Diagram | 21 |
| Figure 4.4.1 | Data Flow Diagram | 22 |
| Figure 4.4.2.1 | Use Case Diagram | 23 |
| Figure 4.4.5 | Sequence Diagram | 24 |
| Figure 4.4.6 | Activity Diagram of the System | 25 |
| Figure 4.6.1 | Home Page | 27 |
| Figure 4.6.2 | About the page | 27 |
| Figure 4.6.3 | Detect Lung Cancer Page | 28 |
| Figure 4.6.4 | Invalid image type | 28 |
| Figure 4.6.5 | Image is not infected with Cancer | 29 |
| Figure 4.6.6 | Image is infected with Cancer | 29 |

**CHAPTER - 1**

# COMPANY PROFILE

## History of the Organization

Inflow Technologies Pvt Ltd is a technology solutions company based in Bangalore, India. The company was founded in 2005 by Byju Pillai, who has a background in the IT industry. In the early years, Inflow Technologies focused on providing networking solutions to businesses in India. As the company grew, it expanded its offerings to include storage and cyber security solutions.

In 2010, Inflow Technologies formed a partnership with Palo Alto Networks, a leading cyber security company. This partnership helped Inflow Technologies to expand its offerings in the cyber security space.

Over the years, Inflow Technologies has formed partnerships with several other leading technology companies, including Cisco, Dell EMC, Symantec, and many others. These partnerships have helped the company to offer a wide range of technology solutions to its clients.

Today, Inflow Technologies is one of the leading technology solutions providers in India, with a team of over 300 employees and a presence in multiple cities across the country. The company continues to focus on providing innovative technology solutions to businesses of all sizes in various industries.

Inflow Technologies is a leading global Technology Distribution Company offering a wide range of networking, infrastructure, cyber security, unified communications, storage, and hyper-converged infrastructure solutions. It was founded in 2005 in Bangalore, India. Over the years, Inflow Technologies has expanded its presence across the Asia-Pacific region, with offices in countries such as India, Singapore, Thailand, Malaysia, Vietnam, Philippines, Indonesia, and Australia. The company has established partnerships with renowned technology vendors and serves a diverse customer base, including enterprises, government organizations, and small to medium-sized businesses. Inflow Technologies continues to evolve and adapt to the changing technological landscape, providing innovative solutions and excellent service to its clients.

#### Objectives

The primary objective of Inflow Technologies Pvt Ltd is to provide innovative and effective technology solutions to businesses in various industries.

To achieve this, the company has set the following objectives:

* + - * Provide comprehensive technology solutions: Inflow Technologies aims to provide end-to-end technology solutions to its clients, from consulting and implementation to support and maintenance. The company's goal is to ensure that its clients have the technology infrastructure they need to succeed in their business.
      * Partner with leading technology companies: Inflow Technologies has formed partnerships with several leading technology companies to ensure that it can offer the latest and most effective technology solutions to its clients.
      * Provide exceptional customer service: Inflow Technologies believes that exceptional customer service is crucial to its success. The company aims to provide prompt and effective support to its clients, ensuring that their technology systems are always running smoothly.
      * Stay ahead of the curve: The technology industry is constantly evolving, and Inflow Technologies aims to stay ahead of the curve by investing in research and development and keeping up with the latest trends and innovations.
      * Foster a culture of innovation and collaboration: Inflow Technologies encourages its employees to think outside the box and come up with innovative solutions to its clients' technology challenges. The company also fosters a culture of collaboration, encouraging its employees to work together to deliver the best possible outcomes for its clients.

#### Operation of the Organization

Inflow Technologies operates as a technology solutions provider, offering a wide range of services to businesses in various industries.

The company's operations can be divided into the following categories:

* + - * Consulting: Inflow Technologies offers consulting services to businesses looking to improve their technology infrastructure. The company's consultants work closely with clients to understand their specific needs and challenges, and develop custom technology solutions to meet those needs.
      * Implementation: Inflow Technologies provides implementation services to ensure that its clients' technology solutions are properly

installed and configured. This includes setting up hardware and software systems, integrating different technologies, and ensuring that everything is working together seamlessly.

* Support: Inflow Technologies offers ongoing support services to ensure that its clients' technology systems are always up and running. This includes monitoring systems for issues, providing remote assistance, and offering on-site support as needed.
* Training: Inflow Technologies provides training services to help clients get the most out of their technology solutions. This includes training employees on how to use new technology systems, as well as providing ongoing education and training to keep up with the latest trends and innovations
* Sales: Inflow Technologies sells a range of hardware and software products from leading technology companies. The company's sales team works closely with clients to understand their specific needs and recommend the best products to meet those needs

## Major Milestones

Inflow Technologies Pvt Ltd has achieved several major milestones since its founding in 2005. Some of the key milestones include:

* Formation of strategic partnerships: Inflow Technologies has formed strategic partnerships with several leading technology companies, including Cisco, Palo Alto Networks, Symantec, and Dell EMC. These partnerships have helped the company to expand its offerings and provide a wider range of technology solutions to its clients.
* Expansion into new markets: Inflow Technologies has expanded its operations to multiple cities across India, including Bangalore, Mumbai, Chennai, and Hyderabad. The company has also expanded its presence in the Asia-Pacific region, with offices in Singapore, Thailand, and the Philippines.
* Recognition and awards: Inflow Technologies has received several awards and recognitions for its innovative technology solutions and exceptional customer service. Some of the awards include the CRN Excellence Award for Best IT Infrastructure Provider, the Palo Alto

Networks Partner of the Year Award, and the Cisco Emerging Partner of the Year Award.

* Growth and revenue: Inflow Technologies has experienced significant growth and revenue over the years. In 2020, the company reported a revenue of INR 1,200 crores ($160 million USD), a 25% increase from the previous year
* Innovation and research: Inflow Technologies has invested heavily in research and development to stay ahead of the curve in the rapidly evolving technology industry. The company has developed several innovative solutions, including its own cyber security platform, which has received widespread recognition and acclaim.

Overall, these milestones demonstrate Inflow Technologies' commitment to providing innovative technology solutions and exceptional customer service, as well as its continued growth and success in the industry.

## Structure of the Organization

Inflow Technologies Pvt Ltd has a hierarchical organizational structure, with various departments and teams working together to deliver technology solutions to clients.

The company's organizational structure can be broken down into the following levels:

* Board of Directors: The Board of Directors is responsible for setting the company's strategic direction, overseeing its operations, and ensuring that it is meeting its objectives.
* Executive Leadership Team: The Executive Leadership Team includes the CEO, COO, CFO, and other top executives. They are responsible for implementing the company's strategy and managing its day-to-day operations.
* Sales and Marketing Team: The Sales and Marketing Team is responsible for identifying and pursuing new business opportunities, promoting the company's products and services, and developing relationships with clients.
* Consulting and Professional Services Team: The Consulting and Professional Services Team includes technology consultants, solution architects, and project managers. They work closely with clients to

understand their needs and develop custom technology solutions to meet those needs.

* Implementation and Support Team: The Implementation and Support Team is responsible for ensuring that clients' technology solutions are properly installed and configured, and that they receive ongoing support to keep their systems running smoothly.
* Research and Development Team: The Research and Development Team is responsible for developing innovative technology solutions, conducting research into emerging trends and technologies, and staying ahead of the curve in the industry.

## Services Offered

Inflow Technologies Pvt Ltd offers a wide range of technology solutions and services to its clients. Some of the key services offered by the company include:

* Infrastructure Solutions: Inflow Technologies provides a range of infrastructure solutions, including servers, storage, networking, and virtualization.
* Security Solutions: The company offers a comprehensive range of cyber security solutions, including endpoint protection, network security, threat intelligence, and security analytics.
* Cloud Solutions: Inflow Technologies helps clients to migrate their systems and applications to the cloud, and offers cloud-based solutions such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).
* Data Management Solutions: The company provides data management solutions, including backup and recovery, data protection, and data analytics.
* Professional Services: Inflow Technologies offers consulting, design, implementation, and support services for its technology solutions.
* Training and Certification: The company provides training and certification programs for its technology solutions to help clients develop the skills and knowledge needed to effectively use these solutions.

# CHAPTER – 2

**ABOUT THE DEPARTMENT**

## Specific Functionalities of the Department

Our department of tech support majorly focused on manage, maintain and repair IT systems. unior Software Developers are entry-level software developers that assist the development team with all aspects of software design and coding. The special functionalities include

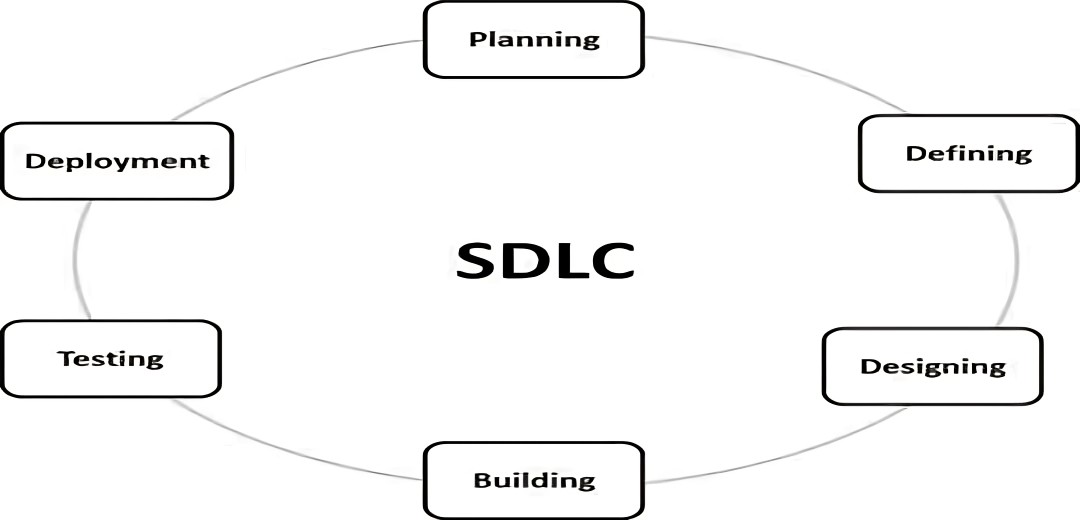
* + - Understanding the work to be completed.
    - Planning the assigned activities in more detail if needed.
    - Completing assigned work within the budget, timeline and quality expectations.
    - Informing the project manager of issues, scope changes, risk and quality concerns.
    - Proactively communicating status and managing expectation.

## Process Adopted

The department aims to first understand the user requirements. Further on, a basic structure of the product that needs to be built is drawn and understood. Eventually, the technologies that would best help in developing the product are understood. If the product has database requirements, the schema and the database design are worked upon. The department believes in "Think before you code"- the requirements and logics are first understood over a paper and then are moved to a code form. Agile processes generally promote a disciplined project management process that encourages frequent inspection and adaptation, a leadership philosophy that encourages teamwork, self-organization and accountability, a set of engineering best practices intended to allow for rapid delivery of high-quality software, and a business approach that aligns development with customer needs and company goals. Agile development refers to any development process that is aligned with the concepts of the Agile Manifesto.

Process as followed mentioned steps:

* + - Planning
    - Defining
    - Designing
    - Building
    - Testing
    - Deployment



## Testing

Figure 2.2: Process adopted

The various testing techniques used by the department can be summarized as follows:

1. Functionality Testing of a Website: it is a process that includes several testing parameters like user interface, APIs, database testing, security testing, client and server testing and basic website functionalities. Functional testing is very convenient and it allows users to perform both manual and automated testing. It is performed to test the functionalities of each feature on the website.
2. Usability Testing: This type of testing includes testing the site navigations and contents of the website.
3. Interface Testing: Three areas to be tested here are Application, Web and Data base Server.
4. Database Testing: Database is one critical component of your web application and stress must be laid to test it thoroughly Testing activities will include Test if any errors are shown while executing queries, Data Integrity is maintained while creating, updating or deleting data in database, Check response time of queries and fine tune

them if necessary, Test data retrieved from your database is shown accurately in your web application.

1. Compatibility testing: Compatibility tests ensures that your web application displays correctly across different devices. This would include-Browser Compatibility Test: Same website in different browsers will display differently. You need to test if your web application is being displayed correctly across browsers, JavaScript, AJAX and authentication is working fine.
2. Pipeline testing: After compatibility testing it is the time to test all the micro- services in pipeline together to check their compatibility and message passing. Thus all the services/functionalities are kept in pipeline and tested together. Afterwards whole pipeline is pushed in the deployment server.

## Structure of the Department

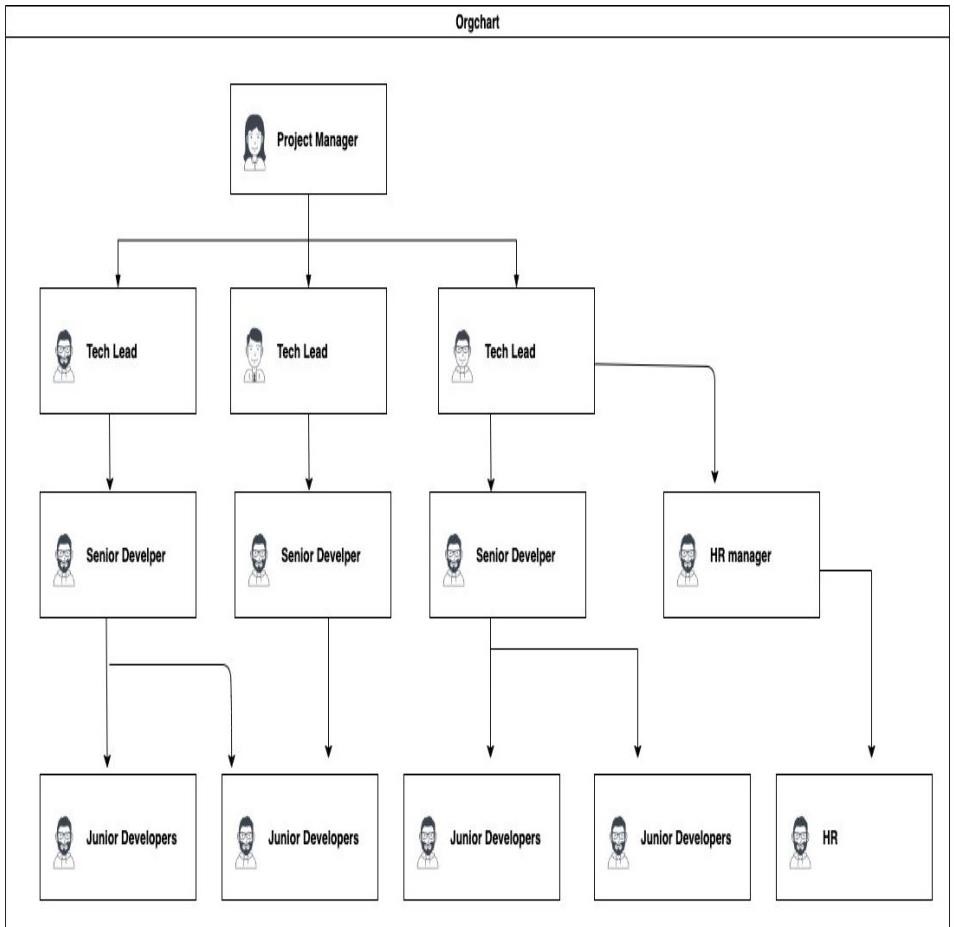
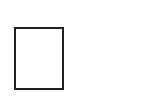
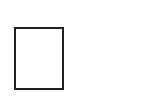
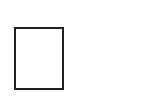


Figure 2.2 Department Structure

Any organization will have a specific structure to function as a whole. The hierarchy of an organization is as shown above. There are multiple levels in an organization hierarchy starting from high level to low level.

The project manager will be the top level person responsible for delivering projects intime.

Tech Leads work at the next level to project managers they provide technical assistanceto the below level peers and check the work at regular intervals.

Senior Developers and Junior Developers work at next two consecutive levels

## Roles and Responsibilities of Individuals

The different roles and responsibilities of individuals are:

1. Project Manager : Project Managers plays the lead role in planning, executing. Monitorizing, controlling, and closing projects.
2. Tech Leads: Technical Lead as the name states is solely responsible for leading a development team. The is not easy. They have to lead a team. Technical Lead is the one who actually creates a technical vision in order to turn it into reality with the help of the team.
3. HR Manager: The Human Resource Manager will lead and direct the routine functions of the Human Resources (HR) department including hiring and interviewing staff, administering pay, benefits, and leave, and enforcing company policies and practices.
4. Senior Developer. Develops software solutions by studying information needs, conferring with users, studying systems flow, data usage, and work processes; investigating problem areas; and following the software development lifecycle. A senior developer may manage a team of developers and will be expected to encourage creativity and efficiency throughout complex digital projects. Due to the pressurised nature of the role, a robust and organised approach to the work is needed to produce the best solutions.
5. Junior Developer: Junior Software Developers are entry-level software developers that assist the development team with all aspects of software design and coding.

# CHAPTER – 3

**TASK PERFORMED**

In this Internship Machine Learning with Python using MI it was a course of making predictions using ml algorithms. In this project author is evaluating performance of various classification/prediction algorithms such as SVM, Naïve Bayes, and Logistic Regression etc to predict lung cancer disease. All this algorithms are good in prediction but accuracy is not good enough. To get better prediction accuracy author is combining two classification algorithms such as Linear Model and Random Forest to build new algorithm called Hybrid Machine Learning to get better prediction accuracy of heart dataset. Hybrid algorithm will form up by using Voting classifier, Internally Voting classifier will build up using Linear Model and Random Forest and while classification voting algorithm will evaluate prediction accuracy of both algorithms and vote for that algorithm which gives better accuracy. So by using hybrid model always we will have better prediction accuracy algorithm which helps in better prediction of lung cancer disease.

It is difficult to identify lung cancer because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) .The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the lung or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation. Training Program The internship is a platform where the trainees are assigned with the specific task. In the initial days of the internship, I was trained on the following: o Python Programming o Machine Learning Algorithms Modules are divided as:

1. Data Collection
2. Data Preprocessing
3. Splitting of Dataset
4. Performance of the Algorithm
   * **Data Collection**: The collecting of the project's dataset is the initial step. The Dataset in this case was obtained from the Kaggle.This section describes, in brief, the data that has been used for the research. Data from traffic was used in this project, the major amount of data was extracted from public website Kaggle (Kaggle.com), data regarding the review and linked was obtained from a leading traffic where many pedestrians will be present. Data from various sources was integrated together to form a staging data-set. For predicting the pedestrian whether he is present or not which is helpful in overcoming the accidents as the data is already predicted.
   * **Data Preprocessing**: Pre-processing data is the second phase. If any missing or null values are found throughout this operation, the entire row has been erased. Gathering original data to be used with a machine learning algorithm is known as data pre- processing.
   * **Splitting of Dataset:** Following pre-processing, the data is divided into training and testing sets according to the specifications. The information has indeed been divided into two sections in this stage. Such as a data collection for training and validation
   * **Performance of the Algorithms:** Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), and Random Forests (RF) are all popular machine learning algorithms with distinct characteristics and applications.

### Some factors which the performance depends is as follows

1. Computational Efficiency:

* ANN: Training ANNs can be computationally intensive, especially for deep architectures with many layers and parameters. The backpropagation algorithm, used for updating weights during training, involves multiple iterations over the entire dataset, which can be time-consuming.
* KNN: KNN requires no training phase, as it stores all training instances in memory. However, predicting the class of a new instance involves computing distances to all stored instances, which can be slow for large datasets. Additionally, the algorithm's memory usage grows linearly with the number of training instances.
* RF: Training RF involves constructing multiple decision trees, each of which can be trained independently. While this parallelization improves training speed compared to single decision trees, RF may still be slower than simpler models like logistic regression, especially for very large datasets due to the need to process multiple subsets of the data.

1. Scalability:

* ANN: ANNs can scale well to large datasets and high-dimensional feature spaces, thanks to their capacity to model complex relationships and learn from diverse data sources. However, training deep neural networks on extremely large datasets may require distributed computing resources and specialized hardware accelerators.
* KNN: KNN's performance can degrade in high-dimensional spaces due to the curse of dimensionality, where the concept of distance loses meaning as the number of dimensions increases. As a result, KNN may not scale well to datasets with a large number of features, requiring dimensionality reduction techniques like principal component analysis (PCA) or feature selection.
* RF:RF is generally scalable to large datasets, as the training time of individual decision trees is typically not affected by the size of the dataset. However, RF may become computationally expensive for very large datasets due to memory requirements and the need to build and maintain multiple trees in memory.

1. Interpretability:

* ANN: Interpretability of ANNs can be challenging, especially for deep architectures with many layers and parameters. While techniques like activation maximization and gradient-based attribution methods can provide insights into model behavior, understanding the contribution of individual features to predictions may be difficult.
* KNN:KNN is relatively easy to interpret, as predictions are based on the majority class of nearest neighbors in feature space. However, understanding why specific instances are classified as they are may require examining the distribution and characteristics of neighboring instances, which can be cumbersome for high- dimensional data.
* RF:RF models are less interpretable than simpler models like logistic regression, as predictions are based on an ensemble of decision trees rather than a single interpretable model. While techniques like feature importance scores can provide insights into variable importance, understanding complex interactions among features may be challenging.

1. Generalization Ability:

* ANN: ANNs have the potential to generalize well to unseen data, especially when trained on diverse and representative datasets. However, over fitting is a common concern, particularly for deep architectures with many parameters. Regularization techniques like dropout and weight decay are often used to improve generalization performance.
* KNN:KNN's performance heavily depends on the choice of distance metric and the value of k. Selecting appropriate hyper parameters through cross-validation is essential to prevent over fitting and achieve good generalization performance. Additionally, KNN may struggle with imbalanced datasets and noisy data, which can affect classification accuracy.
* RF:RF is robust against over fitting, thanks to its ensemble approach and the averaging of predictions from multiple trees. This inherent regularization helps improve generalization performance, especially on noisy or heterogeneous datasets. However, RF hyper parameters like the number of trees and tree depth need to be tuned carefully to optimize performance and prevent under fitting or over fitting.

In summary, while ANNs, KNN, and RF have different computational characteristics, scalability, interpretability, and generalization abilities, selecting the most suitable algorithm depends on the specific requirements of the problem domain, available computational resources, and desired trade-offs between performance and interpretability. Conducting thorough performance analysis, including experimentation with different algorithms and hyper parameters, is crucial to identify the most effective approach for a given task.

**Problem Statement:**

There is ample related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field. The function of back propagation multilayer perception (MLP) of ANN is used to predict lung cancer disease. The obtained results are compared with the results of existing models within the same domain and found to be improved

## Technology used:

* + Goggle Colab /Jupiter Notebook
  + Python Programming Language
  + Different Python Libraries

CHAPTER – 4

# REFLECTION NOTES

## Experience

The internship has been a really useful experience for me that I can learn a lot of new knowledge that will definitely be useful for my future study. I'm grateful that my assignments have a lot of variety instead of just focusing on a specific area. This allows me to be able to learn more and also challenge myself to overcome many different kinds of difficulties encountered during my internship. Having many assignments also required me to manage my work time efficiently prioritizing the urgent task Some tasks require me to do research with less available online documentation other task requires me to make attempts on works that I have never experienced before just by learning from documentations. Although the task may be difficult and overwhelming sometimes, I'm really excited to push my skills to the limit and carry out those task assigned to me. Beside technical skills, I also observed and learned a lot of soft skills from my supervisors and my co-workers such as professional communication and team work. I have also learned a lot from my supervisor who's always willing to help me when I face difficulties and also willing to share a lot of his knowledge and wisdom to me from his post experience. My internship experience has definitely improved my hard skills in IT and sharpen my soft skills a lot more than I expected I have shaped a better mind set in me and motivated me to keep on exploring and challenging myself in the world of information technology.

## Technical Outcomes

* + - Learning the basics of Al and its sub-domains Machine Learning.
    - Understand a wide variety of learning algorithms.
    - Understand how to evaluate models generated from data.
    - Apply, the algorithms to real problem.
    - Optimize the models learned and report on the expectancy accuracy that can be achievedby applying the models.

## System Requirement Specification

#### Hardware Requirements:

* + - * A computer with a multi-core processor (minimum 4 cores) and at least 8GB of RAM.
      * A dedicated GPU is not necessary but can significantly speed up training time for large datasets.
      * Sufficient storage space to store the dataset, preprocessed features, and the trained model.

#### Software Requirements:

* + - * + An operating system such as Linux, macOS or Windows
        + Python 3.x installed along with necessary libraries such as NumPy, Pandas, Scikit-learn, and Matplotlib.- Jupyter Notebook or similar environment for interactive coding and experimentation.

## System Analysis and Design

#### Existing System

One existing problem in lung cancer prediction is the challenge of early detection. Lung cancer is often diagnosed at advanced stages when treatment options are limited and prognosis is poor. This is due in part to the lack of specific symptoms in the early stages of the disease, leading to delayed diagnosis.Additionally, there are limitations in the current screening methods for lung cancer, such as chest X-rays and computed tomography (CT) scans, which may not be accessible to all populations or may have limitations in sensitivity and specificity, leading to false positives or false negatives.Furthermore, there is a need for more accurate and personalized risk assessment models for lung cancer. While certain risk factors such as smoking history are well-established, there are other factors, including genetic predisposition and environmental exposures, that may also contribute to an individual's risk of developing lung cancer. Integrating these factors into predictive models could improve their accuracy and reliability. Data availability and quality also pose challenges in developing robust predictive models for lung cancer. Access to comprehensive and diverse datasets, including clinical, imaging, genetic, and demographic data, is essential for training and validating predictive algorithms. However, there may be limitations in the availability and standardization of such data, which can impact the performance of predictive models. Moreover, there is a need for effective implementation and integration of predictive models

into clinical practice. This involves addressing issues related to workflow integration, clinician acceptance, and patient engagement to ensure that predictive tools are effectively utilized for early detection and personalized management of lung cancer.Addressing these challenges requires interdisciplinary collaboration between clinicians, researchers, data scientists, and healthcare stakeholders to develop innovative solutions for improving lung cancer prediction and ultimately reducing the burden of this disease.

#### Disadvantages of the Existing System

1. Limited Sensitivity: Current prediction methods, such as chest X-rays and CT scans, may have limited sensitivity in detecting early-stage lung cancer, leading to missed diagnoses, particularly in asymptomatic individuals or those with small or indolent tumors.
2. False Positives: Screening tests for lung cancer often result in false-positive findings, leading to unnecessary follow-up tests, invasive procedures, and patient anxiety. False positives can also strain healthcare resources and increase healthcare costs.
3. Radiation Exposure: Imaging-based screening methods like CT scans expose patients to ionizing radiation, which carries inherent risks, especially with repeated screenings over time. This radiation exposure can increase the risk of secondary cancers and other health complications, particularly in individuals with long-term screening protocols.
4. Overdiagnosis and Overtreatment: Overdiagnosis occurs when screening detects indolent or slow-growing cancers that may never progress to cause symptoms or harm during a person's lifetime. This can lead to unnecessary treatments, such as surgery, chemotherapy, or radiation therapy, which carry their own risks and burdens for patients.
5. Limited Accessibility: Access to lung cancer screening and predictive tools may be limited due to factors such as geographical location, socioeconomic status, healthcare disparities, and insurance coverage. This can result in unequal distribution of healthcare resources and disparities in cancer outcomes.
6. Inaccuracy in Risk Assessment: Current risk assessment models for lung cancer may not accurately account for all relevant risk factors, including genetic predisposition, environmental exposures, and comorbidities. This can lead to underestimation or

overestimation of an individual's risk, affecting the effectiveness of screening and prevention strategies.

1. Lack of Personalization: Existing prediction methods may not fully account for individual variations in lung cancer risk and progression. Personalized approaches, tailored to an individual's unique characteristics and biomarkers, are needed to improve the accuracy and effectiveness of prediction and prevention efforts.
2. Data Limitations: The development and validation of predictive models rely on the availability of high-quality, comprehensive datasets. However, there may be limitations in the quantity, quality, and diversity of available data, which can impact the performance and generalizability of predictive algorithms.

#### Proposed System

The proposed system for lung cancer prediction aims to overcome the limitations of existing methods by leveraging advanced data analytics and machine learning techniques to enhance accuracy, sensitivity, and personalization. The system integrates diverse data sources, including clinical, imaging, genetic, and demographic data, to develop robust predictive models that can identify individuals at high risk of developing lung cancer. Utilizing state-of- the-art machine learning algorithms, such as deep learning and ensemble methods, the system can extract meaningful patterns and relationships from complex datasets to improve risk assessment and early detection. Furthermore, the proposed system incorporates personalized risk factors, such as smoking history, genetic predisposition, environmental exposures, and comorbidities, to tailor prediction models to individual characteristics and optimize screening strategies. Through continuous learning and adaptation, the system can refine predictive algorithms over time, incorporating new data and insights to enhance performance and reliability. Moreover, the proposed system prioritizes accessibility and equity by ensuring that predictive tools are widely available and easily accessible to diverse populations, regardless of geographical location or socioeconomic status. By harnessing the power of data-driven analytics and machine learning, the proposed system has the potential to revolutionize lung cancer prediction, enabling early intervention, personalized care, and improved outcomes for patients.

#### Advantages of the Proposed System

The proposed system for lung cancer prediction offers several advantages over existing methods:

1. Improved Accuracy: By leveraging advanced data analytics and machine learning techniques, the proposed system can achieve higher accuracy in predicting lung cancer risk compared to traditional screening methods. It can identify subtle patterns and relationships in diverse datasets, leading to more precise risk assessment and early detection.
2. Enhanced Sensitivity: The system's ability to integrate multiple data sources, including clinical, imaging, genetic, and demographic data, allows for improved sensitivity in detecting early-stage lung cancer. This enables timely intervention and treatment when the disease is more manageable and potentially curable.
3. Personalized Risk Assessment: Incorporating personalized risk factors such as smoking history, genetic predisposition, and environmental exposures enables the system to tailor prediction models to individual characteristics. This personalized approach enhances the accuracy of risk assessment and allows for targeted screening and prevention strategies.
4. Continuous Learning and Adaptation: The system's capability for continuous learning and adaptation enables it to refine predictive algorithms over time. By incorporating new data and insights, the system can continuously improve its performance and reliability, keeping pace with evolving research and clinical practices.
5. Accessible and Equitable: The proposed system prioritizes accessibility and equity by ensuring that predictive tools are widely available and easily accessible to diverse populations. This helps address healthcare disparities and ensures that all individuals, regardless of geographical location or socioeconomic status, have access to advanced lung cancer prediction technologies.
6. Reduction in False Positives and Overdiagnosis: By improving the accuracy and specificity of lung cancer prediction, the proposed system can reduce the incidence of false positives and overdiagnosis associated with traditional screening methods. This minimizes unnecessary

follow-up tests, invasive procedures, and patient anxiety, leading to more efficient use of healthcare resources and improved patient outcomes.

1. Empowering Healthcare Professionals: The system provides healthcare professionals with valuable insights and decision support tools to facilitate informed clinical decision-making. By assisting clinicians in identifying individuals at high risk of lung cancer, the system enables proactive interventions and personalized care delivery, ultimately improving patient outcomes.

## System Architecture

The architecture of the proposed system has the following components:

* + - Training data to train the models.
    - Testing data to apply the models.
    - Data storage- stores data.
    - Classification and prediction algorithms.
    - Forecast engine.

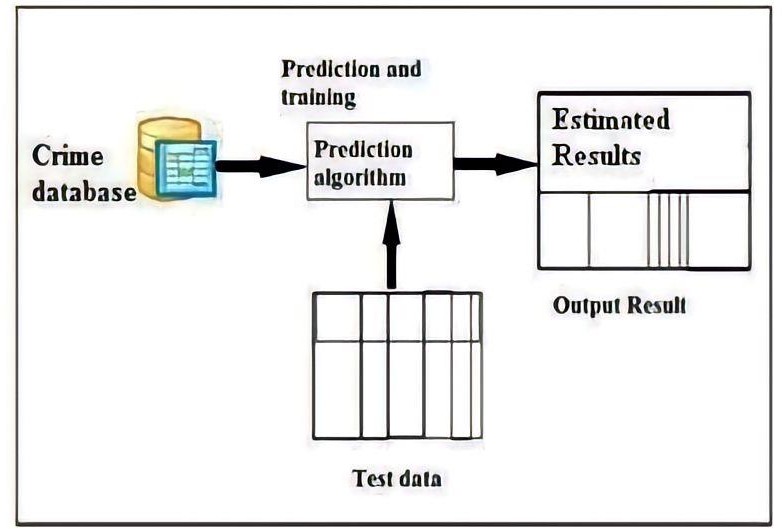


Figure 4.4 :System Architecture Diagram

#### Data Flow Diagram

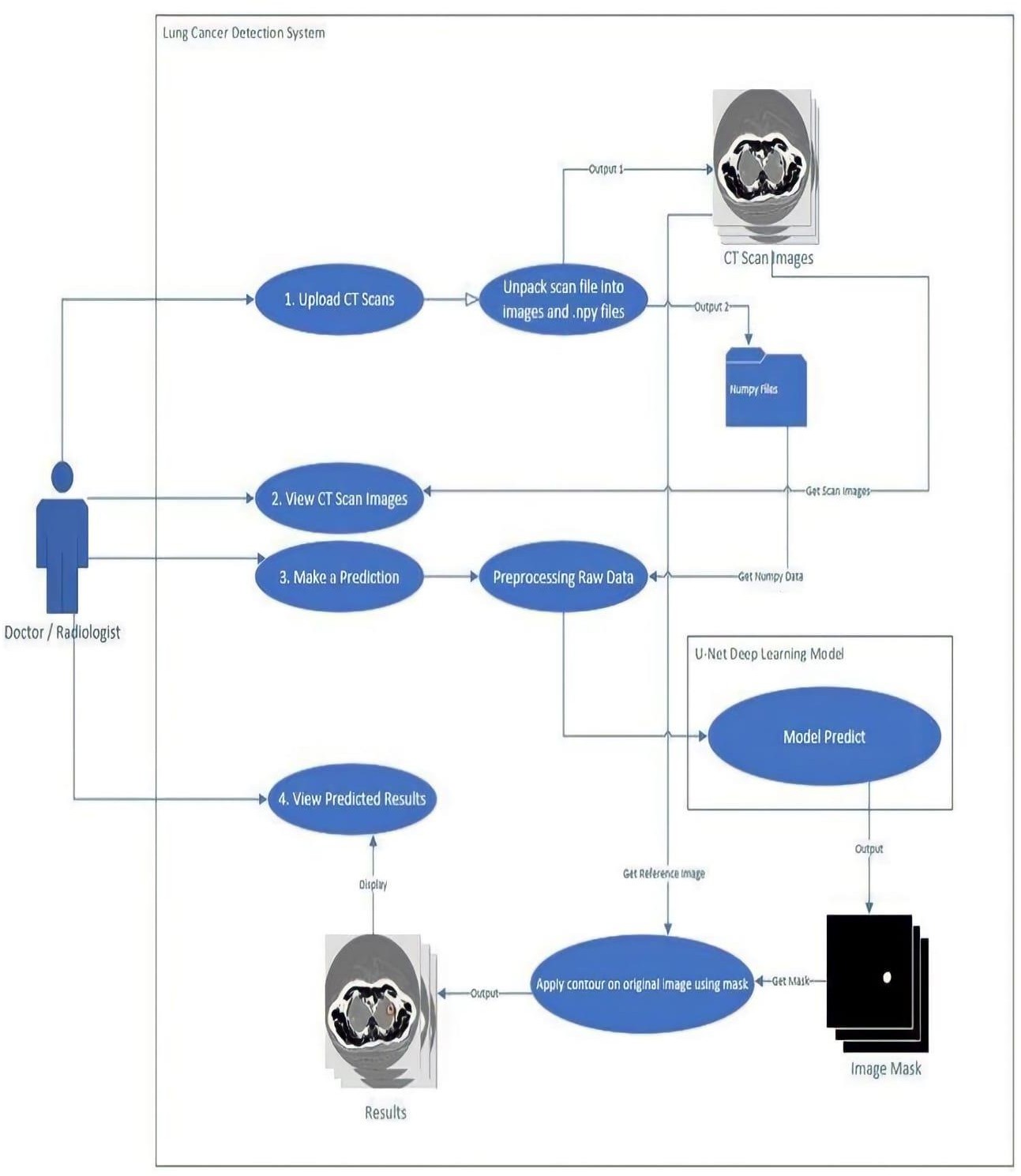


Figure 4.4.1 : Data Flow Diagram

Also known as DFD, Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation. Data flow diagrams can be divided into logical and physical. The logical data flow diagram describes flow of data through a system to perform certain functionality of a business. The physical data flow diagram describes the implementation of the logical data flow.

## UML Diagram

## Use Case Diagram

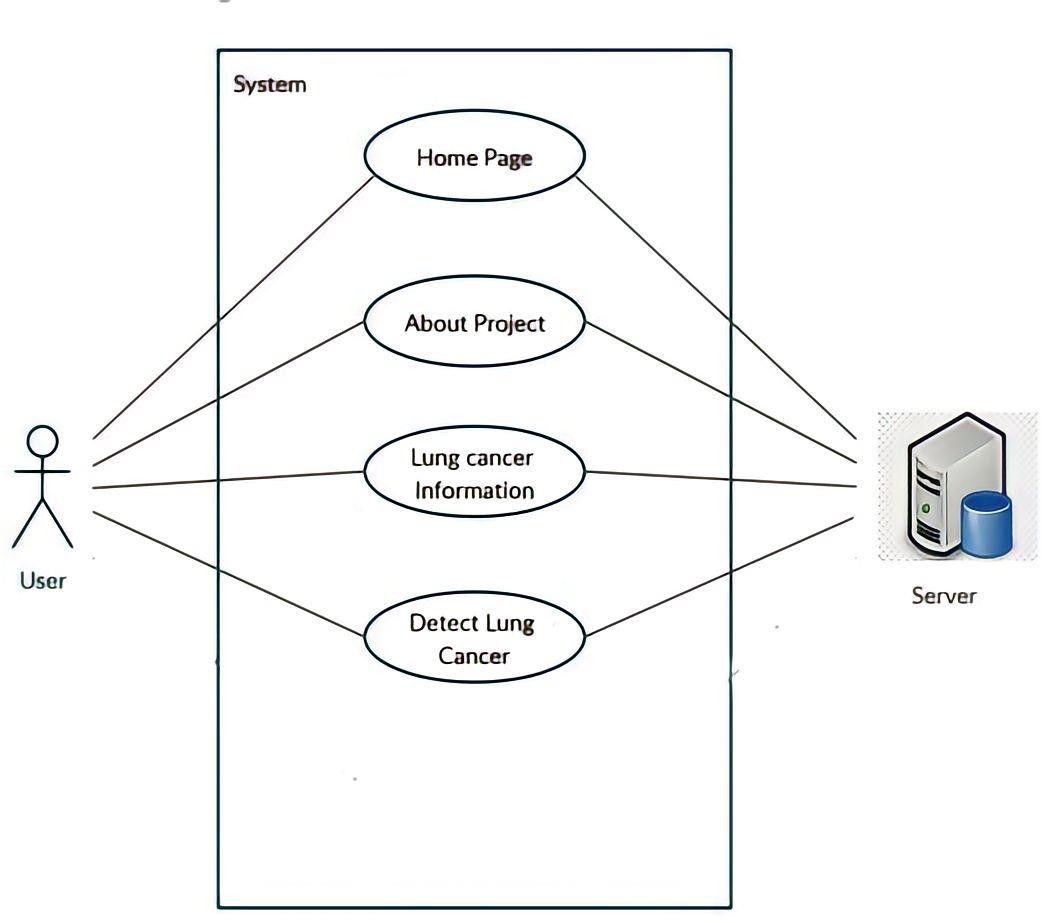


Figure 4.4.2.1 Use Case Diagram

In the Unified Modelling Language (UML), a use case diagram can summarize the details of your system's users (also known as actors) and their interactions with the system. To build one, you'll use a set of specialized symbols and connectors. An effective use case diagram can help your team discuss and represent, Scenarios in which your system or application interacts with people. organizations, or external systems, Goals fy your system or application helps those entities (known as actors) achieve, The scope of your system

## Sequence Diagram

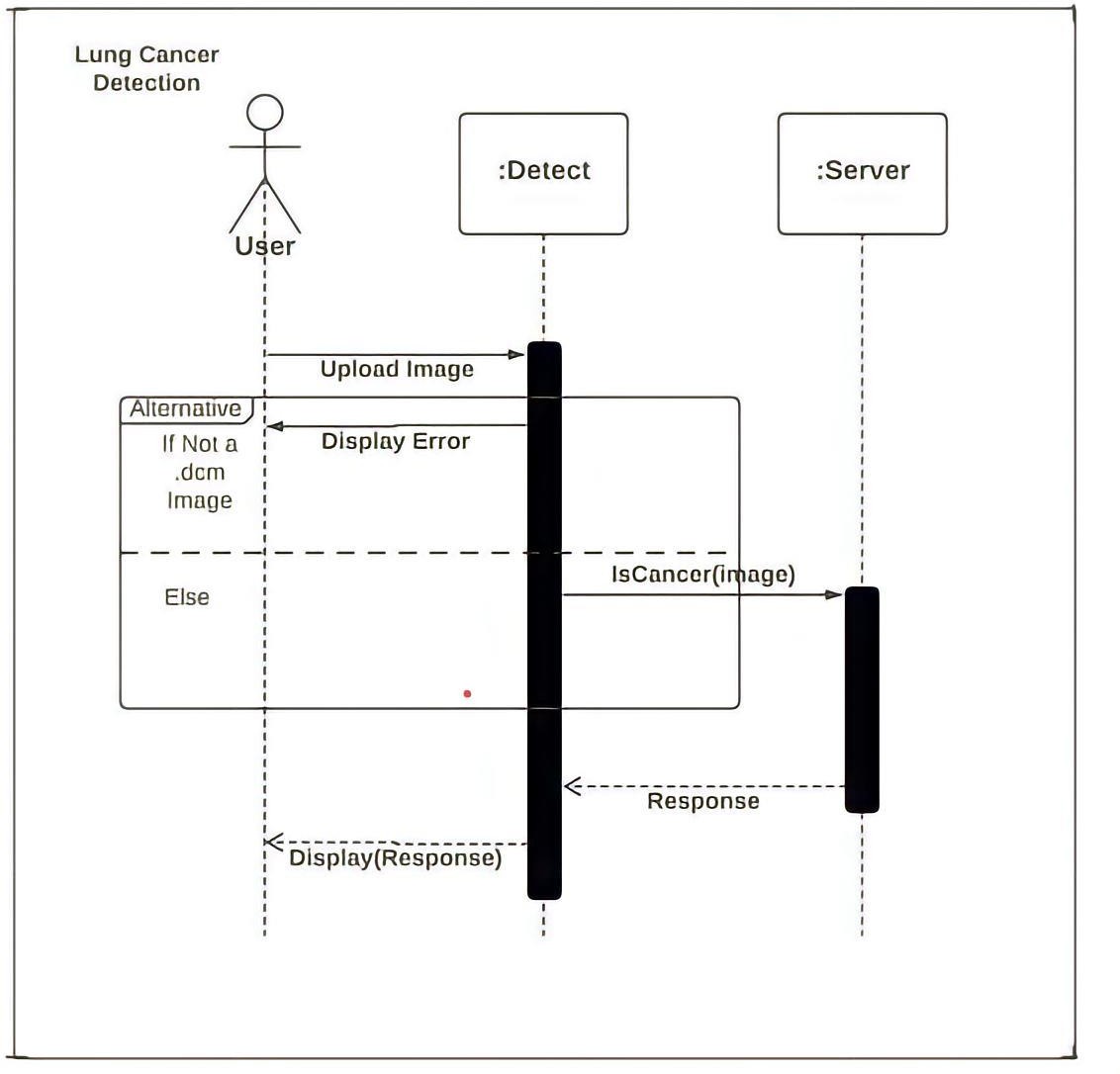


Figure 4.4.5 Sequence Diagram of Lung Cancer Detection module

Figure 4.4.5 specifies the sequence diagram of the Lung Cancer Detection module. This diagram specifies the timeline of different actors and objects and the sequence of steps undertaken to complete the functionality of the module. User will upload an image and if the image is not in .dcm format it will throw an error and asks the user to upload the correct format image. Once the user uploads the correct image it is sent to the server for processing and server processes and the image and send the response to the user whether the image is infected with cancer or not.

## Activity Diagram

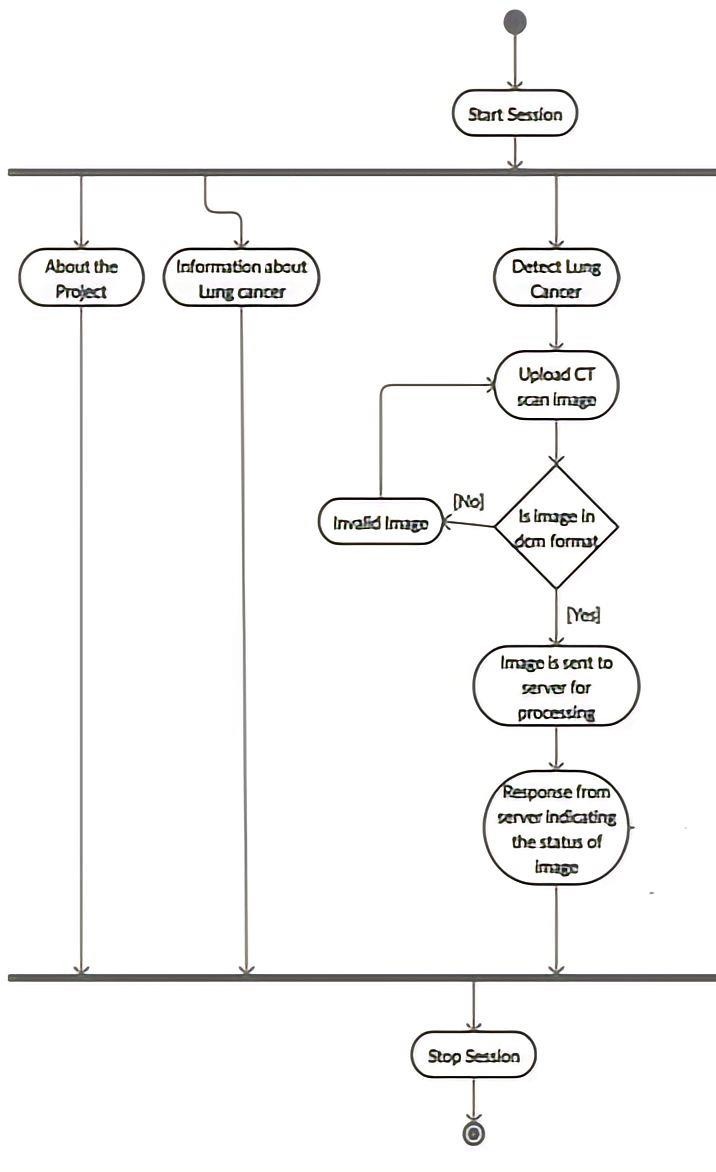


Figure 4.4.6: Activity Diagram of the System

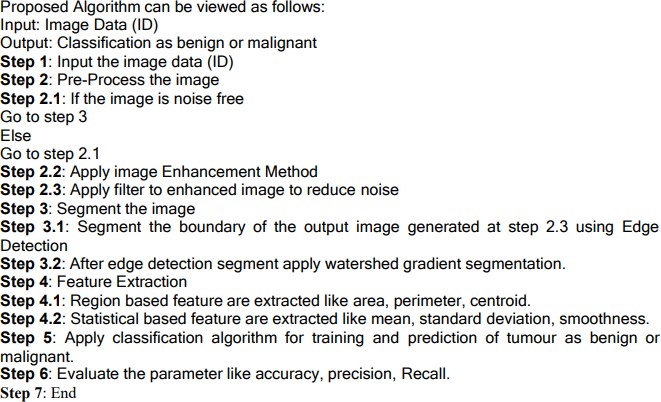
## Implementation

This section shows an accurate classification and prediction of lung cancer using technology that is enabled by machine learning and image processing. To begin, photos need to be gathered. After that, a geometric mean filter is used to perform preprocessing on the images.

This ultimately leads to an improvement in image quality. After that, the *K*-means method is used to segment the images. The identification of the region of interest is facilitated by this segmentation. After that, categorization strategies based on machine learning are

utilized. figure illustrates the classification and prediction of lung cancer utilizing technology that enables machine learning and image processing.

The following algorithm describes the step by step approach for the proposed model.



## 4.5.1 Modules

#### Pre-Processing Layer

Image has been collected from LIDC-LDRI. The original image was full of noise and for that first we have applied histogram equalization on the image to enhance the image and then on the equalized image median filter has been applied to remove the noise which was already present in the image after getting the noise free image we have applied some more noise in the image yield more clearer picture then again noise has been removed using median filter. Generally median filter is non linear digital filtering technique and it is also used as

smoothing 13 of images as it don’t blur the edges completely as compare to other filtration technique like Gaussian filter or average filter.

#### Segmentation Layer

Image segmentation is a method of partitioning the image into various parts. After pre processing the image on the pre-processed image segmentation is applied to acquire the information from the image. For image segmentation first we have applied edge detection technique through edge detection we can segment the boundary of the image for edge detection prewitt operator has been used, on that operator threshold has been applied so that after edge detection the intensity value which is less than threshold is removed and the intensity value which is higher than or equal to threshold will consider for further segmentation after getting the segmented image by edge detection we will apply watershed segmentation on the output image. Watershed segmentation takes the concept topographical landscape with ridge and valley which is defined by a gray level with respective pixel or gradient magnitude. There exist various ways to segment using watershed segmentation here we have used watershed segmentation using gradient. The gradient magnitude is used to preprocess the gray scale image; it has high pixel value along the object edge and low pixel value in another left region. And through this we can get the final segmented image through which we can extract features

#### Feature Extraction Layer

The output generated by segmentation is used for feature extraction. By doing feature extraction we have extracted two types of feature one is region based another is texture based region based we have extracted feature like area in context to image means pixel of the image, perimeter in context image mean vector containing the distance around the boundary of each region in the image, centroid means the centre of mass of the region and it is in 1 X 2 vector form, image and based on texture we have extracted feature like mean is used to find average intensity, standard deviation is used to measure average contrast, smoothness used to measure relative smoothness of the intensity in the region, entropy is used to measure randomness using statistical approach of texture based

## Screen Shots



Figure 4.6.1 Home Page

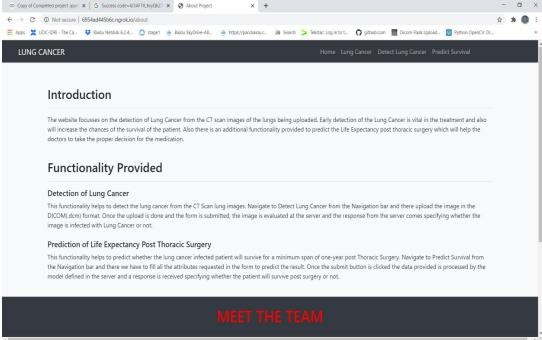


Figure 4.6.2 About the page

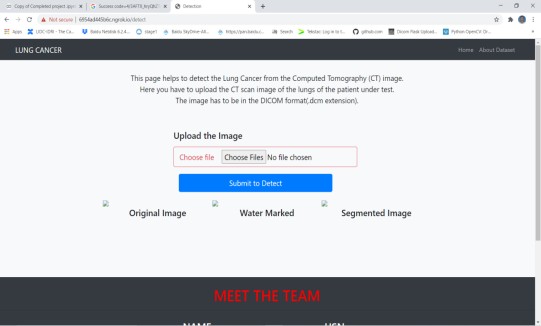


Figure 4.6.3 Detect Lung Cancer Page

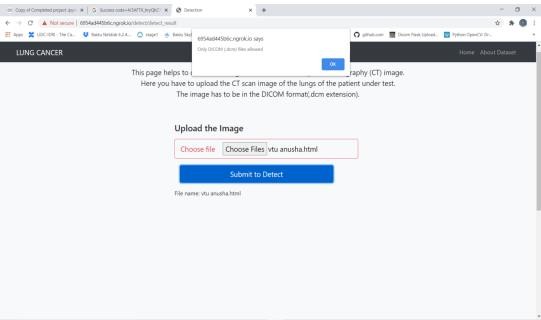


Figure 4.6.4 : Invalid image type

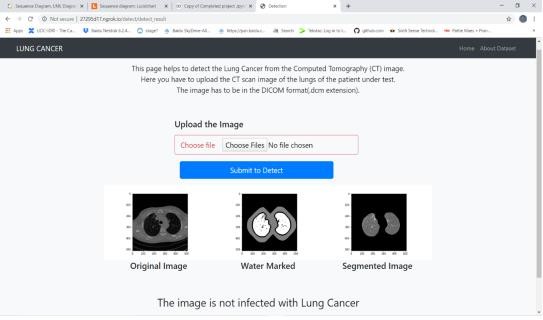


Figure 4.6.5 Image is not infected with Cancer

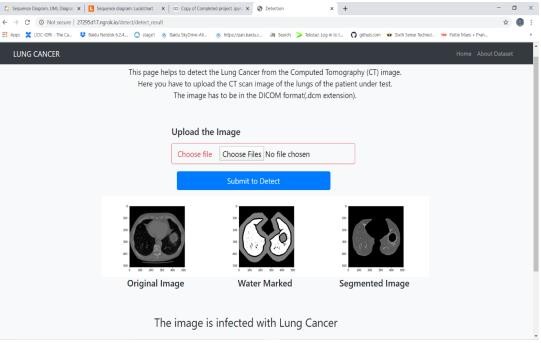


Figure 4.6.5 Image is infected with Cancer

# CHAPTER – 4

**CONCLUSION**

Lung cancer is one of the deadliest types of the disease, claiming the lives of approximately one million people each year. Given the current state of affairs in medicine, it is critical that lung nodule identification be performed on chest CT scans. As a result, the use of CAD systems is crucial for the early detection of lung cancer. Image processing is a necessary activity that is employed in a wide range of economic domains. It is used in X-ray imaging of the lungs to find areas of the body that have developed malignant growths. Image processing techniques such as noise reduction, feature extraction, identification of damaged regions, and maybe comparison with data on the medical history of lung cancer are used to locate sections of the lung that have been affected by cancer. This study demonstrates accurate lung cancer classification and prediction using technologies enabled by machine learning and image processing. To begin, photographs must be collected. Following that, the images are preprocessed using a geometric mean filter. This eventually leads to an increase in image quality. The *K*-means approach is then used to segment the images. This segmentation makes it easier to identify the region of interest. Following that, machine learning-based categorization algorithms are used. ANN predicts lung cancer with more accuracy. This research will help to increase the accuracy of lung cancer detection systems that use strong classification and prediction techniques. This study brings cutting-edge images based on machine learning techniques for implementation purposes.

1. Ahmed Medjahed S., AitSaadi T., Benyettou A., Ouali M. Kernel-based learning and feature selection analysis for cancer diagnosis. *Applied Soft Computing .* 2017;51:39–48. doi: 10.1016/j.asoc.2016.12.010. [[CrossRef](https://doi.org/10.1016%2Fj.asoc.2016.12.010)] [Google Scholar]
2. Deepa N., Prabadevi B., Maddikunta P. K., et al. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *The Journal of Supercomputing .* 2021;77(2):1998–2017. doi: 10.1007/s11227-020-03347-
3. [[CrossRef](https://doi.org/10.1007%2Fs11227-020-03347-2)] [Google Scholar]
4. Jasti V., Zamani A. S., Arumugam K., et al. Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. *Security and communication networks .* 2022;2022:7. doi: 10.1155/2022/1918379.1918379 [[CrossRef](https://doi.org/10.1155%2F2022%2F1918379)] [Google Scholar]
5. Chaudhury S., Krishna A. N., Gupta S., et al. Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer. *Computational and Mathematical Methods in Medicine .* 2022;2022:6. doi: 10.1155/2022/6841334.6841334 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9012610/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/35432588)] [[CrossRef](https://doi.org/10.1155%2F2022%2F6841334)] [Google Scholar] [Retracted](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10356254/)
6. Halder A., Kumar A. Active learning using Fuzzy-Rough Nearest Neighbor classifier for cancer prediction from microarray gene expression data. *Journal of Biomedical Informatics .* 2020;34(1):p. 2057001. doi: 10.1142/S0218001420570013. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30802546)] [[CrossRef](https://doi.org/10.1142%2FS0218001420570013)] [Google Scholar]
7. Zamani A. S., Anand L., Rane K. P., et al. Performance of machine learning and image processing in plant leaf disease detection. *Journal of Food Quality .* 2022;2022:7. doi: 10.1155/2022/1598796.1598796 [[CrossRef](https://doi.org/10.1155%2F2022%2F1598796)] [Google Scholar]
8. Sandhiya S., Kalpana Y. An artificial neural networks (ANN) based lung nodule identification and verification module. *Medico-Legal Update .* 2019;19(1):p. 193. doi: 10.5958/0974-1283.2019.00039.2. [[CrossRef](https://doi.org/10.5958%2F0974-1283.2019.00039.2)] [Google Scholar]
9. Palani D., Venkatalakshmi K. An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification. *Journal of Medical Systems .* 2019;43(2):p. 21. doi: 10.1007/s10916-018-1139-7. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30564924)] [[CrossRef](https://doi.org/10.1007%2Fs10916-018-1139-7)] [Google Scholar]
10. Ahmed Medjahed S., AitSaadi T., Benyettou A., Ouali M. Kernel-based learning and feature selection analysis for cancer diagnosis. *Applied Soft Computing .* 2017;51:39–48. doi: 10.1016/j.asoc.2016.12.010. [[CrossRef](https://doi.org/10.1016%2Fj.asoc.2016.12.010)] [Google Scholar]
11. Deepa N., Prabadevi B., Maddikunta P. K., et al. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *The Journal of Supercomputing .* 2021;77(2):1998–2017. doi: 10.1007/s11227-020-03347-
12. [[CrossRef](https://doi.org/10.1007%2Fs11227-020-03347-2)] [Google Scholar]
13. Jasti V., Zamani A. S., Arumugam K., et al. Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. *Security and communication networks .* 2022;2022:7. doi: 10.1155/2022/1918379.1918379 [[CrossRef](https://doi.org/10.1155%2F2022%2F1918379)] [Google Scholar]
14. Chaudhury S., Krishna A. N., Gupta S., et al. Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer. *Computational and Mathematical Methods in Medicine .* 2022;2022:6. doi: 10.1155/2022/6841334.6841334 [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9012610/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/35432588)] [[CrossRef](https://doi.org/10.1155%2F2022%2F6841334)] [Google Scholar] [Retracted](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10356254/)
15. Halder A., Kumar A. Active learning using Fuzzy-Rough Nearest Neighbor classifier for cancer prediction from microarray gene expression data. *Journal of Biomedical Informatics .* 2020;34(1):p. 2057001. doi: 10.1142/S0218001420570013. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30802546)] [[CrossRef](https://doi.org/10.1142%2FS0218001420570013)] [Google Scholar]
16. Zamani A. S., Anand L., Rane K. P., et al. Performance of machine learning and image processing in plant leaf disease detection. *Journal of Food Quality .* 2022;2022:7. doi: 10.1155/2022/1598796.1598796 [[CrossRef](https://doi.org/10.1155%2F2022%2F1598796)] [Google Scholar]
17. Sandhiya S., Kalpana Y. An artificial neural networks (ANN) based lung nodule identification and verification module. *Medico-Legal Update .* 2019;19(1):p. 193. doi: 10.5958/0974-1283.2019.00039.2. [[CrossRef](https://doi.org/10.5958%2F0974-1283.2019.00039.2)] [Google Scholar]
18. Palani D., Venkatalakshmi K. An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification. *Journal of Medical Systems .* 2019;43(2):p. 21. doi: 10.1007/s10916-018-1139-7. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/30564924)] [[CrossRef](https://doi.org/10.1007%2Fs10916-018-1139-7)] [Google Scholar]