Thesis Report

Development of Regression-Based Mathematical Models for Early Detection of Lung and Breast Cancer Using Clinical Parameters

SUBMITTED BY

NUSRAT JAHAN

Roll No: BKH2011025F

Session: 2019-2020

Year 04, Term 02

Department of Information and Communication Engineering, NSTU

THESIS SUPERVISOR

Dr. Md. Ashikur Rahman Khan

Professor

Department of Information and Communication Engineering, NSTU

DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING

NOAKHLAI SCIENCE AND TECHNOLOGY UNIVERSITY

Development of Regression-Based Mathematical Models for Early Detection of Lung and Breast Cancer Using Clinical Parameters

NUSRAT JAHAN

This thesis report submitted to the Department of Information and Communication Engineering, Noakhali Science and Technology University, in partial fulfillment of the requirements for the degree of B.Sc. (Engineering) in Information and Communication Engineering.

DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING

NOAKHLAI SCIENCE AND TECHNOLOGY UNIVERSITY

JULY 2025

DECLARATION

This thesis report is submitted to the Department of Information and Communication Engineering (ICE), Noakhali Science and Technology University, Noakhali- 3814, in partial fulfillment of the requirements for having the B.Sc. degree in ICE under the course entitled with "ICE- 4218". So, I, here by declare that this thesis report has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

Nusrat Jahan

Roll No: BKH2011025F

Session: 2019-2020

Year 04, Term 02

ACCEPTANCE

This thesis report is submitted to the Department of Information and Communication Engineering (ICE), Noakhali Science and Technology University, Noakhali- 3814, in partial fulfillment of the requirements for having the B.Sc. degree in ICE under the course entitled “ICE- 4218”.

Dr. Md. Ashikur Rahman Khan

Professor

Department of Information and Communication Engineering

Noakhali Science & Technology University

Sonapur, Noakhali

ABSTRACT

Cancer diagnosis at an early stage is vital in enhancing better outcomes of treatment and prevention of fatalities, particularly in low- and middle-income economies, such as that of Bangladesh, where modernized diagnostic methods are less available. The proposed research aims to construct regression models to detect lung and breast cancer in their early stages using clinically and efficiently accessible diagnostic parameters. A study of 1,000 patients with lung cancer and 570 patients with breast cancer was conducted. The most important predictors were identified through correlation matrix analysis to perform the feature selection. The Response Surface Methodology (RSM) establishes both single-order and polynomial regression models to capture both linear and nonlinear correlations between variables. ANOVA was used to test the statistical significance and the predictive power of the models. In the case of lung cancer, factors like history of smoking, genetic risk, and chronic lung disease were identified as 12 significant predictors. In the case of breast cancer, characteristics such as radius, perimeter, concavity, and compactness turned out to be influential. Separate datasets were used to verify the empirical validity and strength of the models. The findings indicate that the second-order polynomial models performed better than the linear models because they were able to capture the complex interaction of the variables. This study demonstrates that modeling based on regressions offers a cost-efficient, interpretable, and scalable approach to early cancer screening. These models present significant opportunities to support clinical decision-making in resource-constrained settings and facilitate early intervention in healthcare environments with limited resources. The innovative procedure has the potential to make a significant contribution to the improvement of cancer screening procedures and the current state of health in underserved communities, such as Bangladesh, by reducing dependence on costly diagnostic equipment.

Keywords: Early detection, mathematical model, regression analysis, ANOVA, Response Surface Methodology (RSM), Lung cancer, Breast cancer.

Table of content

LIST OF FIGURES

LIST OF TABLES

CHAPTER 01
INTRODUCTION

1.1 Introduction

Cancer is a significant health issue in Bangladesh, as well as worldwide. The treatment is more effective when detected early, and the patient's outcome improves. In this chapter, we have structured our work into six subsections to provide an overview of our study. Section 1.2 gives a brief background on lung and breast cancer, including their causes and effects on a global scale as well as in Bangladesh. This section emphasizes the importance of early cancer detection. Section 1.3 presents the problem statement, which outlines the reasoning behind the choice of research area and demonstrates the significance of the topic in the modern healthcare setting. In Section 1.4, we outlined the reasons for conducting this research. Cancer is a significant problem in our country, mainly due to ignorance and the unavailability or high prices of advanced detection instruments. Many cases are therefore diagnosed late, resulting in poor treatment outcomes. Our research objectives are expounded on in Section 1.5. We aim to develop a mathematical model that can detect cancer early using clinical data. Our goal is to determine the key parameters and develop first-order and second-order models for predicting lung and breast cancer. Lastly, Section 1.6 discusses the contributions and lessons learned during the writing of this research project, including the growth of information, research, and analysis, as well as advancements in cancer detection methods.

1.2 Background

Lung cancer is an illness whereby abnormal cells develop and multiply within the lung and may form tumors, which interrupt the process of breathing and may extend to other parts of the body. Tobacco smoking is the primary cause, which is present in approximately 85-90 percent of cases. In contrast, other factors noted as contributing to the disease include air pollution, secondhand smoking, radon, asbestos, a genetic predisposition, and, in Bangladesh, groundwater arsenic contamination. The new diagnoses in 2022 were nearly 2.48 million, and 1.8 million deaths were recorded globally, making lung cancer the worst cause of death in terms of incidence (around 23.6/100,000 population) and mortality (approximately 16.8/100,000 population) [1]. In Bangladesh, cancer is attributed to the death of 10 percent of the population a year; the primary cancer in men, as well as in the first six in women, is lung cancer. Despite not having complete national registries, it is estimated that there are approximately 9-15 per 100,000 incidence and almost parallel mortality, contributing to an overall cancer rate of 106 new cases and 75 deaths per 100,000 in 2020 [2,3]. This fatal outcome occurs because lung cancer, in most cases, has no symptoms at the earlier stages of occurrence. Still, at the later stages, the outcome is detected at a time when treatment is less effective. Coupled with the loss of lives, the result is a heavy loss

economically and socially. Globally, there is a high incidence, mainly due to smoking and air-pollution (which are now known to have contributed to an increase in non-smokers), scarcity of early screening methods, such as low-dose CT, in low- and middle-income countries, and late diagnosis. However, screening technologies, including AI-enhanced CT or X-ray scans, can significantly improve survival rates by detecting disease early. Thus, early prediction is crucial in the case of Bangladesh: once lung cancer is predicted and detected early, the chances of it being treated and cured increase significantly, and lives may be saved.

Breast cancer occurs when the cells of the breast start multiplying uncontrollably, and usually that results in the formation of a tumor that may even spread to the tissues around it or to other parts of the body too. It is caused by a complex combination of factors, including genes (such as those carrying a BRCA mutation), hormonal balance, obesity, alcohol consumption, physical inactivity, radiation exposure, and aging. Breast cancer results in massive personal and social losses (lives, emotional trauma, health expenses, and loss of productivity). The current rates of diagnosis and death in women, respectively, aged-standardized to around 46 and 13 per 100,000 in 2022, are about 2.3 million in total and 670,000 worldwide [4]. Breast cancer is the leading cancer among women in Bangladesh and is estimated to have an incidence of 22.5 per 100,000. Dengue fever among women aged 15-49 causes an estimated 21% of deaths in the region [5].

Another problem is that the disease is often identified in a late stage, more than 80 percent of all cases in Bangladesh are at an advanced stage, which makes the treatment more difficult to carry out, and the odds of survival are much lower. Inability to diagnose early, as well as the lack of screening across the board, results in irreversible progression and the loss of potential medical intervention. Nevertheless, when detected early using such methods as self-exams, clinical exams, mammography, or AI-assisted technologies, breast cancer is much easier to treat, and the survival rates are much higher. Consequently, we have narrowed down to early prediction strategies to prevent suffering and ensure a safe life.

1.3 Problem statement

Indeed, one of the most widespread and fatal cancer types all over the world is lung cancer and breast cancer, which have claimed millions of lives annually. One of the primary reasons why they have a high fatality rate is that, most of the time, they are diagnosed at a high or severe stage when the disease has become imminent, and the drugs are deemed to be extremely hard and ineffective. The first stages of detection are critical because the survival rates are much higher in such cases, and full recovery is more likely. Nevertheless, in most low- and middle-income countries, including Bangladesh, the absence of widespread screening centers, limited awareness, and late visits to clinical centers often result in the diagnosis of the ailment at a late stage among most patients. The current achievements of artificial intelligence (AI) and machine learning (ML) techniques demonstrate their potential in cancer prediction and diagnosis. The prevalent datasets utilize Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) to analyze medical image data, histopathology slides, and clinical parameters for predicting cancer in its early stages. More sophisticated auto diagnostic

techniques are made possible by the use of image processing methods to extract subtle patterns that even the human eye cannot detect. Despite all these, there has never been a standard, acceptable, highly accurate predictive mathematical model of lung and breast cancer. The reason is that the correlation level between clinical, genetic, and imaging parameters is very high and stochastic, making it challenging to build a specific mathematical model or equation. It is in this light that it is essential to develop a viable and precise predictive model to address the complex and non-linear interactions of parameters and enhance the early detection of lung and breast cancer. In this context, our aim in undertaking the project is to formulate a mathematical model/formula at a confidence level that can diagnose premature breast and lung cancer using this equation.

1.4 Motivation

Lung cancer is increasingly becoming a health epidemic in our nation, and it is afflicting both men and women. The occurrence of lung cancer is on the rise year by year. Unfortunately, it is often diagnosed at a very late stage due to inadequate sensitization and limited access to specialized diagnostic equipment. Due to this, the patients have diminished rates of recovery and survival, as well as emotional and economic trauma. Even though CT scans, PET scans, and biopsies are some technologies being utilized globally in the detection of lung cancer, these are also cost-intensive and unavailable to most individuals in our nation. Fear and the high price of diagnosis make many people to delay getting diagnosis resulting in poor results. Hence, a non-expensive, handy, and readily available procedure is desperately required to assist in the diagnosis of lung cancer in its initial phase. With these issues in mind, we are developing a mathematical model to predict lung cancer in its early stages.

Breast cancer is emerging as a significant disease condition among women in our nation. It is growing daily. In this country, it was not caught in time due to a lack of knowledge and the unavailability of advanced medical diagnosis. Consequently, they are left to undergo less successful treatment and have a limited possibility of survival. However, if we can identify the very early stage of cancer, we may be able to save lives and promote social development. Ultrasound, MRL, and mammography are involved in breast cancer detection in the world, but we cannot easily do it. Such processes are expensive, hence our people will not be interested in facing a diagnosis. They require a cost-effective and convenient way of detecting cancer so that we can enhance the survival rate. In consideration of these problems, our study aims to develop a mathematical model. Our model will utilize clinical diagnostic data to enhance its accuracy and precision. We are confident that our model will provide a reliable source for doctors, enabling women to achieve improved outcomes and reduced mortality rates.

1.5 Objectives

The primary objective of this paper is to develop a mathematical model for the early detection of breast cancer using the data collected. The precise goals are as follows:

1.To determine some significant independent variables that affect cancer detection through a correlation matrix.

2. The study aims to come up with a first-order mathematical model to define the relationship between the independent and dependent variables in terms of linearity between Lung cancer detection and breast cancer detection.

3. To build a second-order mathematical model that will be used to establish a non-linear relationship and enhance the level of prediction.

4. To compare the goodness-of-fit and performance of first-order and second-order mathematical model, and identify more appropriate one that can be used to predict cancer.

5. To confirm the chosen mathematical model with the rest of the unused dataset to test the model and assess its predictive capacity.

1.6 Scope

Some of the sophisticated tools will be used to ensure that the analysis used in this research is accurate and provides the anticipated result. Hypothesis testing and regression are two statistical activities performed in Minitab, as they help present relevant data clearly and understandably. Our assessment of the extent to which the results vary, as well as our evaluation of how well our developed model is functioning, is also conducted using ANOVA (Analysis of Variance). ANOVA allows us to determine whether the factors of our experiment have a statistically significant effect. Additionally, we utilize Python to perform manual calculations and verify our equations. Python helps us calculate and verify that our model is correct.

CHAPTER 02

LITERATURE REVIEW

2.1 Introduction

The past few years have seen a lot of research being done in the area of early detection of lung cancer & breast cancer. To present the literature in this manner, this chapter is divided into seven sections. Section 2.2 covers how machine learning algorithms may enhance the accuracy of cancer detection beyond what is possible with traditional techniques. In more detail, Section 2.2.1 targets lung cancer, while Section 2.2.2 targets breast cancer. In Section 2.3, the detections of early cancer using artificial intelligence (AI) are discussed. Considerations on the combination of genetic and imaging information have been increasing in the recent past to improve early diagnosis and better treatment outcomes. This section is further subdivided into Part 2.3.1, which covers lung cancer, and Part 2.3.2, which covers breast cancer. Section 2.4 recalls the techniques of processing a picture to detect early cancer. As in other parts, the section 2.4.1 is dedicated to lung cancer whereas Section 2.4.2 is dedicated to breast cancer. Section 2.5 provides an overview of the related work, and Section 2.6 discusses the research gaps that exist in recent studies. The issues are that they are dependent on small data sets, have intense computational demands, and are widely unavailable in low-income countries.

2.2 Machine learning algorithm for cancer detection

2.2.1 Lung cancer

Teena Sabu and Mr. Jinson Devis developed a paper in 2022 [6]. The model of the paper predicts lung cancer depending on multiple input parameters using machine learning algorithms (Decision Tree, SVM, and Logistic Regression). It concludes that the risks to develop lung cancer are larger in chain smokers and it intends to allow early diagnosis to enhance survival. The models are proposed to provide valid predictions, although the measures of accuracy are not outlined. The main implication is that entries with missing values are eliminated, which could result in data loss. Future work consists in the enhancement of the accuracy of prediction, in a more effective way of dealing with missing data, and in the application of more various datasets.

The paper applies the deep residual learning of UNet and ResNet to the detection of lung cancer in 2019. It is combining features with such classifiers as XGBoost and Random Forest. It scored 84% [7] on the LIDC-IDRI dataset, outperforming the previous algorithms. No particular limitations are presented, although data and computational requirements are prevalent among them. The following steps involve enhancing model robustness, interpretability, and dataset diversity.

In 2022, a study surveyed deep learning procedures on lung images, mainly in identifying and categorizing pulmonary nodules [8]. It lays stress on enhancing the clinical relevance in terms of sensitivity and specificity as opposed to conventional measures of accuracy. The most obvious drawback is that medically trained experts did not participate in the model, thus making them less applicable in clinical practice. In future research, the priority should be placed on the cooperation between clinicians and AI researchers, and results that could be meaningful in healthcare practice.

This study aimed to develop a deep learning model for detecting lung cancer in chest X-rays by utilizing image segmentation. Using 629 images during training, it produced 0.73 sensitivity and 0.13 mFPI. Maximum sensitivity occurred on large tumors (31 mm or more) and decreased when the tumors fell within the blind spot or were ill-defined. Limitations are reduced precision in problematic regions. In the future, the work will focus on improving detection in complex areas and enhancing the accuracy of segmentation.

2.2.2 Breast cancer

In 2015, Murat Karabatak developed a new Naive Bayes classifier for breast cancer identification [9]. Conventional techniques, such as mammography, rely on the radiologist, whose results may be inconsistent. The two main problems evident in mammograms, masses and micro-calcifications, are highlighted. Some of the classification techniques with high accuracy rates, as compared in the review, are decision trees (94.74%), rule induction (94.99%), LDA with neural networks (96.8%), SVM (97.2%), and feed-forward neural networks (98.10%) [9]. Bayesian classifiers have a good chance of accurately determining breast cancer risk based on genetic information. By the use of AI and machine learning, accuracy in the detection of cancer has increased. The offered weighted Naive Bayes classifier addresses the shortcomings of classic techniques. It demonstrates superiority to other available methods [9].

In 2014, Semih Ergin and Onur Kilinc emphasized the necessity of improving diagnostic methods. They suggested, wavelet-based breast cancer detection [10]. The research dwelled on the significance of CAD systems. They analyze different feature extraction techniques, including Gabor Wavelet, DWT, PCA, and LBP. Some techniques, such as the Overcomplete Wavelet Transform, have high accuracy (approximately 90%). It also has certain limitations, such as relying on user-defined parameters, which affect its usability [10]. To address these problems, the authors introduce an alternative framework for feature extraction that combines noise filtration with wavelet decomposition. The approach will resolve the current issues that affect the classification process and limit unnecessary biopsies [10].

Shubham Sharma and colleagues studied various determinants in breast cancer detection, which was done in 2018. They employ a machine learning algorithm [11]. The paper contrasts several algorithms, including the Relevance Vector Machine. They score 97 percent on the Wisconsin data set. The Naive Bayes with weighting got to 92%.

A new machine learning model was developed by Abien Fred M. Agarap in 2018. The GRU-SVM that is an integration of a gated recurrent unit (GRU) and a support machine (SVM). It will enhance the accuracy with which breast cancer is classified as compared to the conventional methods [12]. Paper is compared to

GRU-SVM, Linear Regression, multilayer perceptron (MLP), nearest neighbor (NN), softmax regression, and SVM. These models are compared based on classification accuracy, sensitivity, and specificity [12]. The WDBC dataset contains features based on digitized fine needle aspiration (FNA) images of breast masses and plays a prominent role in testing. The structure of the dataset, as well as the method of partitioning it (70% training data and 30% testing data), is fundamental to executing the algorithms [12]. This research has shown that all the applied algorithms behaved well. The accuracy of the MLP algorithm is the best, at 99.04.

Seddik, Ahmed F. et al. (2015) discussed how the automatic detection of breast cancer can be done through a Logistic Regression Model. In their article, they go through many different machine learning methods of breast cancer diagnosis. They discuss their strengths and weaknesses. The use of artificial neural networks (ANNs) is based on high accuracy. In other studies, ANNs are used in conjunction with a multivariate adaptive regression spline (MARS) to obtain better results. Other hybrid models, such as the K-means-based support vector machine (K-SVM) and a fuzzy Min-Max neural network using Random forests, achieved the highest accuracy rates of 98.84% and 98.48%, respectively.

2.3 Artificial Intelligence technique for cancer detection

2.3.1 Lung cancer

A 2020 study proposes a hybrid deep learning model that combines the LeNet, AlexNet, and VGG-16 architectures, utilizing the mRMR feature selection method to classify lung cancer. This model demonstrates high accuracy and fewer false positives compared to conventional methods. Its effectiveness will be assessed using parameters such as sensitivity, specificity, and AUC (Area Under the Curve). However, the study has some limitations, including the use of outdated data from 2008 to 2019 and the absence of specific patient groups, which affects the generalizability of the results. Future directions include conducting more in-depth analyses of the models and developing AI-powered tools to assist physicians..

This paper discusses the use of an Artificial Neural Network (ANN) study to classify X-ray images and identify early lung cancer through image processing and feature extraction. It strongly performed with 100 percent [14] accuracy on 10 test samples. Nonetheless, the data set is small, so it can inhibit generalizability and cause overfitting. Future research should focus on increasing the size of the data, testing other ANN alternatives, and working towards real-time clinical implementation.

The paper adopts machine learning to identify the role of COVID-19 in lung change in patients and differentiate them from cancer therapies. Although accuracy measures are not operationalized, it is apparent in the study that data of very high quality and a representative one will be required. Such weaknesses as a lack of data and delays in its gathering are key. The next step is to enhance the data preparation process and initiate the ethical processing of patient data to achieve improved accuracy in future outbreaks.

Publication of the CT scan-based CAD system based on 3D CNNs to classify lung cancer with 97.4 percent accuracy. The lung segmentation is done via thresholding, and the U-Net is modified to locate a nodule. Weaknesses have a possibility of bias in outcome due to single radiologist-based ratings, as well as segmentation errors. The prospective research targets enhanced segmentation, depth of model, and applicability to different cancers with the application of 3D images.

2.3.2 Breast cancer

According to Seral Sahan et al., a fuzzy-artificial immune system was created, and this was integrated together with the k-NN algorithm in detecting breast cancer [15]. In their research, they examine the other classification approaches that were employed in this area in the past. They show the development and the efficiency of these methods. The new algorithm reported the best accuracy of 99.14% on the Wisconsin Breast Cancer Dataset [15]. Some of the more advanced techniques, such as neuro-fuzzy systems and genetic algorithms, are also addressed in the paper.

Leila Abdelrahman and others created a Convolutional Neural Network (CNN) to identify breast cancer in 2021. They reveal to us the shortcomings of the conventional Computer-Aided Detection (CAD) systems. This is why they have unsuccessful achievements. Nevertheless, recent CNN models have surmounted the drawbacks. Their model is now effective and precise in detection. There are four important tasks in the research that include classifying breast density, detecting asymmetry, identifying calcifications, and finding masses. It also underlines the need for further research to clarify various types of masses further and says something about the problems of training datasets.

Zakia Sultana and her colleagues researched in 2021 how artificial neural networks could assist in the early identification of breast cancer. In their article, they write about the increasing significance of the Computer-Aided Detection (CAD) systems. The system also helps the radiologists detect abnormalities as compared to the conventional approaches. This plays an important role in early diagnosis and ready treatment. The study established that probabilistic neural networks qualify as one of the most effective models of detecting breast cancer with a high rate of accuracy (98.24 percent). It can also be observed in the paper how breast cancer detection is being advanced with the use of artificial intelligence in terms of getting better results in terms of accuracy with regard to detection.

In 2022, Ali Bou Nassif and co-workers examined the application of artificial intelligence (AI) to breast cancer detection. The study groups base their studies on the datasets applied to research and the efficiency of AI models. They demonstrated that deep convolutional neural networks (CNNs) have demonstrated success in early detection. It mentions that the predominant use of CNNs and Naive Bayes is in imaging-related research. However, genetic researches extensively depend on SVMs and decision trees. Before COVID-19, there was an increase in AI research on breast cancer, according to the review, an indication of increasing interest in using AI in detection and treatment. In the study, the conclusion is that genetic sequencing is likely to be used together with imaging data to enhance early detection and therapy. The review is done in reference to very few studies published that combine genetic and imaging information.

2.4 Image processing technique for cancer detection

2.4.1 Lung cancer

Sean Blandin Knight1 and collaborators in 2017 are working on the paper, "Early lung cancer detection: a review of methods." The paper starts with a review of past approaches, those being sputum examination and genetic marker analysis, as well as low-dose CT (LDCT) screening. The mortality is also significantly lowered with LDCT, and it increases the survival significantly, and the sputum analysis is promising to be able to detect a mutation early [16]. But LDCT screening criteria can fail to capture a large number of early-stage cases, and the rates of detection through sputum are very low. Future analysis needs to make screening criteria more complete and incorporate both molecular and imaging methods so that screening can be more accurate.

In 2019, an article rates the profuse clustering technique (improved) and an instantaneously trained neural network to diagnose lung cancer using CT pictures, giving a performance of 98.42% [17], accurate with a minor error in classification. The task of preprocessing leads to an improvement in image quality, but noise is an issue. In the future, we will aim to add noise robustness and make the segmentation further precise.

The research compares five optimization algorithms for lung cancer image segmentation, where GCPSO performed the best in accuracy with a score of 95.89 percent on 20 CT images. Adaptive median filtering and histogram equalization were performed during preprocessing. Although the limitations, such as small sample size and the possibility of overfitting, were mentioned, in the future, one can consider further increasing the datasets, researching new algorithms, and elaborating on real-time clinical tools.

2.4.2 Breast cancer

Methods on mammogram images were proposed by Anuj Kumar Singh and Bhupendra Gupta in 2015. Their works present a style of techniques from naive image processing to sophisticated machine learning. It helps to boost accuracy during the detection and segmentation of cancers in mammograms.

H.D. Cheng and others investigated, through their research work, the application of ultrasound in the automated detection of breast cancer and its classification. The increasing trend in the use of ultrasound in place of mammography, especially in young females and those with dense breast tissue, was also noted. It has been demonstrated that ultrasound can increase the rate of detection of cancer by 17 percent and decrease unnecessary biopsies by 40 percent. Nonetheless, as seen in the study, since the success of ultrasound procedures is solely dependent on the skills of the operator, the quality of radiologists should thus be highly trained. In response to this, the study pointed out how Computer-Aided Detection (CAD) systems could be used to minimize disparity among radiologists and enhance accuracy in terms of making a diagnosis.

Rupali Sood and others (2019) focused on how ultrasound is part of an alternative breast cancer screening tool to mammography. The systematic review and meta-analysis of 26 studies

encompassed 76,058 patients, and the objective was the determination of the diagnostic performance of ultrasound against histologic confirmation and mammography.

2.5 Summary

Having screened the available literature, we grouped the papers dealing with the aspect into four major categories. The former group has done research on the basis of machine learning methods, whereas the latter group is concerned with artificial intelligence (AI) applications. The third group is composed of papers that integrate image processing approaches and conventional methods of diagnostics. By conducting this review, we could familiarize ourselves with the various methods that were adopted by each study, the methods adopted in their study, as well as the limitations that these studies faced. An important point that could be made after this review is that the number of studies that center on the mathematical modeling of cancer detection is very small. As a matter of fact, the specific mathematical model to predict cancer early, particularly with the information gathered in a clinical setting, is yet to be established. This brings out a research gap in the previous studies, which is to be filled in this research by formulating and calibrating a mathematical model to detect the early detection of breast and lung cancer.

Summary shown in table:

Table 2.1: Lung cancer related work

Ref No.	Title	Authors	Method	Findings	Limitations
[18]	Lung Cancer Detection Using Machine Learning	Teena Sabu and Mr.Jinson Devis et al	ML algorithms: Decision Tree, SVM, Logistic Regression	Predicts higher cancer risk in smokers	Ignores missing values (0), possibly losing important data
[19]	Automated Lung Cancer Detection Using AI	Kaviya Sathyakumar et al	Hybrid CNN (LeNet, AlexNet, VGG-16) + mRMR + multiple classifiers	High-performance accuracy, low false positives	Excludes >65 age group, PET scans, radiographs; limited to studies from 2008–2019
[20]	Lung Cancer Detection: A Deep Learning Approach	Siddharth Bhatia, Yash Sinha and Lavika Goel et al	Deep ResNet + UNet for preprocessing, XGBoost & RF for classification	Achieved 84% accuracy	Needs large datasets, risks overfitting, resource-intensive
[21]	Progress and Prospects of Early Detection in Lung Cancer	Sean Blandin Knight1, Phil A. Crosbie et al	Literature review: LDCT, sputum analysis	LDCT increases survival, KRAS mutation detectable early	Screening criteria limited; low mutation detection in sputum
[22]	Deep Learning Techniques to Diagnose Lung Cancer	Lulu Wang	Review of DL for nodule detection, segmentation, classification	DL improves sensitivity, specificity	Lack of radiologist oversight, clinically irrelevant metrics used
[23]	Deep CNNs for Lung Cancer Detection	Albert Chon et al	3D CNN (vanilla & Googlenet) +	Efficient with fewer labels	High false positives, non-SOTA accuracy

			modified U-Net on Kaggle dataset		
[24]	Early Detection Using Data Mining	Kawsar Ahmed and Abdullah-Al-Emran et al	K-means + AprioriTid + Decision Tree	Simple, cost-effective prediction tool	Lack of awareness in Bangladesh, late-stage diagnosis
[25]	Early Lung Cancer Detection Using ANN	T. Pandiangan1, I. Bali et al	ANN + image processing (edge detection, segmentation)	100% detection on test samples	Very small dataset (10 test images), overfitting likely
[26]	Liquid Biopsy for Early Detection	Mariacarmela Santarpia1 , Alessia Liguori1 et al	Blood-based biomarkers: miRNA, ctDNA, exosomes, etc.	87% sensitivity, 81% specificity	Study heterogeneity; mostly small, single-center studies
[27]	DL-Based Lung Cancer Detection & Classification	N Kalaivani1, N Manimaran et al	DenseNet + AdaBoost on 201 CT images	90.85% accuracy	Small dataset; limited imaging modality
[28]	DL for Lung Cancer on Chest Radiographs	Akitoshi Shimazaki1 , Daiju Ueda1et al	Segmentation model on 629 train/151 test X-rays	Sensitivity varies by tumor position; 87% for easy regions	Blind-spot tumors hard to detect; low dice score (0.52)
[29]	Lung Cancer Detection Using ANN	Ibrahim M. Nasser, Samy S. Abu-Naser et al	ANN on 14 features from survey dataset	96.67% accuracy	Small, possibly biased dataset; generalizability concerns

Table 2.2: Breast cancer related work

Ref No	**Title**	**Autor**	**method**	**Observation**	**limitations**
[9]	A new classifier for breast cancer detection based on Naïve Bayesian, 2015	Murat Karabatan	Naïve Bayesian Classifier, Weighted Naïve Bayesian Approach.	Sensitivity of 99.11%, specificity of 98.25%, and an overall accuracy of 98.54%	computational expense, dependency on weight initialization, inherent drawbacks of previous method
[30]	A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram,2015	Anuj Kumar Singh and Bhupendra Gupta	Image processing approach	Improved segmentation compared to prior methods; better tumor region isolation visually	Limited to low-resolution images; results depend on intensity differences

[11]	Breast Cancer Detection Using Machine Learning Algorithms,2018	Shubham Sharma et al.	Random Forest, k-NN, Naïve Bayes on Wisconsin Diagnosis Dataset	k-NN showed best accuracy (95.90%), Random Forest (94.74%), Naïve Bayes (94.47%)	Small dataset,
[31]	Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis, 2016	Dana Bazazeh et al.	three machine learning 12echnique—Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks (BN)	accuracy rate of 97% for the classifiers	overfitting due to the small dataset size.
[32]	Convolutional neural networks for breast cancer detection in mammography: A survey,2021	Leila Abdelrahman et al.	convolutional neural networks (CNNs)	mean accuracy of 92.73%, with sensitivity 94.58% and specificity 91.67%	Lack of explainability in CNN models. Challenge in devices and datasets
[33]	Early detection and Screening for breast cancer, 2017	Cathy coleman et al.	combination of mammography, (CBE), and additional imaging 12echnique	accuracy rate of 85% to 90% in detecting early-stage breast cancer	Variability in early detection practices. Conflicting screening guidelines
[34]	Infrared imaging technology for breast cancer detection – Current status, protocols and new directions,2017	Satish G. Kandlikar et al.	Review of infrared imaging, inverse modeling, and AI for breast cancer detection.	Sensitivity: 71%-94%; Specificity: 14%-85% depending on thresholds.	Variability in devices, algorithms, and threshold choices.
[35]	Delays in Breast Cancer Detection and Treatment in Developing Countries, 2017	Monica M Rivera-Franco and Eucario Leon-	Review of healthcare delays in breast cancer detection and treatment in LMICs	High delay rates: 17%-42.5% globally.	Limited data from LMICs Focus on specific regions

		Rodrigue z			
[36]	Ultrasound for Breast Cancer Detection Globally: A Systematic Review and Meta-Analysis, 2019	Rupali Sood, MPH et al.	Meta-analysis of ultrasound as a primary breast cancer detection	Sensitivity: 80.1%-89.2%; Specificity: 88.4%-99.1%	High heterogeneity across studies; limited direct comparisons to mammography.
[37]	Logistic Regression Model for Breast Cancer Automatic Diagnosis,2015	Ahmed F. Seddik et al.	data analysis followed by logistic regression	Accuracy: 98.9%; Sensitivity: 98.5%; Specificity: 99.1%.	Reliance on WDBC dataset, limiting generalizability. Potentially missed important features.

2.6 Research Gap

Cancer is one of the greatest medical issues in the world, and early detection is highly significant in increasing the rate of survival. Several cancer-detecting techniques have been established over the years to diagnose them, including the employment of imaging tools like ultrasound, mammography, computer-based approaches like machine learning and artificial intelligence (AI), etc. Although these methods have assisted in numerous aspects, they still have quite significant issues. To begin with, the majority of these techniques are developed on small datasets. The model fails to learn when the data are scarce, and hence, there are chances that predictions are not accurate. In addition, although some of these models may be accurate, most of them are costly, elaborate, and require sophisticated technology. This renders them unreliable in low-income areas or in rural settings where high-tech equipment and qualified technicians are unavailable. Second, such instruments as ultrasound and mammography are highly dependent on the level of the machine and the skill of its operator. This may cause them to fail in their detection, and more so where the facilities available are minimal in places. Machine learning and AI have been applied by many researchers in recent years to detect cancer. Such methods may provide a satisfactory

performance in some cases; however, they can mostly resemble the so-called black box, i.e., their decision-making process is not completely understandable. They also need huge volumes of data and spacious computers, which may be lacking in a low-resource scenario. The use of mathematical models to detect cancer has received very little attention. Relationships between various factors explained with the help of formulas and equations are called a mathematical model. These models are simplified to understand, less data is required, and are applicable in limited resources as well. They are also more transparent, that is, we also know how they make decisions. This indicates the obvious loophole in the existing research. A cancer detection tool that is minimalistic, affordable, interpretable, and operates on a relatively small amount of data is needed. This is the reason why in this research, we are trying to come up with a new mathematical model for detecting cancer. We would like to establish a framework that is able to provide relevant results and is useful in settings where sophisticated technology and large datasets are not accessible.

CHAPTER 03

METHODOLOGY

3.1 Introduction

The chapter contains an elaborate description of the parameters, data collection, and building of the model employed in our study. Section 3.2 represents a description of the parameters and data collection procedure in the case of lung cancer. More precisely, the chosen parameters are described in Section 3.2.1, and sample data is provided in Section 3.2.2. In the same way, Section 3.3 deals with breast cancer. The description of the involved parameters is given in Section 3.3.1, whereas Section 3.3.2 covers sample data to be analyzed. Section 3.4 gives an explanation of the data preprocessing methods that were used to prepare the dataset to develop a model. Section 3.5 focuses on the development of mathematical models (both first-order, which are linear, and second-order or polynomial). In this case, Section 3.5.1 is the part of the study devoted to the lung cancer model, and Section 3.5.2 to the breast cancer model. The third section is Part 3.6, which is reserved for implementation. Response Surface Method (RSM) is applied and described in 3.6.1. In Section 3.6.2, the process of obtaining model coefficients is explained, and Section 3.6.3 is devoted to measuring coefficients. The effectiveness of the model will be checked by checking the fitness of the model in the developed mathematical equations as discussed in Section 3.7. Lastly, Section 3.8 covers the testing phase. Section 3.8.1 is concerned with how a program was used to compute outputs based on the developed equations, and Section 3.8.2 makes a comparison, as well as finalized mathematical equations.

3.2 Parameter and data collection for lung cancer

In the case of lung cancer, the risk level of our dataset contains 1000 records and 24 variables, such as different features of patients, clinical measures, risk factors, lifestyle, and symptoms. Our thesis helps us to choose variables. Predictor variable such as Age, Gender, Air pollution, Alcohol use, Dust Allergy, Alcohol use, Dust Allergy, OccuPational Hazards, Genetic Risk, Chronic Lung Disease, Balance Diet, Obesity, Smoking, Passive smoker, Chest Pain, Coughing of Blood, Fatigue, Weight loss, Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of finger Nails, Frequent Cold, Dry cough, Snoring was determined and the response variable level of cancer The source of collection of this dataset is Rishi Damarla of Windsor, Canada. In this case, 980 data points will be applied in the building of models, and the remaining 980 will be employed in the testing of the models.

3.2.1 Parameter for lung cancer

Parameters (predictor & response) for lung cancer are shown below:

Table 3.1: Predictor variables for lung cancer

Category	Parameter	Description	Symbol
Patient Information	Age	Age will be a significant factor since many diseases such as cancer risk are age-dependent.	X_1
	Gender	Some diseases have different impacts on males and females.	X_2
Risk Factors & Lifestyle	Air pollution	One of the known risk factors includes long-term exposure to polluted air particularly on lung diseases such as lung cancer.	X_3
	Alcohol Use	Such a habit of frequent alcohol intake could also lead to the deterioration of health in general and the risk of diseases, in particular.	X_4
	Dust Allergy	The allergic reactions to dust may be indicative of sensitivity to airborne particles which may pollute the pilgrimage of the lungs.	X_5
	Occupational Hazards	Occupational exposure to the risk of lung problems through exposure to, among others, chemicals, asbestos, fumes.	X_6
	Genetic Risk	An individual can be at risk through the family history of cancer and related diseases related to the genes that they have inherited.	X_7
	Chronic Lung Disease	The risk of developing lung cancer increases because of existing lung conditions: asthma, COPD, or bronchitis.	X_8
	Balanced Diet	The immune system is supported by a healthy balanced diet that could reduce the chances of diseases.	X_9
	Obesity	Overweight is likely to affect the lung cancer in individuals who have never smoked.	X_{10}
	Smoking	The greatest risk factor of lung cancer is smoking.	X_{11}
	Passive Smoker	The risk of developing lung cancer is also associated with inhalation of smoke of other people (secondhand smoke) including incidences that involve non-smokers.	X_{12}
Symptoms	Chest Pain	Chest pains that can be as a result of heart or chest complications.	X_{13}
	Coughing of Blood	One of the very extreme symptoms is the coughing up of blood and it is likely associated with lung disease.	X_{14}
	Fatigue	Feeling tired or weak on a regular basis could be the symptom of some health complications.	X_{15}
	Weight Loss	Accidental weight loss is also alarming to define serious health problems including cancer.	X_{16}
	Shortness of Breath	The difficulty in breathing, which might be an indication of the lung and heart issues.	X_{17}

	Wheezing	Grating breathing which usually is associated with constricted airways.	X_{18}
	Swallowing Difficulty	Difficulty in swallowing may be the symptom of the throat or lung complication.	X_{19}
	Clubbing of Finger Nails	Chronic lung disease could be suggested by thickened fingertips and hooked nails.	X_{20}
	Frequent Cold	Standard colds could indicate an inherent problem with the immune system, or a long-term problem.	X_{21}
	Dry Cough	The lasting dry cough is one of the initial signs of illnesses of the lungs.	X_{22}
	Snoring	Snoring is frequent and loud, which is an indication of difficulties in breathing during sleeping.	X_{23}

Table 3.2: Response Variable for lung cancer

Category	**Parameter**	**Description**	**Symbol**
Result	Level	This indicate cancer risk level(high, low, medium)	Y

3.2.2 Data collection for lung cancer

In this research, 1 000 data samples were utilized. Among these, The font of the training data sets was 980 samples in which the mathematical model was constructed and tuned according to the input variables and output outcomes. The other 20 samples were to be tested, so that the performance of the model on data that will not be seen before can be observed, and the accuracy and reliability of the model can be checked. The separation assists in making sure that the model is not only developed well, but also that the model works efficiently when presented with new cases.

Table 3.3: sample data of lung cancer

index	Age	Gender	Air Pollut	Alcohol u	Dust Alle	Occu Pati	Genetic F	chronic L	Balanced	Obesity	Smoking	Passive S	Chest Pai	Coughing	Fatigue	Weight L	Shortnes	Wheezin	Swallowi	Clubbing	Frequent	Dry Coug	Snoring	Level
0	33	1	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	Low
1	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	Medium
2	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
3	37	1	7	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
4	46	1	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	High
5	35	1	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
6	52	2	2	4	5	4	3	2	2	4	3	2	2	4	3	4	2	2	3	1	2	3	4	Low
7	28	2	3	1	4	3	2	3	4	3	1	4	3	1	3	2	2	4	2	2	3	4	3	Low
8	35	2	4	5	6	5	6	5	5	5	6	6	6	5	1	4	3	2	4	6	2	4	1	Medium
9	46	1	2	3	4	2	4	3	3	3	2	3	4	4	1	2	4	6	5	4	2	1	5	Medium
10	44	1	6	7	7	7	7	6	7	7	7	8	7	7	5	3	2	7	8	2	4	5	3	High
11	64	2	6	8	7	7	7	6	7	7	7	8	7	7	9	6	5	7	2	4	3	1	4	High
12	39	2	4	5	6	6	5	4	6	6	6	6	6	6	5	3	2	4	3	1	7	5	6	Medium
13	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
14	27	2	3	1	4	2	3	2	3	3	2	2	4	2	2	2	3	4	1	5	2	6	2	Low
15	73	1	5	6	6	5	6	5	6	5	8	5	5	5	4	3	6	2	1	2	1	6	2	Medium
16	17	1	3	1	5	3	4	2	2	2	2	4	2	3	1	3	7	8	6	2	1	7	2	Medium
17	34	1	6	7	7	7	6	7	7	7	7	7	7	8	4	2	3	1	4	5	6	7	5	High
18	36	1	6	7	7	7	7	7	6	7	7	7	7	7	8	5	7	6	7	8	7	6	2	High
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
997	25	2	4	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High
998	18	2	6	8	7	7	7	6	7	7	8	7	7	9	3	2	4	1	4	2	4	2	3	High
999	47	1	6	5	6	5	5	4	6	7	2	3	4	8	8	7	9	2	1	4	6	7	2	High

(1-10) scale show the severity, frequency, or exposure level scale for variables.

(1-very low; 2-low; 3- slightly low; 4- below average; 5-average; 6- slightly high; 7-high; 8-very high; 9-severe; 10- extremely severe)

(low-0; medium-1; high-2)

3.3 Parameter and data collection for Breast cancer

To predict breast cancer, a total of 570 records and 32 variables were used to include features such as tumor size related features, uniformity of cells, Shapes features, Standard Error Features, Worst-case Features and others are the dependent variables to predict cancer. Dependent variable is represented by diagnosis. Erdem Taha of Istanbul, Turkey gathers this data. In this case, 550 data points will be employed in the development of the model, and 20 data points will be employed in testing.

3.3.1 Parameter for breast cancer

Parameters (predictor & response) for breast cancer are shown below:

Table 3.4: Predictor variables for breast cancer

Category	Parameter	Description	Symbol
A tumor cells Mean values	Radius_mean	The mean value between the center of the cell and the edge (perimeter) of the cell.	X_1
	Texture_mean	The standard deviation of gray-scale values that indicates the difference in texture of the cell.	X_2
	perimeter_mean	Mean length along boundary of the cell.	X_3
	area_mean	Mean area (one) of the cell.	X_4
	smoothness_mean	the roughness of the edges on the cell according to the variation in the lengths of radii.	X_5
	compactness_mean	A dimension that is a measure of perimeter and of area as an indicator of the degree to which the cell shape is compact. ((perimeter 2 / area) 1.0)3	X_6
	concavity_mean	Mean inward sweep attempting (concave section) of the boundary of the cell.	X_7
	concave points_mean	The number of inward-directed (convex) constituents of the perimeter of the cell.	X_8
	symmetry_mean	The cell shape is how symmetrical.	X_9
	fractal_dimension_mean	An amount on the level of complexity or coarseness of the cell edge.	X_{10}
Standard error of values	radius_se	Measures change of radius; should this value be large, it can indicate inconsistent tumor growth.	X_{11}
	texture_se	Suggests difference in internal texture, which aids in the identification of abnormal tissue patterns.	X_{12}

	perimeter_se	Displays non uniformity along the tumor edge.	X_{13}
	area_se	Fluctuating tumor mass size is detectable by capturing the changes in the mass size; this can be a warning.	X_{14}
	smoothness_se	Measures irregularities of smoothness of the tumor edge.	X_{15}
	compactness_se	Signals variability in the compactness of the tumor in various areas.	X_{16}
	concavity_se	Records difference in the depth of the concave (inward) points on the boundaries of the tumor.	X_{17}
	concave points_se	Quantifies the inconstancy of the number of concave regions.	X_{18}
	symmetry_se	Displays the extent of variation in symmetry in tumor shape; the bigger the variation the more indicational of malignancy.	X_{19}
	fractal_dimension_se	Records the degree of variations in roughness or complexity of the edge of the tumor in various regions	X_{20}
Worst-case Values	radius_worst	It is quite common to relate higher values of radius to malignant (cancerous) tumors.	X_{21}
	texture_worst	Tough texture can show terrible deformation in tissue architecture.	X_{22}
	perimeter_worst	When the boundary is longer this could indicate a more aggressive or malignant tumor.	X_{23}
	area_worst	The increase of the tumor region may indicate malignant cancer, which grows rapidly.	X_{24}
	smoothness_worst	A major indication of malignancy is the highly irregular boundary of tumor.	X_{25}
	compactness_worst	Higher compactness irregularity can be related with cancer.	X_{26}
	concavity_worst	The boundary makes deep and repeated inward curvatures that are tightly connected to cancer.	X_{27}
	concave points_worst	Large concave areas are usually an indication of some cancerous tumor.	X_{28}
	symmetry_worst	By the fact that malignant tumors have more asymmetry in shape, compared to benign, they are more pronounced in the former, being one of the most typical.	X_{29}

	fractal_dimension_worst	Tumors possessing extensive or rugged appearance of their edges are usually a sign of cancer.	X_{30}

Table 3.5: Response variables for breast cancer

Category	**Parameter**	**Description**	**Symbol**
Result	diagnosis	This indicate cancerous & non- cancerous value(Malignant & Benign)	Y

3.3.2 Data collection for breast cancer

This paper used 570 data samples. Among them: The model development and training took 550 samples. These data were required to construct and optimize mathematical model through studying the links between input variables and target outcome. The rest 20 samples were employed in testing. The model was not exposed to these data during its training, and their presence allowed assessing the ability of the model to identify new unseen cases. That method assist in testing the capacity of the model to generalize it and make sure that the outcome is not prejudiced by the training information.

Table 3.6: Sample data for breast cancer

Index	Radius_n	Texture_	perimete	area_me	smoothne	compactn	concavity	concave p	symmetr	fractal_di	radius_se	texture_s	perimete	area_se	smoothnes	compactne	concavity	concave pc	symmetry	fractal_dim	radius_w	texture_	perimete	area_wo	smoothne	compactn	concavity	concave p	symmetr	fractal_di	diagnosis
0	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189	M
1	20.57	21.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902	M
2	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758	M
3	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173	M
4	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678	M
5	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985	0.1244	M
6	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063	0.08368	M
7	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196	0.1151	M
8	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378	0.1072	M
9	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366	0.2075	M
10	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948	0.08452	M
11	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792	0.1048	M
12	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176	0.1023	M
13	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809	0.06287	M
14	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596	0.1431	M
15	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218	0.1341	M
16	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029	0.08216	M
17	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706	0.1142	M
18	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768	0.07615	M
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
566	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	0.05302	0.159	0.05648	0.4564	1.075	3.425	48.55	0.005903	0.03731	0.0473	0.01557	0.01318	0.003892	18.98	34.12	126.7	1124	0.1139	0.3094	0.3403	0.1418	0.2218	0.0782	M
567	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	0.152	0.2397	0.07016	0.726	1.595	5.772	86.22	0.006522	0.06158	0.07117	0.01664	0.02324	0.006185	25.74	39.42	184.6	1821	0.165	0.8681	0.9387	0.265	0.4087	0.124	M
568	7.76	24.54	47.92	181	0.05263	0.04362	0	0	0.1587	0.05884	0.3857	1.428	2.548	19.15	0.007189	0.00466	0	0	0.02676	0.002783	9.456	30.37	59.16	268.6	0.08996	0.06444	0	0	0.2871	0.07039	B

(These variables values are derived from the digital analysis of cell nuclei in fine needle aspirate (FNA) images of breast tissue)

[M-4(cancerous); B-2(non- cancerous)

3.4 Correlation matrix to determine significant parameter

The number of parameters of independent variables is very large and it can be too burdensome. This is the reason why we need to select significant parameters. In order to select a significant parameter, we will utilize a correlation matrix. A correlation matrix literally refers to a table that indicates the relationship between variables. It is one of the influential statistical tools that is employed to calculate the intensity and alignment of the linear association among various variables in a data set. It will assist us in realizing the relationship that exists between all the 23/30 predictor variables (x1-x23 OR x1-x30) and the target variable (y). This was a matrix, which was used to summarize a vast set of data. Our big data will cause complications in formulating equations as well as subsequent calculations. A correlation matrix can assist us in finding the main variables that are related to the response variable. Thus, we will summarize our data through a correlation matrix. Correlation Coefficient, r, varies from -1 to +1. +1: very good positive linear relationship

0: no linear relation

–1: perfect negative linear relationship

Pearson Correlation coefficient formula-

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\times\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3.1)$$

Here,

x_i, y_i = values of variables x & y

$\bar{x}$, $\bar{y}$ = mean of x & y

n = number of observations

Using this mathematical formula, we will find out significant variable for our lung cancer and breast cancer model.

3.5 Mathematical model development

We seek to come up with a mathematical model for the early detection of cancer. Mathematical models translate the problems of reality into equations. Mathematical models come in various forms that could be used to analyze and predict a result using input data. These models allow people to comprehend connections between variables and make sound choices. The most usual examples of mathematical models are: Linear models, which are based on the presupposition that the input variables and the output are connected with a straight line. Non-linear models are used because they describe more of the curved relationships that cannot be described through a straight-line explanation. Polynomial models - An expansion of linear models in that they too can describe a more complicated pattern by incorporating squared or terms of higher degree. In this paper, we shall concentrate on linear and polynomial functions to fashion a model that detects cancer. The models were selected because they are simple and easy to interpret, and they are used to generate important patterns in the data. Here is a general form of a mathematical equation:

$$Y = f(X_1, X_2, X_3, X_4 \dots \dots \dots X_n) + \varepsilon \quad (3.2)$$

This is a normal equation. The function f can be linear, non-linear, polynomial etc. Here, Y is the response variable, f is the response variable of $(X_1, X_2, X_3, X_4 \dots \dots \dots X_n)$; values of are mention before and ε is the error of this equation. We would like to deal with low order polynomial to process variables.

i) single order Equation

A given mathematical relation that describes the interdependency between the numerous independent variables (inputs) of a system and independent variable (output) is referred to as a first-order model. Single order model is look like –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots \beta_n X_n + \varepsilon \quad (3.3)$$

If we simply the equation then it will be look like this-

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \quad (3.4)$$

Here, X_i(1,2,3,4……n) are the predictor variable. β_0 is the constant of single order equation and β_i is the regression co-efficient. Y is response variable of single order model.

ii) Polynomial Equation

The polynomial model is an extension of the single-order modeling concept with an addition of not only linear terms but also the quadratic terms (squared terms) and interaction terms (products of two or more variables). This enables the model to calculate nonlinear dependencies between the independent variables and the dependent, whereas it appears much more realistic to a lot of real-life systems.

Here is a general form of a mathematical equation:

$$Y = f(X_1, X_2, X_3, X_4 \dots \dots \dots X_n) + \varepsilon \quad (3.2)$$

Here f is polynomial function. Now the Polynomial equation look like-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots \beta_n X_n + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{44} X_4^2 + \cdots \beta_{nn} X_n^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{14} X_1 X_4 + \beta_{15} X_1 X_5 + \cdots \beta_{(n-)(n)} X_{n-1} X_n + \varepsilon \quad (3.5)$$

A second-order model consists following terms-

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \beta_{ii} X_i^2 + \sum_{i,j=1, i \neq j}^{n} \beta_{ij} X_i X_j + \varepsilon \quad (3.6)$$

Here, X_i(1,2,3,4……n) are the predictor variable. β_0 is the constant of polynomial equation and β_i is the regression co-efficient. Y is response variable of single order model.

3.5.1 Mathematical model for lung cancer

For lung cancer,According to preprocessing value we re-arrange our parameter. $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$ and X_{12} are obesity**(O_b)**,coughing of blood**(C_b)**, alcohol use**(A_u)**, dust allergy**(D_a)**, passive smoker**(P_s)**, balance diet**(B_d)**, genetic risk**(G_r)**, occupational hazards**(O_h)**, chest pain**(C_p)**, air pollution**(A_p)**, Fatigue**(F_a)** and chronic lung cancer**(C_l)** respectively for level. Equation can be present as,

$$Y = f(O_b, C_b, A_u, D_a, P_s, B_d, G_r, O_h, C_p, A_p, F_a, C_l) + \varepsilon \quad (3.7)$$

i) single order Equation

General formula for first-order model according to pre-process lung cancer data:

$$Y = C_0+C_1X_1+C_2X_2+C_3X_3+C_4X_4+C_5X_5+C_6X_6+C_7X_7+C_8X_8+C_9X_9+C_{10}X_{10}+C_{11}X_{11}+C_{12}X_{12}+\varepsilon \quad (3.8)$$

Here, C_0 is constant for level. Where $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}$ are regression coefficient of connecting factor.

When we will simplify our equation then it look like-

$$Y = C_0 + \sum_{i=1}^{12} C_i X_i + \varepsilon \quad (3.9)$$

ii) Polynomial equation

General formula for second-order model according to pre-process lung cancer data:

$$Y = C_0+C_1X_1+C_2X_2+C_3X_3+C_4X_4+C_5X_5+C_6X_6+C_7X_7+C_8X_8+C_9X_9+C_{10}X_{10}+C_{11}X_{11}+C_{12}X_{12}+C_1X_1^2+C_2X_2^2+C_3X_3+C_4X_4^2+C_5X_5^2+C_6X_6^2+C_7X_7^2+C_8X_8^2+C_9X_9^2+C_{10}X_{10}^2+C_{11}X_{11}^2+C_{12}X_{12}^2+C_{12}X_1X_2+C_{13}X_1X_3+C_{14}X_1X_4+C_{15}X_1X_5+C_{16}X_1X_6+C_{17}X_1X_7+C_{18}X_1X_8+C_{19}X_1X_9+C_{110}X_1X_{10}+C_{111}X_1X_{11}+C_{112}X_1X_{12}+C_{23}X_2X_3+C_{24}X_2X_4+C_{25}X_2X_5+C_{26}X_2X_6+C_{27}X_2X_7+C_{28}X_2X_8+C_{29}X_2X_9+C_{210}X_2X_{10}+C_{211}X_2X_{11}+C_{212}X_2X_{12}+C_{34}X_3X_4+C_{35}X_3X_5+C_{36}X_3X_6+C_{37}X_3X_7+C_{38}X_3X_8+C_{39}X_3X_9+C_{310}X_3X_{10}+C_{311}X_3X_{11}+C_{312}X_3X_{12}+C_{45}X_4X_5+C_{46}X_4X_6+C_{47}X_4X_7+C_{48}X_4X_8+C_{49}X_4X_9+C_{410}X_4X_{10}+C_{411}X_4X_{11}+C_{412}X_4X_{12}+C_{56}X_5X_6+C_{57}X_5X_7+C_{58}X_5X_8+C_{59}X_5X_9+C_{510}X_5X_{10}+C_{511}X_5X_{11}+C_{512}X_5X_{12}+C_{67}X_6X_7+C_{68}X_6X_8+C_{69}X_6X_9+C_{610}X_6X_{10}+C_{611}X_6X_{11}+C_{612}X_6X_{12}+C_{78}X_7X_8+C_{79}X_7X_9+C_{710}X_7X_{10}+$$

$C_{711}X_7X_{11}+ C_{712}X_7X_{12}+C_{89}X_8X_9+ C_{810}X_8X_{10}+ C_{811}X_8X_{11}+ C_{812}X_8X_{12}+C_{910}X_9X_{10} + C_{911}X_9X_{11} + C_{912}X_9X_{12} +C_{1011}X_{10}X_{11} + C_{1012}X_{10}X_{12} +C_{1112}X_{11}X_{12}$ (3.10)

After simplifying this equation,we get this-

$$\text{level} = C_0 + \sum_{i=1}^{12} C_i X_i + \sum_{i=1}^{12} C_{ii} X_i^2 + \sum_{i,j=1,i\neq j}^{12} C_{ij} X_i X_j + \varepsilon \quad (3.11)$$

Here, C_0 is constant for level. Where $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}$ ……..are regression coefficient of connecting factor.

3.5.2 Mathematical model for breast cancer

For breast cancer, For lung cancer,According to preprocessing value we re-arrange our parameter. $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}\ and\ X_{13}$ are radius_mean(**R_m**), perimeter_mean(**P_m**), area_mean(**A_m**),compactness_mean(**C_m**),concavity_mean(**C_{vm}**), concave point_mean(**C_{pm}**), symmetry_mean(**S_m**), radius_wrost(**R_w**), perimeter_wrost(**P_w**), area_wrost(**A_w**),compactness_worst(**C_w**),concavity_worst(C_{vm}), concave point_worst(C_{pw}), respectively for diagnosis.

$$Y = f(R_m, P_m, A_m, C_m, C_{pm}, C_{pm}, S_m, R_w, P_w, A_w, C_w, C_{vw}, C_{pw}) + \varepsilon \quad (3.12)$$

Also can be written as,

$$\text{Diagnosis} = f(R_m, P_m, A_m, C_m, C_{pm}, C_{pm}, S_m, R_w, P_w, A_w, C_w, C_{vw}, C_{pw}) + \varepsilon \quad (3.13)$$

i) Single-order equation:

General formula for first-order model according to breast cancer data:

$$Y = D_0+D_1X_1+D_2X_2+D_3X_3+D_4X_4+D_5X_5+D_6X_6+D_7X_7+D_8X_8+D_9X_9 +D_{10}X_{10} +D_{11} X_{11} +D_{12} X_{12}+D_{13} X_{13}+ \varepsilon \quad (3.13)$$

Here, Here, D_0 is constant for level. $D_0, D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}$ regression coefficient of related variables for single order model.

When we simplify equation can be written as-

$$\text{diagnosis} = D_0 + \sum_{i=1}^{10} D_i X_i + \varepsilon \quad (3.14)$$

ii) Polynomial equation:

General formula for second-order model according to breast cancer data:

$$Y=D_0+D_1X_1+D_2X_2+D_3X_3+D_4X_4+D_5X_5+D_6X_6+D_7X_7+D_8X_8+D_9X_9+D_{10}X_{10}+....D_1X_1^2+D_2X_2^2+D_3X_3+D_4X_4^2+D_5X_5^2+D_6X_6^2+D_7X_7^2+D_8X_8^2+D_9X_9^2+D_{10}X_{10}^2+....D_{12}X_1X_2+D_{13}X_1X_3+D_{14}X_1X_4+D_{15}X_1X_5+ D_{16}X_1X_6+ D_{17}X_1X_7+ D_{18}X_1X_8+ D_{19}X_1X_9+ D_{110}X_1X_{10}+ D_{23}X_2X_3+D_{24}X_2X_4+D_{25}X_2X_5+D_{26}X_2X_6+ D_{27}X_2X_7+ D_{28}X_2X_8+ D_{29}X_2X_9+ D_{210}X_2X_{10}+ D_{34}X_3X_4+ D_{35}X_3X_5+ D_{36}X_3X_6+ D_{37}X_3X_7+ D_{38}X_3X_8+ D_{39}X_3X_9+ D_{310}X_3X_{10} +D_{45}X_4X_5+ D_{46}X_4X_6+ D_{47}X_4X_7+ D_{48}X_4X_8+ D_{49}X_4X_9+ D_{410}X_4X_{10}+ D_{56}X_5X_6+D_{57}X_5X_7+D_{58}X_5X_8+D_{59}X_5X_9+ D_{510}X_5X_{10}+ D_{67}X_6X_7+ D_{68}X_6X_8+D_{69}X_6X_9+ D_{610}X_6X_{10}+ D_{78}X_7X_8+ D_{79}X_7X_9+ D_{710}X_7X_{10} +D_{89}X_8X_9+ D_{810}X_8X_{10}+ D_{910}X_9X_{10} + \varepsilon \quad (3.15)$$

If we simplify the equation then it would be look like-

$$\text{diagnosis} = D_0 + \sum_{i=1}^{13} D_i X_i + \sum_{i=1}^{13} D_{ii} X_i^2 + \sum_{i,j=1,i\neq j}^{13} D_{ij} X_i X_j + \varepsilon \quad (3.16)$$

Here, D_0 is constant for level. Here,$D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}$…… regression coefficient of related variables.

.

3.6 Implementation

This paper will focus on the prediction of breast cancer and lung cancer using the regression equations formulated earlier. These functions will be based on estimating the probability of possessing cancer with the help of the several input characteristics. But in order to provide these equations with functionality and their potential to be used by the real data we have to find out the suitable coefficients of each of these models. In the case of lung cancer, the coefficients will be written as: $C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}\ldots$ in which the coefficients represent the effect of a particular variable or interaction term of the outcome that is set to be predicted. Equally, in the case of breast cancer the coefficients shall be considered as $D_0, D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}$. These coefficients play a very critical role since they determine the type of relationship that exists between every variable and the risk of cancer. We are going to apply the Response Surface Methodology (RSM) to estimate these coefficients. We can use RSM to be able to fit the mathematical model of our dataset and estimation of coefficients and assess the performance of the model. After the determination of coefficients, final equations may be applied to forecast the existence of lung cancer or breast cancer in a patient on the basis of their clinical and diagnostic properties.

3.6.1 Response Surface Methodology (RSM)

Response Surface Methodology (RSM) refers to a set of statistical and mathematical processes that are employed in the search of the connection between a number of input parameters and single or multiple output parameters. It assists us to know the influence of the inputs on the output and the way the result can be enhanced or geared. The primary aim of RSM is to develop the model that can predict the output and apply the model to determine optimal conditions to achieve the required result.

3.6.2 Coefficient Determination

A coefficient is a figure that multiplies a variable. It may be a parameter with an equation. Coefficients are of some types. Different section represents different coefficient. In RMS model we shall consider regression coefficient. Regression coefficient is a numerical index which give an expression of the relationship between the independent variable(predictor) and the dependent variable(response) in a model.

Manually regression coefficient calculation formula:

$$\beta = (X^T X)^{-1} X^T Y \qquad (3.17)$$

Here, X = design matrix (with intercept column); Y = response vector; β = vector of coefficient.

We have to minimizes the sum of square residuals between predicted and actual Y. Sum of square error:

$$SSE = \sum(y_i - \bar{y}_i)^2 \quad (3.18)$$

Here, y_i = actual observe value; $\bar{y}_i$= predicted value from the model

Our coefficients of mathematical models will be calculated both manually and by means of the Response Surface Methodology (RSM). This is going to be done in two ways the single order and also in the polynomials. This will enable us to conclude on the best and valid coefficients to be used in predicting the lung and breast cancer by comparing the results of the two methods. The synergistic method will ensure that we construct a better and more efficient model of early cancer detection.

3.6.3 coefficient evaluation

In regression, we can evaluate the coefficients in order to determine the significance of each of the variables in the prediction of the output. To do this, we use three primary things: Standard Error (SE), T-value and P-value. The Standard Error tells us the extent to which the estimating coefficient may change when we had other samples taken - the smaller the SE the more precise. The coefficients become T-value divided by SE; greater value of T-value has indicated the prominence of the variable. P-value indicates how likely the variable has no actual influence (it is just a matter of chance). When the P-value is below 0.05 we tend to state that the variable is statistically significant i.e. it has actual influence in the model. Thus, considering all these three values together, we can make a decision as to what variables are important and which will not be needed.

i)Standard Error: The standard error (SE) of a related matrix of coefficient indicates the exactness of the coefficient in reality. It informs us the extent of the value of a coefficient that may change in case we repeat the regression to be performed in the same dataset using different variables. When SE is small, then it is likely that the coefficient is valid and it is not going to fluctuate much. However, when SE is big the coefficient may vary significantly with other data. That is why it is not more reliable. Mathematically, SE can be traced to the disparities between the observed and the predicted values (so-called residuals), as well as the dispersion and interrelation of the predictor variables.

$$SE = \sqrt{\sigma^2[(X^T X)^{-1}]_{jj}} \ ; \quad (3.19)$$

here, $\sigma^2 = {}^{SSR}/_{n-p}$

ii)T-value: To track reliability of coefficient, t-value is applied. It tells whether the coefficient is not equal the zero; The value of t in tells us how well a variable relates to the outcome. It is that we are attempting to forecast. It is obtained through the removal of standard error of a variable divided by the coefficient. Speaking in layman terms, it indicates the number of times larger the effect ought to be relative to the degree of a given uncertainty in that estimation. Leading to a high t-value (regardless of being positive or negative) implies that we are more sure that the variable actually does impact. When the value of t is low, near zero, the variable may not be very much important and may be simply noise. To determine whether result is significant or not we usually compare the t-value with some standard cutoff (e.g. 2 or -2). The greater the t- value, the higher is the possibility of the variable to be important in the model. The t-value therefore assists us to determine to which predictors we can have confidence and to which we may be willing to disregard. The formula for t-value-

$$T = \frac{Regression\ coefficient}{Standard\ error} \tag{3.20}$$

Larger values indicate more likely significant.

iii)P-value: When p-value is less than 0.05, then it has significant effects on Y. The p- value assists us in deciding whether a variable in a regression model does impact or the finding was merely a matter of chance. It goes hand in hand with the t- value and informs us of the ability to be certain that a certain relationship is real between the predictor and the outcome. When p is small (i.e. less than 0.05) then it is very unlikely that the outcome is non-random, and, therefore, we tend to say that variable is statistically significant. However, when the p-value is large then this is an indication that the use of the variable may not be so important; there is a high possibility that, its effect may simply arise by chance in the data. Shortly stated, p-value is an instrument which assists us in identifying which variables are worth hearing about or not, in a regression model.

3.7 Verification of fitness of the mathematical equations

When developing a regression model or designing a mathematical equation to describe a system, one should test the efficiency of the equation in terms of portraying the relationship between input variables (independent variables) and the output (dependent variable). At this point, ANOVA (Analysis of Variance) comes in handy. ANOVA, i.e., Analysis of Variance, is a technique through which it is verified that a mathematical equation or model is sufficient to explain the data. It assists in making us know whether the output (result) variation is actually dependent on the input variables or whether it occurred by mere coincidence. In simple terms, it will answer the question, which is: Is this model meaningful and reliable? This is done by ANOVA by comparing the adequacy of the model in explaining the portion of the overall variance in the data and the portion that is unexplained (error). When the model explains most of the variation and the p value is below 0.05, then the model can be termed statistically significant. It implies that the equation fits well and may be trusted to perform some prediction or analysis.

Parameters of ANOVA:

a)Sum of squares (SS): There are methodologies of understanding the variation in the data which is known as the Sum of Squares. In regression and ANOVA, it demonstrates the differences between observed values of data and the predicted or mean values. The Total Sum of Squares (SST) is what informs us of the extent to which the values differ with the overall average. This total variation is split into two parts, the Sum of Squares (SSR) indicates how much of the variation is covered by the model, and the Residual Sum of Squares (SSE) tells how much of it the model failed to cover (the errors). These are connected through the following formula, SST = SSR + SSE. When SSR is small and SSE is large then there is a good fit between data and the model. Thus, examining these values, we may realize the quality of the model and its answerability to the results.

Total sum of squares:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.21}$$

Here, y_i is actual value and $\bar{y}$ is mean value of all y_i .

Regression sum of squares:

$$SSR = \sum_{i=1}^{n}(\bar{y_i} - \bar{y})^2 \tag{3.22}$$

Here, y_i is predicted variables.

Residual sum of squares:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y_i})^2 \tag{3.23}$$

b)Degree of freedom (DF): Degrees of Freedom (DF) is a method through which values free to be changed are counted in data. It gives us the number of values in calculation that are free to change so that they can satisfy some conditions. We have like the number of independent data of dataset. DF in ANOVA assists in decomposing the source of variation in the data, namely, some part of it due to the model, and some due to the error. It is also used in calculations such as mean squares, or to test whether the model results have any meaning or are simply by chance, and in tests (such as the F-test).

Total Degrees of Freedom:

$$DF_{total} = n\text{-}1 \tag{3.24}$$

In this case, It is the total amount of observations decreased by 1. This is how one can start analysis of variability.

Regression Degrees of Freedom:

$$DF_{regression} = p \tag{3.25}$$

In this case, p represents the amount of predictors of the model (terms). This indicates the number of model terms utilized in interpreting the data.

Residual Degree of Freedom:

$$DF_{residual} = n\text{-}p\text{-}1 \tag{3.26}$$

This means the number of values remaining after the model has been accounted.It indicates the remaining freedom allowing the estimation of error or noise in the model.

Lack-of- Fit Degree of fitness:

$$DF_{Lof} = DF_{residual\ error} - DF_{pure\ error} \tag{3.27}$$

This informs us on the number of degrees of freedom associated with an inappropriate model. The high DF of lack-of-fit may be used to check whether the model is incomplete.

Pure Error Degrees of Freedom (DF_PureError):

It is due to repeated measurements in the same variable. It aids to deconfound random error (measurement noise) and systematic error (bad model fit).

DF is used in determining Mean Squares (MS), F-values and p-values that can be utilized to provide answers to whether or not your model and model terms are significant. It also provides fair comparisons particularly between models which have different number of predictors.

c)Mean Squares(MS): Mean Square (MS) is a statistical value and it is used to quantify the average of the amount of variation of particular source, such as a model or an error. It is obtained

by dividing the Sum of Squares (SS) and its Degrees of freedom (DF). Regression or ANOVA may often have two simple types, Mean Square of Regression (MSR) which displays an average variation explained by the model and Mean Square of Error (MSE) which displays an average variation that could not be explained by the model (the residuals). These values are significant since they are used to calculate the F-ratio which is used to test the importance of the model as well as whether it is significant or not.

$$MS = \frac{Sum\ of\ square(SS)}{Degree\ of\ freedom(DF)} \tag{3.28}$$

for MSR = Regression MS = SSR/$DF_{regression}$

MES = Residual MS = SSE/$DF_{residual}$

d)F-ration: The F-ratio stands as a statistic value that is applied to compare two kinds of variations: the one explained by an apparent model and the umxplained one (the error). It is computed as the ratio between the mean square of model and mean square of the error. When F-ratio is significantly larger than 1, it implies that model accounts by far more variance than remainder left as error and this normally implies that the model is quite useful. The small F-ratio (it is close to 1) implies that the model may not be performing better than random guessing. F-ratio is applied to find out whether there is a need to reject a null hypothesis in a regression or ANOVA analysis.

$$F = \frac{MS_{regression}}{MS_{residual}} \tag{3.29}$$

e)P-value: The p-value assists us in determining whether the observed results in our data are the actual results or they are mere coincidences. In ANOVA, we are informed about the probability of achieving equal results in the case the model would not have a true effect at all. When the p-value is small (typically smaller than 0.05), the model or the variable is very likely important and really effective. When the p-value is big, it implies that the result would have occurred as a chance like event and the model or variable may not be important. The p-value therefore assists us in determining whether we should believe what the model is saying or not.

->If $p < 0.05$: model or term is statistically significant

->If $p > 0.05$: not significant (may be due to chance)

f)R^2(coefficient of determination): The number R^2 is an indication of the accuracy of a regression model to explain the variability of the outcome (dependent variable). It gives the information on how the variation in the result can be explained by the input variables of the model. Values of R^2 will vary between 0 and 1; the higher it can be up to 1 the better is the fit with the data. But what about a high R^2 not a good model nor that one thing causes another. In order to get a clear picture of how good a model is, however, we must also consider other measures such as adjusted R^2, p-values as well as error patterns.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (3.30)$$

->R^2 = 1: perfect fit

->R^2 = 0: model explains nothing

->R^2 = 0.8 → model explains 80% of variance

g)Adjusted – R^2: Adjusted R^2 is a modification of R^2 which considers the number of predictors apply in a regression model. Although R^2 will always increase as we introduce more variables, including those that will not be of benefit, adjusted R^2 will increase only when the new variable passes the test of significance by improving the model. It presents a superior and more reasonable quantification of the fit of the model to the data particularly when it involves a large number of variables. This prevents overfitting and also facilitates a comparison of the models with different values of predictors..

$$\boldsymbol{R^2_{adj}} = 1 - \left(\frac{\frac{SSE}{DF_{res}}}{\frac{SST}{DF_{total}}}\right) = 1 - (1 - R^2)\frac{n-1}{n-p-1} \quad (3.31)$$

Adjusted R^2 can decrease if new terms don't improve the model.

h)PRESS (Prediction Error Sum of Squares): PRESS (Prediction Error Sum of Squares) indicates the skill of a regression model to predict novel data. It is calculated by omitting every data point in turn, then using the remaining data to build the model, and then predicting the left-out point. The squared difference in the actual and the predicted values are added together. The lower PRESS, the better it is predicted in the model and also works on new unseen data. It aids in detecting overfitting as well as comparisons between models.

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}_i}{1 - h_{ii}}\right)^2 \quad (3.32)$$

g)S (Standard Error of Regression): S (Standard Error of Regression) is a value which demonstrates the extent that the tested data are varying with the expected values of a regression formula. It is the mean distance of the observed value with the regression line. The smaller the S value the more the data points fit the line thus a good fit to the model. It is provided in the units of response variable and assists in assessing the correctness of model forecasts. In simple words, S explains how well the regression line is indicative of the data.

$$S = \sqrt{\frac{\text{SSE}}{\text{DF}_{\text{res}}}} = \sqrt{\text{MSE}} \quad (3.33)$$

With this term we will ANOVA fitness test of our developed model. ANOVA assists to compute the fitness of our model. Based on various Anova terms, we will be in a position to know more of model of lung cancer & breast cancer.

3.8 Testing

The process of testing is also relevant when it comes to the construction of any mathematical model. Once we have trained a model and have it perform on a particular subset of data (called the training set), we should examine how it does on new data that it is encountering for the first time. This new information is referred to as a test set. The primary objective of testing is whether the model is able to make the right predictions using the new, not observed data, as it should in the real world. In testing, we will use the input values contained in the test set to give the model and see how accurate its predictions are against the real outcomes. When the forecasts are near the actual solutions, then the model has acquired good patterns during the training exercise. However, on the off chance that it does not perform all that well, then this could be because the model has learnt incorrectly or it has just committed the training data rather than portraying the way the data relates to one another- this is known as overfitting. Through testing, we can determine the accuracy, reliability, and generalizability of the model. This assists us in determining whether the model is ready to be implemented in real-life scenarios or whether it requires further enhancement. Briefly speaking, testing will make sure that the model is able to really solve the problem it should solve and not just repeat what it has been trained to see.

3.8.1 Program to calculate the output from the developed equation

This is a program that is constructed to compute the output or the predicted result using a mathematical expression, which was formulated in the model development process. It is written in the form of an equation that depends on the pattern observed in the training data, and it includes fixed values written as coefficients and input variables (or otherwise called features). These coefficients indicate the extent and the orientation of the relationship between every variable and the outcome. This is where, in this program, we feed it new figures of the variables, it plugs them into the equation, and gives a prediction. The process is vital in that it gives us an opportunity to run the developed model on real or test data and see what comes out of it. With the use of this program, we are in a position to test whether the equation is functioning appropriately and providing relevant results in new cases. It is particularly helpful to perform predictions in applications in the real world (like disease detection, sales forecasting, or risk analysis). It also allows us to make sure the model is performing as it is supposed to, and not only working with the training data it was exposed to, but can also operate with new, unseen data correctly.

Code for lung cancer:

```
# Function for output based on range

def classify_output(y):

    if 0.0 <= y <= 0.49: return 0

    elif 0.5 <= y <= 1.49: return 1

    elif 1.5 <= y <= 2.49: return 2

    else:

        return "-"

# Function to calculate first-order model

def first_order(X):

    coeffs = [0.0863, 0.05512, 0.0839, 0.0629, 0.05057, 0.0199, 0.0638, -0.1105, -0.0366, -0.00248, 0.10171, 0.0357]

    intercept = -0.7361

    Y = intercept + sum(c * x for c, x in zip(coeffs, X))

    return Y

# Function to calculate second-order model

def second_order(X):

    a = [-0.385, 0.3581, -0.3454, 0.4505, -0.0375, 0.5708, 0.5287, -0.2208, -0.0429, -0.0739, 0.1103, 0.0621]

    b = [0.08705, -0.03338, -0.04663, -0.04133, 0.01314, -0.02630,  -0.06584, -0.0410, -0.00851, -0.01069, -0.01135, -0.0019]

interactions = { (0, 2): -0.06383, (0, 9): 0.03402, (2, 3): -0.05441, (2, 5): 0.08858, (2, 7): 0.1091, (2, 9): 0.1544, (3, 9): 0.0163, (5, 9): -0.1639, (7, 9): -0.0089, (7,11): 0.0065 }

   intercept = -1.581

    Y = intercept

    Y += sum(a[i] * X[i] for i in range(12))

    Y += sum(b[i] * (X[i] ** 2) for i in range(12))

    for (i, j), coef in interactions.items():

        Y += coef * X[i] * X[j]
```

```
    return Y

# Main Program

try:

    user_input = input("Enter 12 values for X1 to X12 separated by spaces: ")

    X = list(map(float, user_input.strip().split()))

    if len(X) != 12:

        print("enter exactly 12 values.")

    else:

        y1 = first_order(X)

        y2 = second_order(X)

        class_y1 = classify_output(y1)

        class_y2 = classify_output(y2)

        print(f"\nFirst-Order Model customize output: {class_y1}")

        print(f"Second-Order Model customize output: {class_y2}")

except ValueError:

    print("-")
```

Code for breast cancer:

```
def calculate_single_order(values):

    v = values

    return (0.182 - 0.864 * v[0] + 0.0651 * v[1] + 0.002579 * v[2] - 8.53 * v[3] + 0.82 * v[4] + 10.85 * v[5] + 2.57 * v[6] + 0.6473 * v[7] - 0.01339 * v[8] - 0.003089 * v[9] + 1.419 * v[10] + 0.120 * v[11] + 1.32 * v[12])

def calculate_polynomial(values):

    v = values

    return (7.87 - 5.08 * v[0] + 0.916 * v[1] - 0.02523 * v[2] - 14.87 * v[3] + 9.85 * v[4] + 21.69 * v[5] + 7.71 * v[6] - 1.015 * v[7] - 0.1278 * v[8] + 0.02294 * v[9] + 0.10 * v[10] - 1.16 * v[11] - 7.37 * v[12] + 1.143 * v[0]**2 + 0.01807 * v[1]**2 - 0.00001103 * v[2]**2 - 10.7 * v[3]**2 - 19.4 * v[4]**2 + 17.8 * v[5]**2 - 16.0 * v[6]**2 -
```

```
0.0584 * v[7]**2 - 0.002169 * v[8]**2 + 0.00000738 * v[9]**2 - 3.14 * v[10]**2 - 1.27 * v[11]**2 + 57.3 * v[12]**2 - 0.304 * v[0] * v[1] + 0.000353 * v[1] * v[2] + 0.0100 * v[1] * v[3] - 30.5 * v[3] * v[4] +  23.0 * v[3] * v[10] + 62.3 * v[4] * v[5] - 171.7 * v[5] * v[12] + 0.0403 * v[7] * v[8] - 0.001548 * v[7] * v[9] - 0.0000721 * v[8] * v[9] + 1.25 * v[10] * v[11] + 8.21 * v[11] * v[12])

def classify(result):

    if 1.5 <= result <= 3.5:

        return 2

    elif result > 3.5:

        return 4

    else:

        return result

# Input

raw_input = input("Enter 13 space-separated values: ")

values = list(map(float, raw_input.strip().split()))

if len(values) != 13:

    print("Error: Please enter exactly 13 values.")

else:

    y_single = calculate_single_order(values)

    y_poly = calculate_polynomial(values)

    print(f"\nSingle Order Output: {y_single:.4f}, Classified: {classify(y_single)}")

    print(f"Polynomial Output: {y_poly:.4f}, Classified: {classify(y_poly)}")
```

3.8.2 Comparison and final mathematical equation selection

Having come up with several mathematical equations, including linear (first-order) and polynomial (second-order) models of breast cancer and lung cancer .it is worth comparing the performance between each model in order to determine the better one. This comparison is made by such evaluation measures as accuracy, R-squared (R^2), mean squared error (MSE) or other statistical figures. These assist us in seeing how the various models predict the outcome, as well

as whether the given model captures the actual pattern in the data. This step will be to find which of the equations produces the most reliable and accurate result when used with the real data or one that cannot be seen. We pay a lot of attention to the performance of each model and how simple or complex (under- or overfitting) it is, and how effectively it can be applied in previously unseen cases. Judging by this comparison, we would choose the most successful equation as the last mathematical model. The chosen equation will then be used to continue the analysis, prediction work, or life application, such as the detection of cancer or the risk prediction, since the equation indicates the highest potential of providing accurate and repeatable findings.

Chapter 04
Results and Discussion

4.1 introduction

This chapter represents and comments on the research findings. In Section 4.2, we explain the preprocessing procedure undertaken on lung cancer data, and give the finding associated with the given. In the same way, Section 4.3 discusses the breast cancer data preprocessing and results. In section 4.4, the process of calculating the coefficients of the lung cancer model has been highlighted, and the resulting mathematical equation, along with the coefficients of the equation, is presented. The breast cancer model is set in Section 4.5 in the same way. Section 4.6 will check the fitness of the single-order and polynomial models constructed for lung cancer. The same verification is done in section 4.7 on the models of breast cancer. Section 4.8 is concerned with model testing. In particular, Section 4.8.1 provides the results of the tests conducted on the lung cancer model based on the test data set, whereas Section 4.8.2 contains the results of the testing behaved on breast cancer model.

4.2 Correlation matrix for lung cancer

In this case we attempt to identify meaningful variable in our data set. Due to the huge amounts of data causing complexity to model prediction. After calculating our 23 predictor and 1 response variable, we got this…

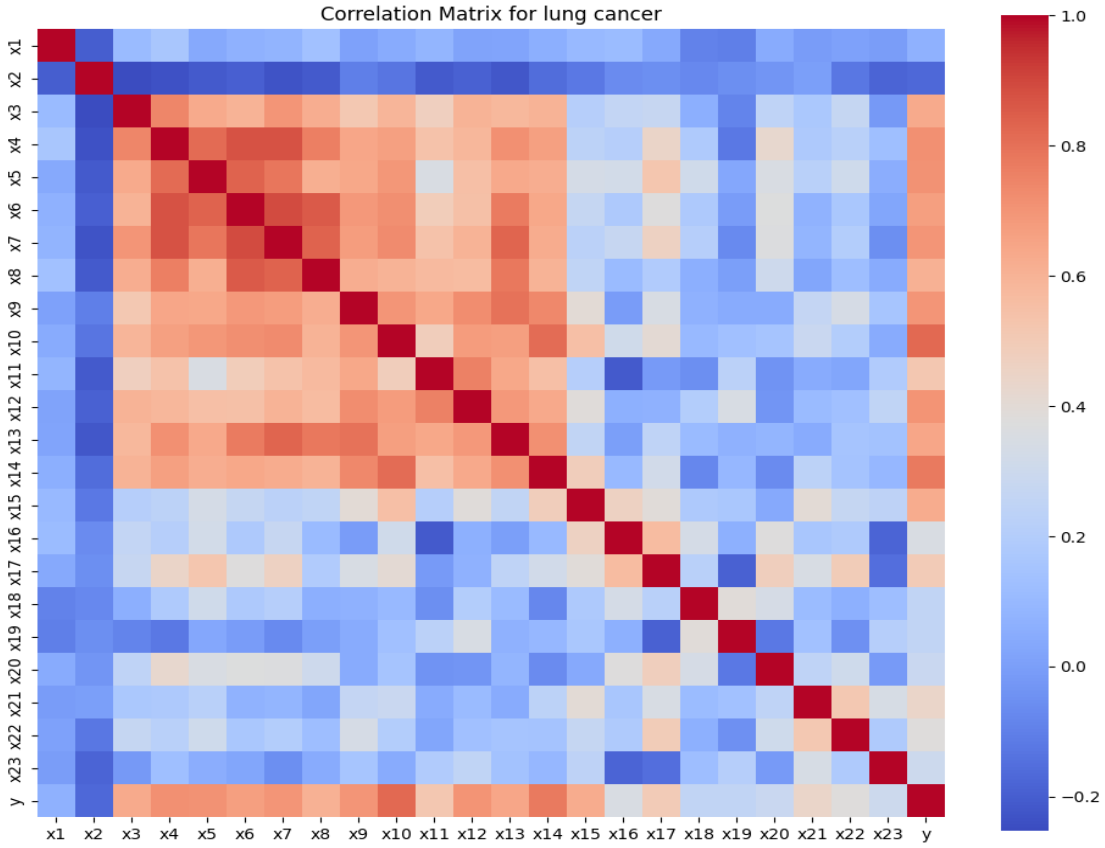


Fig 4.1: correlation matrix heatmap for lung cancer

This heatmap describe the correlation coefficient all the pairs (x_1 – x_{23} and y). Now see about the figure-

i. **Dark Red:** It located at diagonal line which represent self-correlation. That's why value is equal to 1.
ii. **Red:** it located at right side which represent strong positive correlation. Values of red zone is **(0.5 – 1)**. X_3 to x_{10} are in this cluster.
iii. **Blue:** It represents negative correlation. Values of this zone is **(-0.5 to -0.2)**. Variables like x1, x2, x16, x17, x20, and x22 show blue shades with others and with y. They have inverse relationship with cancer risk.
iv. **White or light shades:** It represent no linear relationship. Values of this region mostly 0.

Correlation coefficient with y:

Table 4.1: Correlation coefficient for lung cancer

Variable	X_{10}	X_{14}	X_4	X_5	X_{12}	X_9	X_7	X_6	X_{13}	X_3	X_{15}
Correlation with y	0.823	0.777	0.716	0.710	0.703	0.700	0.698	0.669	0.643	0.630	0.626
Category	strong	strong	strong	strong	strong	strong	moderate -high	moderate	moderate	moderate	moderate

Variable	X_8	X_{11}	X_{17}	X_{21}	X_{22}	X_{16}	X_{23}	X_{20}	X_{19}	X_{18}	X_1	X_2
Correlation with y	0.607	0.517	0.497	0.440	0.378	0.353	0.297	0.285	0.254	0.252	0.068	-0.171
Category	moderate	moderate	moderate	moderate	low	low	low	low	low	low	Very-low	Very-low

In order to create a practical and stream-lined model, we chose important variables after examining their correlation with the target variable Y. Correlation coefficient demonstrates the intensity of correlation between each variable and the outcome. A correlation value that is nearer to 1 or -1 is a stronger correlation whereas a value that is close to 0 shows little or no correlation. Our data contained 23 predictive variables. We used the correlation matrix to determine the relationship each of the variables has with Y and picked the variables that had the correlation value of 0. 60 or more. We have therefore reduced it to 12 notable variables that have higher chance of adding much of a value to the model. By turning to these variables that are highly correlated with each other, we are less complex, we are not over fitting, and our model can be more successful and can lead to us making the correct predictions. Selected parameters for model are-

Table 4.2: Selected parameter for Lung Cancer

Selected parameter	X_{10}	X_{14}	X_4	X_5	X_{12}	X_9	X_7	X_6	X_{13}	X_3	X_{15}	X_8
Represented by	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}

4.3 Correlation matrix for breast cancer

In this case, we attempt to identify the presence of significant variables based on our data through the use of correlation matrix. Due to the fact that large amount of data makes it difficult to predict model. After calculating our 30 predictor and 1 response variable, we got this…

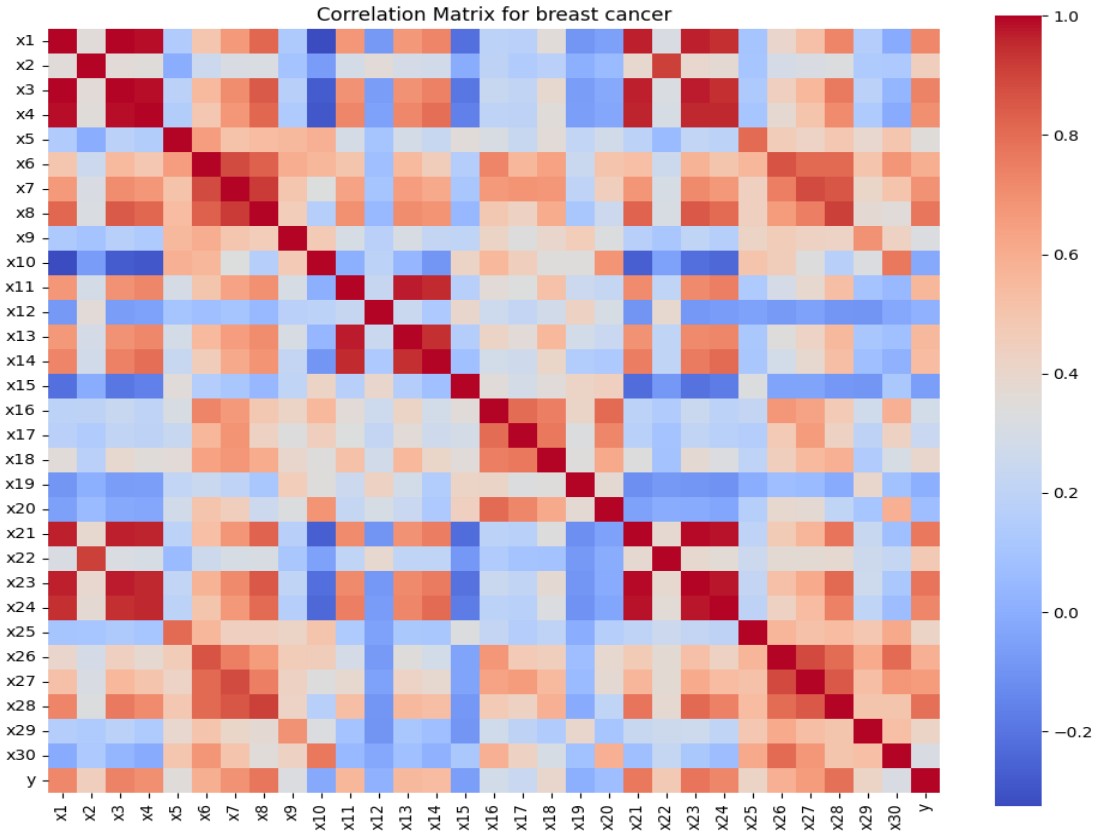


Fig 4.2: correlation matrix heatmap for breast cancer

This heatmap describe the correlation coefficient all the pairs (x_1 – x_{30} and y). Now see about the figure-

i. **Dark Red:** It located at diagonal line which represent self-correlation. That's why value is equal to 1.
ii. **Red:** it located at right side which represent strong positive correlation. Values of red zone is **(up – 1)**. x1, x3, x4, x8, x23, x24, x27 – these variables are near to deep read.
iii. **Blue:** It represents negative correlation. Values of this zone is **(down to -0.3/lower)**. Variables like x10, x11, x12 are near to deep blue shades with others and with y. They have inverse relationship with cancer risk.
iv. **White:** Values of this region mostly 0. These values x15, x19, x22, x29, x30 are near white. These variables have little influence in y.

Correlation coefficient with y:

Table 4.3: Correlation coefficient for lung cancer

Variable	X_{28}	X_{23}	X_8	X_{21}	X_3	X_{24}	X_1	X_4	X_7	X_{27}	X_6	X_{26}	X_{11}	X_{13}	X_{14}
Correlation with y	0.791	0.778	0.773	0.771	0.737	0.728	0.724	0.702	0.692	0.658	0.591	0.589	0.563	0.551	0.54
Category	strong	strong	strong	strong	strong	strong	strong	strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate

Variable	X_{22}	X_2	X_{25}	X_{29}	X_{18}	X_5	X_9	X_{30}	X_{16}	X_{17}	X_{20}	X_{12}	X_{19}	X_{10}	X_{15}
Correlation with y	0.478	0.447	0.417	0.415	0.397	0.348	0.322	0.315	0.234	0.243	0.071	0.012	0.004	-0.018	-0.056
Category	low	low	low	low	low	low	low	low	Very-low	Very-low	Very-low	Very-low	Very-low	Very-low	Very-low

In order to construct a good model, we should concentrate on those variables which are related to another variable (Y) in the best way possible. To do this, what we employ is a correlation matrix where we indicate how well each variable relates to Y. Those variables that have a higher correlation value towards 1 or -1 are then said to be more important as they have a better effect on the output. To give our dataset an idea, we had initially 30 predictive variables. Having used the correlation matrix, we picked the 13 variables that had the correlation value of up to 0.56 with Y. The variables were selected in this manner because the more they will be able to do something important to the model, and those less correlated have been discarded so as not to add complexity to this. We reduce the model to the minimum by choosing only the most considerable features so as to have a greater opportunity of making the correct predictions. Our selected parameter now-

Table 4.4: Selected parameter for Breast cancer

parameter	X_{28}	X_{23}	X_8	X_{21}	X_3	X_{24}	X_1	X_4	X_7	X_{27}	X_6	X_{26}	X_{11}
Represented by	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}

4.4 Coefficient calculation for single-order model and polynomial model for lung cancer:

In order to create a proper predictive model of lung cancer, we estimated the coefficients of our first-order (linear) and second-order (polynomial) equations established earlier. To do this, we applied the Response Surface Methodology (RSM) and the manual methods of calculation(python). RSM assisted us in understanding the dependence between the input variables and the output by fitting a mathematical formulation whereas the process of going through the steps was able to help us understand and achieve better accuracy of results through manual calculation. In this mixed strategy, we were able to arrive at values of the coefficients relating the coefficients defining the impact of each variable in the model. These coefficients are very crucial since they determine the end equations that will be utilized in prediction and analysis in detection of lung cancer.

Table 4.5: Coefficient for lung cancer model

Polynomial model				single-order model			
Term	**Coef**	**T-Value**	**P-Value**	**Term**	**Coef**	**T-Value**	**P-Value**
Constant	-1.581	-12.40	0.000	Constant	-0.7361	-19.88	0.000
X1	-0.3852	-8.79	0.000	X1	0.0863	6.65	0.000
X2	0.3581	13.08	0.000	X2	0.05512	5.55	0.000
X3	-0.3454	-7.53	0.000	X3	0.0839	7.03	0.000
X4	0.4505	6.16	0.000	X4	0.0629	5.09	0.000
X5	-0.0375	-1.10	0.270	X5	0.05057	6.12	0.000
X6	0.5708	10.91	0.000	X6	0.0199	1.97	0.050
X7	0.5287	13.75	0.000	X7	0.0638	3.71	0.000
X8	-0.2208	-3.69	0.000	X8	-0.1105	-5.61	0.000
X9	-0.0429	-1.29	0.197	X9	-0.0366	-3.02	0.003
X10	-0.0739	-1.77	0.077	X10	-0.00248	-0.26	0.792
X11	0.1103	5.52	0.000	X11	0.10171	15.86	0.000
X12	0.0621	1.45	0.147	X12	0.0357	2.34	0.019
X1*X1	0.08705	9.47	0.000				
X2*X2	-0.03338	-10.76	0.000				
X3*X3	-0.04663	-4.68	0.000				
X4*X4	-0.04133	-4.54	0.000				
X5*X5	0.01314	3.03	0.003				
X6*X6	-0.02630	-3.23	0.001				
X7*X7	-0.06584	-12.25	0.000				
X8*X8	-0.0410	-2.34	0.020				
X9*X9	-0.00851	-1.98	0.049				
X10*X10	-0.01069	-1.57	0.118				
X11*X11	-0.01135	-5.91	0.000				
X12*X12	-0.0019	-0.16	0.874				
X1*X3	-0.06383	-7.63	0.000				
X1*X10	0.03402	4.54	0.000				
X3*X4	-0.05441	-5.83	0.000				

X3*X6	0.08858	10.54	0.000	
X3*X8	0.1091	5.15	0.000	
X3*X10	0.1544	9.46	0.000	
X4*X10	0.0163	1.06	0.288	
X6*X10	-0.1639	-15.15	0.000	
X8*X10	-0.0089	-0.59	0.557	
X8*X12	0.0065	0.33	0.744	

Single-order model

Based on the findings, it appears that the overall model variables are slightly significant except some variables whose t-values are rather small as opposed to that or p-values, which are considerably low suggesting $p < 0.05$ (statistically significant). To be more precise, highly important predictors are included in X1, X2, X3, X4, X5, X7, X8, X9, X11 and X12 variables, since their t-values are significantly higher than 2 (or less than -2), with p-values being close to zero. It has the implication that name and character of these variables really do influence the response variable. It is noted that the t-value on the constant term is also large and the p-value is 0.000 which will mean that the constant term too is significantly different than zero so the model is of use. The t-value 1.97 and p-value equals 5.00 is barely out of line to be described as borderline. it is coming just out of the threshold of significance 5.00 as the norm. Even though this has a slight effect, the weight of the variable is low as far as its influence is considered relative to other variables.

After all, now our equation is look like-

$$Y = -0.7361 + 0.0863X_1 + 0.05512X_2 + 0.0839X_3 + 0.0629X_4 + 0.05057X_5 + 0.0199X_6 + 0.0638X_7 - 0.1105X_8 - 0.0366X_9 - 0.00248X_{10} + 0.10171X_{11} + 0.0357X_{12} \quad (4.1)$$

Polynomial model

With the regression outputs all the model terms are statistically significant except some of the terms. It was guided by t-values that were higher than or lower than the interval of 2 and the p-value that was less than the value 0.05. The main effects that are most important in the case like X1, X2, X3, X4, X6, X7, X8 and X11 have an exceptionally thick value of t and a p- value of 0.000 meaning that they are vital and important. Additionally there exist squared terms and their interaction with all others e.g. X1X1, X2X2, X3X3, X4X4, X5X5, X6X6, X7X7, X8X8, X9X9,

X11X11 and some of the product terms as in X1X3, X1X10 and some new ones like and also X3X4 and X3X6 and X3X8 and X3X10 and X6X10 so higher orders of interaction and the curve are There are also some of the interaction terms that are not significant e.g. X4X10, X8X10 and X8X12 and therefore will not add much to the model. X10X10 is in the borderline (p = 0.118) and X9X9 has reached the extreme end of scale (p = 0.049). All in all, the model appears to be largely influenced by both primary predictors and their higher-order (quadratic and interaction) components, and certain of the variables can be potentially eliminated in case they are making a negligible contribution to the model.

We preprocess our data then we found some significant quadratic terms and interaction terms. By using significant term, our calculation will be easy and we will find a fit model.

$Y = C_0+C_1X_1+C_2X_2+C_3X_3+C_4X_4+C_5X_5+C_6X_6+C_7X_7+C_8X_8+C_9X_9 +C_{10}X_{10} +C_{11}X_{11} +C_{12}X_{12} + C_1X_1^2+C_2X_2^2+C_3X_3+C_4X_4^2+C_5X_5^2+C_6X_6^2+C_7X_7^2+C_8X_8^2+C_9X_9^2+C_{10}X_{10}^2+C_{11}X_{11}^2+C_{12}X_{12}^2+ C_{13}X_1X_3+ C_{110}X_1X_{10}+C_{34}X_3X_4+ C_{36}X_3X_6+ C_{38}X_3X_8+ C_{310}X_3X_{10}+ C_{410}X_4X_{10}+ C_{68}X_6X_8+C_{610}X_6X_{10}+ C_{810}X_8X_{10}+ C_{812}X_8X_{12}$ (4.2)

After allour equation look like-

$$Y = -1.581 - 0.385X_1 + 0.3581X_2 - 0.3454X_3 + 0.4505X_4 - 0.0375X_5 + 0.5708X_6 + 0.5287\,X_7 - 0.2208X_8 - 0.0429X_9 - 0.0739X_{10} + 0.1103X_{11} + 0.0621X_{12} + 0.08705\,X_1 * X_1 - 0.03338X_2 * X_2 - 0.04663X_3 * X_3 - 0.04133X_4 * X_4 + 0.01314X_5 * X_5 - 0.02630X_6 * X_6 - 0.06584X_7 * X_7 - 0.0410X_8 * X_8 - 0.00851X_9 * X_9 - 0.01069X_{10} * X_{10} - 0.01135X_{11} * X_{11} - 0.0019X_{12} * X_{12} - 0.06383X_1 * X_3 + 0.03402X_1 * X_{10} - 0.05441X_3 * X_4 + 0.08858X_3 * X_6 + 0.1091X_3 * X_8 + 0.1544X_3 * X_{10} + 0.0163X_4 * X_{10} - 0.1639X_6 * X_{10} - 0.0089X_8 * X_{10} + 0.0065X_8 * X_{12}$$

(4.3)

4.5 Coefficient calculation for second-order model & first-order model for breast cancer:

We get the coefficient of our earlier developed single order and polynomial equation in the breast cancer, using surface response methodology and manual calculation of coefficient(python).

Table 4.6: Coefficient for breast cancer model

Polynomial -order model

Term	Coef	T-Value	P-Value
Constant	7.87	3.26	0.001
X1	-5.08	-3.48	0.001
X3	0.916	3.52	0.000
X4	-0.02523	-2.54	0.011
X6	-14.87	-2.08	0.038
X7	9.85	2.40	0.017
X8	21.69	3.02	0.003
X9	7.71	1.14	0.255
X21	-1.015	-1.62	0.106
X23	-0.1278	-1.79	0.074
X24	0.02294	4.64	0.000
X26	0.10	0.09	0.929
X27	-1.16	-0.94	0.348
X28	-7.37	-2.05	0.041
X1*X1	1.143	2.76	0.006
X3*X3	0.01807	2.13	0.033
X4*X4	-0.000011	-2.21	0.027
X6*X6	-10.7	-0.39	0.700
X7*X7	-19.4	-1.73	0.084
X8*X8	17.8	0.22	0.830
X9*X9	-16.0	-0.90	0.371
X21*X21	-0.0584	-0.98	0.329
X23*X23	-0.002169	-2.24	0.025
X24*X24	0.000007	4.82	0.000
X26*X26	-3.14	-0.89	0.372
X27*X27	-1.27	-0.97	0.335
X28*X28	57.3	2.28	0.023
X1*X3	-0.304	-2.56	0.011
X3*X4	0.000353	2.13	0.034
X3*X6	0.0100	0.14	0.886
X6*X7	-30.5	-0.92	0.356
X6*X26	23.0	1.86	0.063
X7*X8	62.3	1.26	0.208
X8*X28	-171.7	-3.10	0.002
X21*X23	0.0403	2.64	0.009

Single -order model

Term	Coef	T-Value	P-Value
Constant	0.182	0.44	0.657
X1	-0.864	-2.65	0.008
X3	0.0651	1.34	0.179
X4	0.002579	2.93	0.003
X6	-8.53	-4.15	0.000
X7	0.82	0.49	0.626
X8	10.85	3.27	0.001
X9	2.57	2.46	0.014
X21	0.6473	7.27	0.000
X23	-0.01339	-1.40	0.162
X24	-0.003089	-6.25	0.000
X26	1.419	2.69	0.007
X27	0.120	0.23	0.819
X28	1.32	0.95	0.341

X21*X24	-0.001548	-3.84	0.000	
X23*X24	-0.000072	-1.35	0.178	
X26*X27	1.25	0.35	0.728	
X27*X28	8.21	0.93	0.353	

Single -order model

The main effects that are really influential in this model are X1 (p=0.008), X4 (p=0.003), X6 (p<0.001), X8 (p=0.001), X9 (p=0.014), X21 (p<0.001), X24 (p<0.001), and X26 (p=0.007), meaning that they influence the response in a remarkable way. In particular, X1, X6, and X24 are negative, whereas X4, X8, X9, X21 and X26 are positive.

Our equation-

$$Y = 0.182 - 0.864\,X1 + 0.0651\,X3 + 0.002579\,X4 - 8.53\,X6 + 0.82\,X7 + 10.85\,X8 + 2.57\,X9 + 0.6473\,X21 - 0.01339\,X23 - 0.003089\,X24 + 1.419\,X26 + 0.120\,X27 + 1.32\,X28 \quad (4.4)$$

Polynomial model

The main effects that are significant in this model are X1, X3, X4, X6, X7, X8, X24, and X28, which imply that each of them has a significant (positive or negative) linear influence on the response variable, with the negative impact of X1, X4, X6, and X28, and the positive impact of X3, X7, X8, and X24. Of the quadratic terms, , $X1^2$, $X3^2$, $X4^2$, $X23^2$, $X24^2$, and $X28^2$ are important, indicating the existence of non linear curvature of these variables whereas of others such as $X6^2$, $X7^2$, and $X8^2$ are not important. In the case of interaction terms, X1X3, X3X4, X8X28, X21X23 and X21X24 based on their significance emerge as having interactions between any two variables. In general, the model indicates that the response is affected by the linear and nonlinear effects including significant curvature in X1, X3, X24, and X28 as well as the presence of major interaction between X1, X3, X8, X21 and X24.

We preprocess our data that time we found some significant quadratic terms and interaction terms. By using significant term, our calculation will be easy and we will find a fit model.

$$Y=D_0+D_1X_1+D_2X_2+D_3X_3+D_4X_4+D_5X_5+D_6X_6+D_7X_7+D_8X_8+D_9X_9+D_{10}X_{10}+D_1X_1^2+D_2X_2^2+D_3X_3+D_4X_4^2+D_5X_5^2+D_6X_6^2+D_7X_7^2+D_8X_8^2+D_9X_9^2+D_{10}X_{10}^2+D_{12}X_1X_2+D_{14}X_1X_4+D_{15}X_1X_5+D_{23}X_2X_3+D_{24}X_2X_4+D_{34}X_3X_4+\varepsilon \quad (4.5)$$

Now our equation look like-

$$Y = 7.87 - 5.08 X_1 + 0.916 X_3 - 0.02523 X_4 - 14.87 X_6 + 9.85 X_7 + 21.69 X_8 + 7.71 X_9 - 1.015 X_{21} - 0.1278 X_{23} + 0.02294 X_{24} + 0.10 X_{26} - 1.16 X_{27} - 7.37 X_{28} + 1.143 X_1*X_1 + 0.01807 X_3*X_3 - 0.00001103 X_4*X_4 - 10.7 X_6*X_6 - 19.4 X_7*X_7 + 17.8 X_8*X_8 - 16.0 X_9*X_9 - 0.0584 X_{21}*X_{21} - 0.002169 X_{23}*X_{23} + 0.00000738 X_{24}*X_{24} - 3.14 X_{26}*X_{26} - 1.27 X_{27}*X_{27} + 57.3 X_{28}*X_{28} - 0.304 X_1*X_3 + 0.000353 X_3*X_4 + 0.0100 X_3*X_6 - 30.5 X_6*X_7 + 23.0 X_6*X_{26} + 62.3 X_7*X_8 - 171.7 X_8*X_{28} + 0.0403 X_{21}*X_{23} - 0.001548 X_{21}*X_{24} - 0.0000721 X_{23}*X_{24} + 1.25 X_{26}*X_{27} + 8.21 X_{27}*X_{28} \quad (4.6)$$

4.6 Verification for single-order &polynomial model for lung cancer:

Table 4.7: Fitness test of single order mathematical equation for lung cancer

SOURCE	DF	SUM OF SQUARE	MEAN SQUARES	F- RATIO	P-VALUE
REGRESSION	11	509.105	46.2823	324.7221	1.1102e-16
LINEAR	11	509.105	46.2823	324.7221	1.1102e-16
RESIDUAL ERROR	967	137.8256	0.1425	-	
LUCK-OF-FIT	961	137.8251	0.1434	1692.6889	0.0
PURE ERROR	6	0.00051	8.4728e-05		
TOTAL	978	646.9315			

S= 0.3775, R^2 = 0.79, R^2- adjusted = 0.78, PRESS= 142.23

Table 4.8: Fitness test of polynomial mathematical equation for lung cancer

SOURCE	DF	SUM OF SQUARE	MEAN SQUARES	F- RATIO	P-VALUE
REGRESSION	32	596.540	18.641886	349.96630	1.110e-16
LINEAR	11	509.105	46.282354	324.72210	1.110e-16
SQUARE	11	4463.337	405.75791		
INTERACTION	10	3417.992	341.79920		
RESIDUAL ERROR	946	50.391	0.0532676		
LUCK-OF-FIT	946	50.390	0.0536070	582.30484	0.164
PURE ERROR	6	0.00055	9.2060162		
TOTAL	978	646.931			

S = 0.2307979, R^2 = 0.92, R^2-adjusted = 0.92, PRESS = 55.2340

Using the single-order model, the regression sources (all linear terms) compute on 509.11 units of the total sum of squares, with a mean square of 46.28 and a very large F-ratio of 324.72, which was entirely significant (p-value 40002E-00). The amount of the residual error is 137.83, of which

the lack-of-fit contributes 137.83 and the pure error is mere 0.00051, so nearly all residual error is associated with lack-of-fit (F = 1692.69, p-value = 0.0), meaning that the model might be lacking significant terms. The standard error (S) is 0.3775, R 2 is 0.79, and adjusted R^2 is 0.78 indicating that the model can explain around 79 percent of the variability in the response. The PRESS value high of 142.23, however indicates poor predictive ability on new data.

In the second-order model, the total regression sum of squares is raised to 596.54, the mean square is 18.64, and the F-ratio is significantly higher at 349.97 (again it indicates a strong significance of the model; the p-value 12 is very small 0). The linear terms value remains the lowest at 509.11, whilst square terms value is huge at 4463.34 and interaction terms depict an important final contribution 3417.99 and as such higher order and interaction effects are found to improve model performance greatly. the residual error decreases to 50.39 with the lack-of-fit but still 50.39 but with the p-value of 0.164 that indicates that the model fits the data. The standard error goes down on 0.2308, and R^2 increases to 0.92, and adjusted R^2 is also 0.92 meaning the model explains 92 percent of the variance. More predictive accuracy is displayed as well, dropping PRESS value to 55.23.

The polynomial model is admittedly better than the single-order model. It has square and interaction terms that enhance a significant fit and minimized error. Adjusted R^2 improves to 0.92, the standard error value falls and PRESS is almost one tenth of what it was at the beginning, which means the model is better at making predictions on the unknown data. And, most importantly, the lack-of-fit is not significant, or in other words the polynomial model fits the structure in the data and the single-order model is visibly underfit.

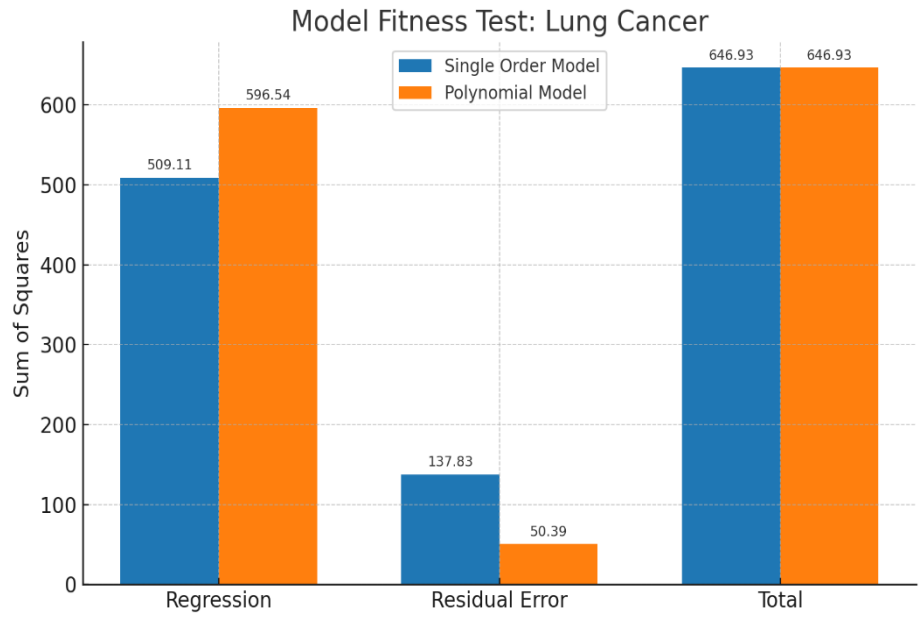


Fig 4.3 : Fitness test of lung cancer

4.7 Verification of single-order & polynomial model for Breast cancer:

Table 4.9: Fitness test of single order mathematical equation for Breast cancer

SOURCE	DF	SUM OF SQUARE	MEAN SQUARES	F- RATIO	P-VALUE
REGRESSION	15	389.38639	25.9590	110.313	1.1102e-16
LINEAR	15	389.38639	25.9590	110.313	1.1102e-16
RESIDUAL ERROR	533	125.42599	0.2353		
LUCK-OF-FIT	527	125.42540	0.2379	2404.43	0.0
PURE ERROR	6	0.00059	9.8983		
TOTAL	548	514.81239			

S = 0.4850988, R^2 = 0.76, R^2-adjusted = 0.75, PRESS = 134.141320

Table 4.10: Fitness test of polynomial mathematical equation for breast cancer

SOURCE	DF	SUM OF SQUARE	MEAN SQUARES	F- RATIO	P-VALUE
REGRESSION	40	428.7074	10.71768	63.23196	1.110e-16
LINEAR	15	389.3863	25.95909	110.3136	1.110e-16
SQUARE	15	34076.105368301e+8	227174037886.7		
INTERACTION	10	39935.07547e+8	399350754.77		
RESIDUAL ERROR	508	86.10494	0.169497		
LUCK-OF-FIT	502	86.10436	0.1715	1783.1477	0.164
PURE ERROR	6	0.00052	9.61909		
TOTAL	548	514.81239			

S = 0.4117012, R^2 = 0.83, R^2-adjusted = 0.82, PRESS = 94.725391

The single-order model had a reasonable fit (R^2 = 0.76, adjusted R^2 = 0.75) and explains over three-quarters (76 percent) of the variation in the response variable with a residual standard deviation S = 0.485 out of the total change in the response variable of 514.81. Its F-value (110.31 $p<0.0001$) affirms that the regression is very much significant. However, a lack of Fit test is significant (p=0.0) and it means that the linear model is not adequate to describe the structure of responses.

By contrast, the polynomial model fits much better, increasing the R^2 =0.83 (adjusted R^2 = 0.82), decreasing the residual variation (S = 0.412) and the PRESS statistic to 94.73, which means a greater predictive ability. The overall F-test remains highly significant (F = 63.23, $p<0.0001$) and

the lack-of-fit test indicates p=0.164 (not significant), the polynomial model therefore explaining the data without serious misspecification. To conclude, even though the single-order model is acceptable, it is evident that the polynomial model is significantly superior compared to the single-order model in explaining variance, predicting error, and eliminating significant lack-of-fit, and thus it is more desirable in terms of both linear and nonlinear relationships existing in the data.

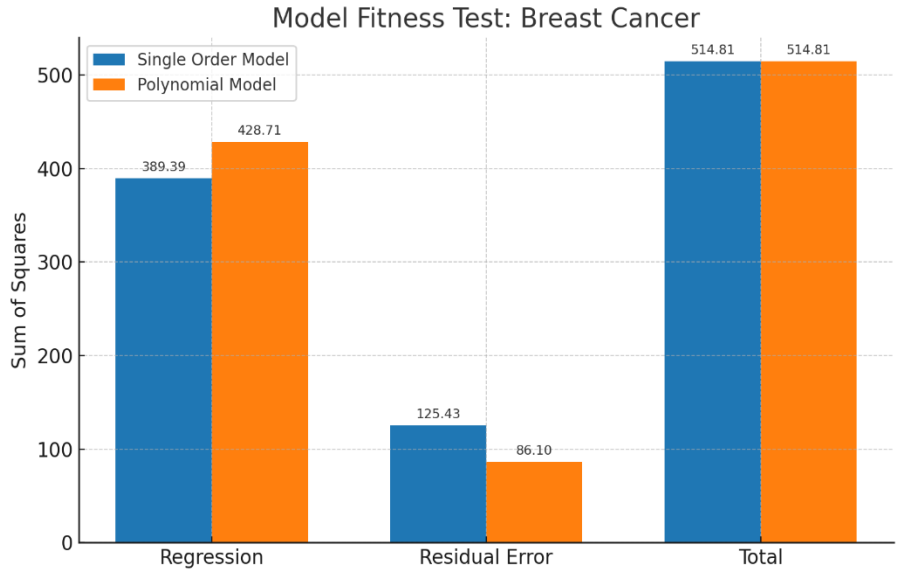


Fig 4.4: Fitness test of breast cancer

Final model for lung cancer & breast cancer for prediction is-

Lung cancer:

$$Y = -1.581 - 0.385X_1 + 0.3581X_2 - 0.3454X_3 + 0.4505X_4 - 0.0375X_5 + 0.5708X_6 + 0.5287X_7 - 0.2208X_8 - 0.0429X_9 - 0.0739X_{10} + 0.1103X_{11} + 0.0621X_{12} + 0.08705X_1 * X_1 - 0.03338X_2 * X_2 - 0.04663X_3 * X_3 - 0.04133X_4 * X_4 + 0.01314X_5 * X_5 - 0.02630X_6 * X_6 - 0.06584X_7 * X_7 - 0.0410X_8 * X_8 - 0.00851X_9 * X_9 - 0.01069X_{10} * X_{10} - 0.01135X_{11} * X_{11} - 0.0019X_{12} * X_{12} - 0.06383X_1 * X_3 + 0.03402X_1 * X_{10} - 0.05441X_3 * X_4 + 0.08858X_3 * X_6 + 0.1091X_3 * X_8 + 0.1544X_3 * X_{10} + 0.0163X_4 * X_{10} - 0.1639X_6 * X_{10} - 0.0089X_8 * X_{10} + 0.0065X_8 * X_{12}$$

(4.3)

Breast cancer:

$$Y = 7.87 - 5.08\,X_1 + 0.916\,X_3 - 0.02523\,X_4 - 14.87\,X_6 + 9.85\,X_7 + 21.69\,X_8 + 7.71\,X_9 - 1.015\,X_{21} - 0.1278\,X_{23} + 0.02294\,X_{24} + 0.10\,X_{26} - 1.16\,X_{27} - 7.37\,X_{28} + 1.143\,X_1*X_1 + 0.01807\,X_3*X_3 - 0.00001103\,X_4*X_4 - 10.7\,X_6*X_6 - 19.4\,X_7*X_7 + 17.8\,X_8*X_8 - 16.0\,X_9*X_9 - 0.0584\,X_{21}*X_{21} - 0.002169\,X_{23}*X_{23} + 0.00000738\,X_{24}*X_{24} - 3.14\,X_{26}*X_{26} - 1.27\,X_{27}*X_{27} + 57.3\,X_{28}*X_{28} - 0.304\,X_1*X_3 + 0.000353\,X_3*X_4 + 0.0100\,X_3*X_6 - 30.5\,X_6*X_7 + 23.0\,X_6*X_{26} + 62.3\,X_7*X_8 - 171.7\,X_8*X_{28} + 0.0403\,X_{21}*X_{23} - 0.001548\,X_{21}*X_{24} - 0.0000721\,X_{23}*X_{24} + 1.25\,X_{26}*X_{27} + 8.21\,X_{27}*X_{28}$$

(4.6)

4.8 Testing result

In order to test our first-order (linear) and second-order (polynomial) model performance, we utilized an additional 20 testing points. To come up with the valued predictions, we made use of the formulas that were developed in conjunction with the calculations of the coefficients in relation to every data point. These predicted values were then compared with the actual results that confirmed whether each of the predictions was correct or incorrect. Depending on such a comparison, we determined significant performance measures like accuracy (percentage of correct answers) and error (the difference between the forecasted and the actual values).

4.8.1 Lung cancer

In order to categorize the predicted values, we were to employ a condition forming on the level output range. Where level is in a range of 0.0-0.49, it is printed as 0; 0.5-1.49, it is printed as 1; 1.5-2.49, it is printed as 2, otherwise it is marked as "-". Applying this rule, we read the continuous prediction values as class labels, and we obtained a result table that contains the model prediction results including the prediction class, actual class, correctness, accuracy, and error.

Table 4.11: Test dataset for lung cancer

index	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	level
981	3	2	1	4	2	3	3	2	4	3	2	2	Low
982	7	8	5	6	3	6	5	5	4	4	8	4	High
983	7	9	8	7	7	7	7	7	7	6	3	6	High
984	7	7	7	7	7	7	7	7	7	6	2	6	High
985	3	3	2	4	2	3	3	2	3	3	4	2	Medium
986	3	4	2	3	2	3	2	4	3	1	4	4	Medium
987	7	7	7	7	8	7	7	7	7	6	5	6	High
988	7	7	8	7	8	7	7	7	7	6	9	6	High
989	7	8	5	6	3	6	5	5	4	4	8	4	High
990	7	9	8	7	7	7	7	7	7	6	3	6	High
991	7	8	7	7	7	7	6	7	7	7	4	7	High
992	7	8	5	6	3	6	5	5	4	6	8	4	High
993	7	9	8	7	7	7	7	7	7	8	3	6	High
994	7	7	7	7	7	7	7	7	7	7	2	6	High
995	7	8	7	7	7	7	6	7	7	7	4	7	High
996	7	7	7	7	7	6	7	7	7	6	8	7	High
997	7	7	7	7	8	7	7	7	7	6	5	6	High
998	7	7	8	7	8	7	7	7	7	6	9	6	High
999	7	8	5	6	3	6	5	5	4	4	8	4	High
1000	7	9	8	7	7	7	7	7	7	6	3	6	High

Table 4.12: Test the single-order model to predict lung cancer

SI NO	Expected value		Predicted value using mathematical equation		Correct/ incorrect	Accuracy	Error
	Value	level	value	level			
1	0	Low	0	Low	✓	95%	5%
2	2	High	2	High	✓		
3	2	High	2	High	✓		
4	2	High	2	High	✓		
5	1	Medium	1	Medium	✓		
6	1	Medium	0	Low	×		
7	2	High	2	High	✓		
8	2	High	2	High	✓		
9	2	High	2	High	✓		
10	2	High	2	High	✓		
11	2	High	2	High	✓		
12	2	High	2	High	✓		
13	2	High	2	High	✓		
14	2	High	2	High	✓		
15	2	High	2	High	✓		
16	2	High	2	High	✓		
17	2	High	2	High	✓		
18	2	High	2	High	✓		
19	2	High	2	High	✓		
20	2	High	2	High	✓		

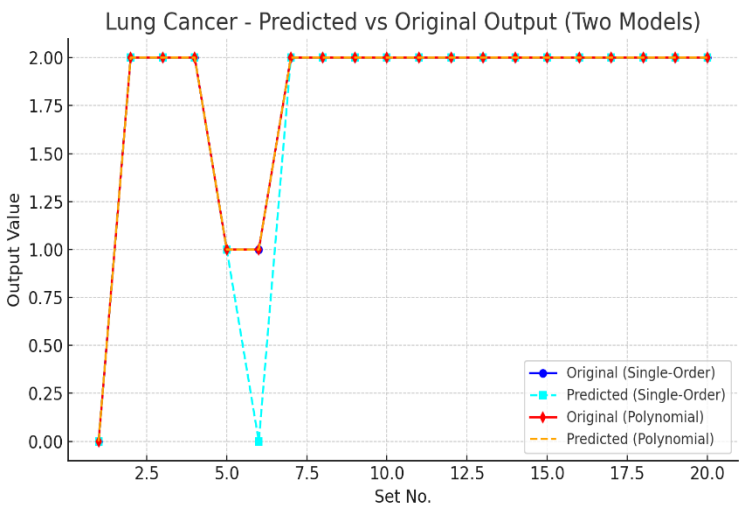
Table 4.13: Test the polynomial model to predict lung cancer

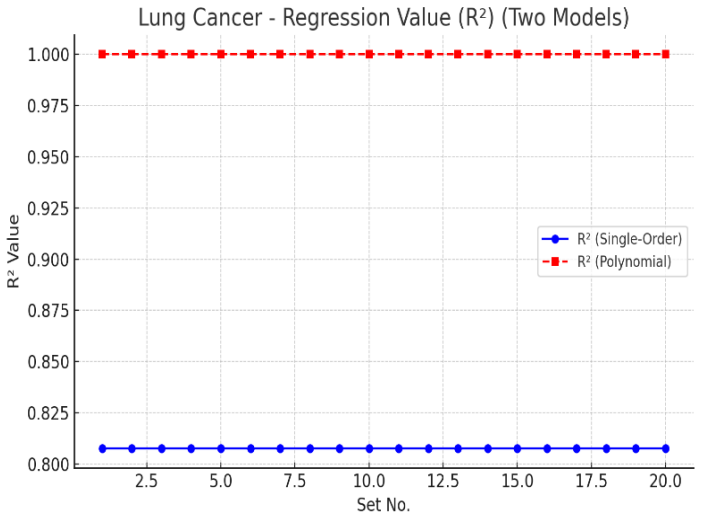
SI NO	Index	Expected value		Predicted value using mathematical equation		Correct/ incorrect	Accuracy	Error
		Value	level	value	level			
1	981	0	Low	0	Low	✓	100%	0%
2	982	2	High	2	High	✓		
3	983	2	High	2	High	✓		
4	984	2	High	2	High	✓		
5	985	1	Medium	1	Medium	✓		
6	986	1	Medium	1	Medium	✓		
7	987	2	High	2	High	✓		
8	988	2	High	2	High	✓		
9	989	2	High	2	High	✓		
10	990	2	High	2	High	✓		
11	991	2	High	2	High	✓		
12	992	2	High	2	High	✓		
13	993	2	High	2	High	✓		
14	994	2	High	2	High	✓		
15	995	2	High	2	High	✓		
16	996	2	High	2	High	✓		

17	997	2	High	2	High	✓		
18	998	2	High	2	High	✓		
19	999	2	High	2	High	✓		
20	1000	2	High	2	High	✓		

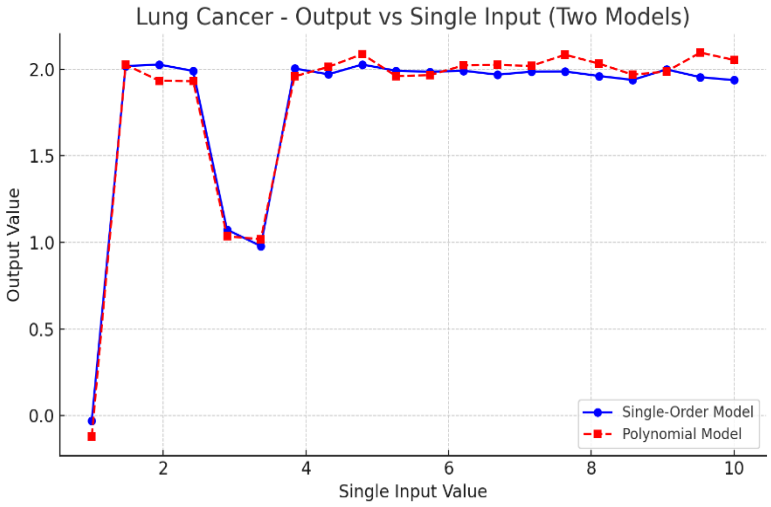
Among the single-order model (95% accuracy) and the polynomial model (100% accuracy), the model that showed to be more accurate was the polynomial model for lung cancer.

Results are shown in graph-





a) **b)**



c)

Fig 4.5: Testing output of lung cancer visualize by graph - **a)** Predicted vs Original Output vs Set No **b)**Regression Value (R^2) bar chart **c)**Curve of Output vs Single Input (others averaged)

4.8.2 Breast cancer

To categorize the output we were predicting, we used a condition on the value of diagnosis in our formulated equation. When Diagnosis lies between 1.5 up to 3.5 we declare it as B (benign). In case diagnosis exceeds 3.5 or is equal to it, it is considered as M (malignant). All other values are represented as '-', meaning that this value can not be classified into a specific group. This rule of classification was applied in creation of the prediction table herein, whereby we were able to draw comparison between conclusion classes and real results and also measure the performance of the given model.

Table 4.14 : Test dataset for breast cancer

index	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	Diagnosis
551	10.82	68.89	361.6	0.06602	0.01548	0.00816	0.1976	13.03	83.9	505.6	0.1633	0.06194	0.03264	B
552	10.86	68.51	360.5	0.04227	0	0	0.1661	11.66	74.08	412.3	0.07348	0	0	B
553	11.13	71.49	378.4	0.08194	0.04824	0.02257	0.203	12.02	77.8	436.6	0.1782	0.1564	0.06413	B
554	12.77	81.35	507.9	0.04234	0.01997	0.01499	0.1539	13.87	88.1	594.7	0.1064	0.08653	0.06498	B
555	9.333	59.01	264	0.05605	0.03996	0.01282	0.1692	9.845	62.86	295.8	0.08298	0.07993	0.02564	B
556	12.88	82.5	514.3	0.05824	0.06195	0.02343	0.1566	13.89	88.84	595.7	0.162	0.2439	0.06493	B
557	10.29	65.67	321.4	0.07658	0.05999	0.02738	0.1593	10.84	69.57	357.6	0.171	0.2	0.09127	B
558	10.16	64.73	311.7	0.07504	0.005025	0.01116	0.1791	10.65	67.88	347.3	0.12	0.01005	0.02232	B
559	9.423	59.26	271.3	0.04971	0	0	0.1742	10.49	66.5	330.6	0.07158	0	0	B
560	14.59	96.39	657.1	0.133	0.1029	0.03736	0.1454	15.48	105.9	733.5	0.3171	0.3662	0.1105	B
561	11.51	74.52	403.5	0.1021	0.1112	0.04105	0.1388	12.48	82.28	474.2	0.2517	0.363	0.09653	B
562	14.05	91.38	600.4	0.1126	0.04462	0.04304	0.1537	15.3	100.2	706.7	0.2264	0.1326	0.1048	B
563	11.2	70.67	386	0.03558	0	0	0.106	11.92	75.19	439.6	0.05494	0	0	B
564	15.22	103.4	716.9	0.2087	0.255	0.09429	0.2128	17.52	128.7	915	0.7917	1.17	0.2356	M
565	20.92	143	1347	0.2236	0.3174	0.1474	0.2149	24.29	179.1	1819	0.4186	0.6599	0.2542	M
566	21.56	142	1479	0.1159	0.2439	0.1389	0.1726	25.45	166.1	2027	0.2113	0.4107	0.2216	M
567	20.13	131.2	1261	0.1034	0.144	0.09791	0.1752	23.69	155	1731	0.1922	0.3215	0.1628	M
568	16.6	108.3	858.1	0.1023	0.09251	0.05302	0.159	18.98	126.7	1124	0.3094	0.3403	0.1418	M
569	20.6	140.1	1265	0.277	0.3514	0.152	0.2397	25.74	184.6	1821	0.8681	0.9387	0.265	M
570	7.76	47.92	181	0.04362	0	0	0.1587	9.456	59.16	268.6	0.06444	0	0	B

Table 4.15: Test the single-order model to predict breast cancer

SI NO	**Expected value**		**Predicted value using mathematical equation**		**Correct/ incorrect**	**Accuracy**	**Error**
	Value	**diagnosis**	**value**	**diagnosis**			
1	2	Benign	2	Benign	✓		
2	2	Benign	2	Benign	✓		
3	2	Benign	2	Benign	✓		
4	2	Benign	2	Benign	✓		

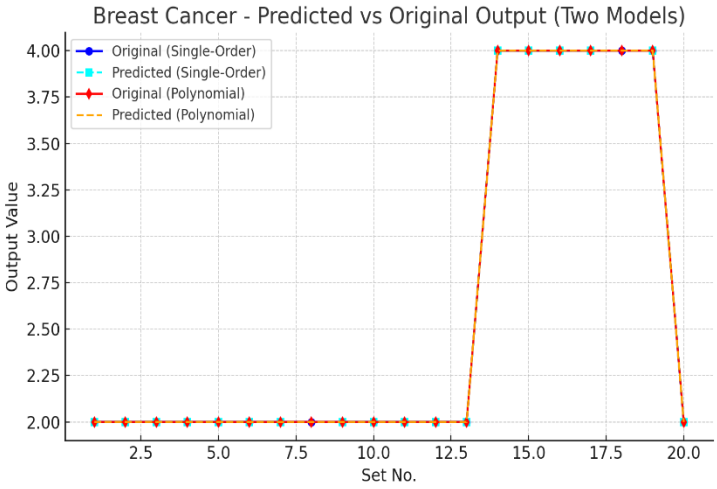
5	2	Benign	2	Benign	✓	90%	10%
6	2	Benign	2	Benign	✓		
7	2	Benign	2	Benign	✓		
8	2	Benign	4	Malignant	×		
9	2	Benign	2	Benign	✓		
10	2	Benign	2	Benign	✓		
11	2	Benign	2	Benign	✓		
12	2	Benign	2	Benign	✓		
13	2	Benign	2	Benign	✓		
14	4	Malignant	4	Malignant	✓		
15	4	Malignant	4	Malignant	✓		
16	4	Malignant	4	Malignant	✓		
17	4	Malignant	4	Malignant	✓		
18	4	Malignant	2	Benign	×		
19	4	Malignant	4	Malignant	✓		
20	2	Benign	2	Benign	✓		

Table 4.16: Test the polynomial model to predict breast cancer

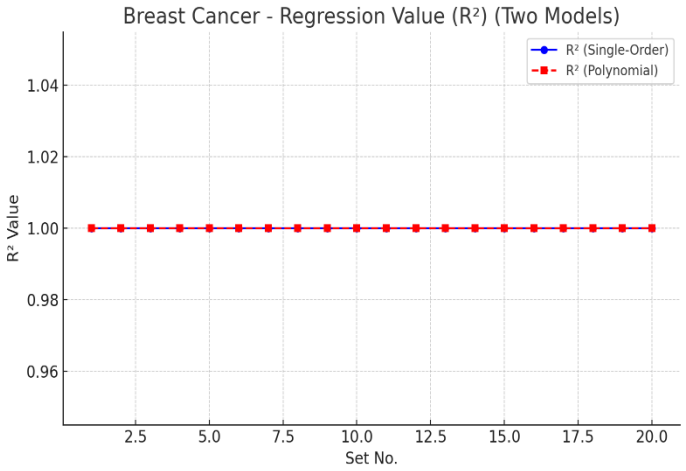
SI NO	**Expected value**		**Predicted value using mathematical equation**		**Correct/ incorrect**	**Accuracy**	**Error**
	Value	**diagnosis**	**value**	**diagnosis**			
1	2	Benign	2	Benign	✓	100%	0%
2	2	Benign	2	Benign	✓		
3	2	Benign	2	Benign	✓		
4	2	Benign	2	Benign	✓		
5	2	Benign	2	Benign	✓		
6	2	Benign	2	Benign	✓		
7	2	Benign	2	Benign	✓		
8	2	Benign	2	Benign	✓		
9	2	Benign	2	Benign	✓		
10	2	Benign	2	Benign	✓		
11	2	Benign	2	Benign	✓		
12	2	Benign	2	Benign	✓		
13	2	Benign	2	Benign	✓		
14	4	Malignant	4	Malignant	✓		
15	4	Malignant	4	Malignant	✓		
16	4	Malignant	4	Malignant	✓		
17	4	Malignant	4	Malignant	✓		
18	4	Malignant	4	Malignant	✓		
19	4	Malignant	4	Malignant	✓		
20	2	Benign	2	Benign	✓		

Among the single-order model (90% accuracy) and the polynomial model (100% accuracy), the model that showed to be more accurate was the polynomial model for Breast cancer.

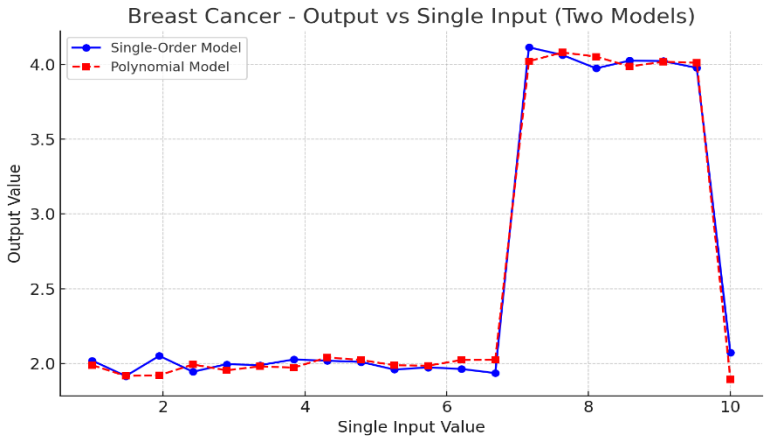
Results are shown in graph-



a)



b)



c)

Fig 4.6: Testing output of breast cancer visualize by graph - **a)** Predicted vs Original Output vs Set No **b)**Regression Value (R^2) bar chart **c)**Curve of Output vs Single Input (others averaged)

Chapter 05

Conclusion

5.1 Introduction

This chapter consists of the results of our research as a conclusion. Section 6.2 gives the general and the main conclusion of all the findings and results achieved during the creation of the given models. In 6.3, the following section explains the future presence of the project including the possible improvements, applications, and further study undertakings of the practice of early cancer detection through mathematical modeling.

5.2 Conclusion

Within the framework of this thesis, we tried to come up with good mathematical models that can identify lung and breast cancer at an early stage based on clinical data. Four models were created, two to represent lung cancer and the other lung cancer; they consisted of single order (linear) equations plus second order (polynomial) equations. Every model has been built attentively on the chosen clinical values and was measured using different statistical tools, such as ANOVA analysis, coefficient calculation, equation fitness tests, and numerous testings of different sets of data. In the case of lung cancer, we obtained a polynomial equation which proved a good fit with great precision and small error rates.

In the same way, in the case of breast cancer the polynomial equation will be performed better than its single order counterpart. In the two cases, the R^2 and adjusted R^2 values were greater, the standard errors (S) and PRESS values were lesser and accuracy was 100 percent in testing in the polynomial models, compared to the lower accuracy and higher residual errors in the single-order models. Using these overall findings, we can arrive at the conclusion that the polynomial models are more precise, dependable and effective in forecasting the outcome of both lung and breast cancer. The models could be therapeutically useful even in a situation where resources are limited as in the case of access in expensive medical imaging or other diagnostic tools. This study also makes contributions to the emerging area of early cancer screening, and given the effectiveness of the identified solution could help healthcare providers to make effective decisions in time

5.3 Future Work

This study provides a great basis to form mathematical models in predicting occurrence of lung and breast cancer in their early stages through clinical parameters. Nevertheless, to be further developed and improved in future work, there are a few vectors that need to be explored.

Compare with existing model: Existing model will compare with mathematical equation.

Integration with the Real-Time Clinical Systems: Mathematical models can connect to the clinical decision-making system or healthcare software and help the physicians diagnose the patients early at the early phase of disease in the low resource areas, where there are minimal access to imaging tools.

Hybrid Modeling and Artificial intelligence: Though the present research used the deterministic mathematical models, a hybrid solution, where deterministic-mathematical models would be used in tandem with the machine learning or artificial intelligence would allow to increase the power of predictions, flexibility, and automation of the detection process.

A User Interface or an App can be created: A small program or application can be created based on the final models in the form of the polynomial, and the healthcare worker can add patient data into it and get predictive estimates in real time.

Real World Data Model Validation: Real-time data at a hospital or a diagnostic center will also be required to further validate the developed models and ascertain the applicability and strength of the developed models in the real world.

Applicability to Other Cancers: The current methodology of the research can be applied to other cancer studies by use of other cancer and appropriate parameters. The methodology could be applicable to the prostate cancer or even cervical cancer.

Addition of Cost and Accessibility Analysis: A limitation to this research could be included by examining the cost-effectiveness and availability of application of this mathematically based detection system in the future, preferably in a rural or undeveloped healthcare environment.

References

[1] Y. X. ,. J. L. Jialin Zhou, "Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN," *pubmed,* 2024.

[2] M. M. a. ·. S. M. S. M. M. ·. R. I. M. M. ·. …. ·. B. S. M. M. Muhammad Rafiqul Islam, "Lung Cancer in Bangladesh," *IASLC,* 2023.

[3] C. Z. L. C. X. Z. Lanwei Guo guolanwei1019@126.com, "Global burden of lung cancer in 2022 and projected burden in 2050," *Chin Med J,* 2024.

[4] Y. Z. 1. †, "Global burden of female breast cancer: new estimates in 2022, temporal trend and future projections up to 2050 based on the latest release from GLOBOCAN," *Journal of the National Cancer Center,* 2025.

[5] M. A. R. Forazy, "Incidence of breast cancer in Bangladesh," 2015.

[6] T. S. M. Devis, "Lung Cancer Detection Using Machine Learning," *Proceedings of the National Conference on Emerging Computer Applications,* 2022.

[7] Y. S. a. L. G. Siddharth Bhatia, "Lung Cancer Detection: A Deep learning," 2019.

[8] L. Wang, "Deep Learning Techniques to Diagnose Lung Cancer," *cancers,* 2022.

[9] M. Karabatak, "A new classifier for breast cancer detection based on Naïve," *Measurement,* pp. 32-36, 2015.

[10] n. Semih Ergin a and O. K. b, "A new feature extraction framework based on wavelets," *Computers in Biology and Medicine,* pp. 171-182, 2014.

[11] A. A. Shubham Sharma and T. Choudhury, "Breast Cancer Detection Using Machine Learning," *IEEE,* 2018.

[12] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," *ACM Digitial Library,* 2018.

[13] M. M. ,. J. S. ,. N. H. ,. B. A. B. Kaviya Sathyakumar, "Automated Lung Cancer Detection Using artificial intellegece(AI)," *CUREUS,* 2020.

[14] I. B. T. Pandiangan1*, "Early Lung Cancer Detection Using Artificial," *Atom Indonesia,* 2018.

[15] ∗. K. P. H. K. S. G. Seral ¸Sahana, "A new hybrid method based on fuzzy-artificial immune system and k-nn," *Computers in Biology and Medicine,* pp. 415-423, 2007.

[16] †. P. A. C. H. B. Sean Blandin Knight1, "Progress and prospects of early detection," *open biology,* 2017.

[17] ⇑. P. Mohamed Shakeel a, "Lung cancer detection from CT image using improved profuse clustering," *ScienceDirect,* 2019.

[18] T. S. M. Devis, "Lung Cancer Detection Using Machine Learning," *NCECA,* 2022.

[19] M. M. ,. Kaviya Sathyakumar, "Automated Lung Cancer Detection Using ai," *cureus,* 2020.

[20] Y. S. a. L. G. Siddharth Bhatia, "Lung Cancer Detection: A Deep," 2019.

[21] †. P. A. C. Sean Blandin Knight1, "Progress and prospects of early detection," 2017.

[22] L. Wang, "Deep Learning Techniques to Diagnose Lung Cancer," *cancers,* 2022.

[23] A. Chon, "Deep Convolutional Neural Networks for Lung Cancer Detection".

[24] K. Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining," *Asian Pacific Journal of Cancer Prevention,,* 2013.

[25] I. B. T. Pandiangan1*, "Early Lung Cancer Detection Using Artificial," *atom indonesia,* 2018.

[26] M. Santarpia1, "Liquid biopsy for lung cancer early detection," 2018.

[27] N. M. N Kalaivani1*, "Deep Learning Based Lung Cancer Detection and".

[28] A. Shimazaki1, "Deep learning-based algorithm," 2022.

[29] S. S. A.-N. Ibrahim M. Nasser, "Lung Cancer Detection Using Artificial Neural Network," *IJEAIS,* 2019.

[30] A. K. S. a. B. Gupta, "A Novel Approach for Breast Cancer Detection and," *Procedia Computer Science,* pp. 676-682, 2015.

[31] D. Bazazeh, "Comparative Study of Machine Learning Algorithms," *IEEE,* 2016.

[32] M. A.-M. Leila Abdelrahman, "Convolutional neural networks for breast cancer detection in," *Computers in Biology and Medicine,* 2021.

[33] C. COLEMAN, "EARLY DETECTION AND," *ScienceDirect,* 2017.

[34] S. G. Kandlikar, "Infrared imaging technology for breast cancer detection – Current status,," *ScienceDirect,* 2017.

[35] M. M. Rivera-Franco, "Delays in Breast Cancer Detection and Treatment in," *sage journals,* 2018.

[36] Rupali Sood, "Ultrasound for Breast Cancer Detection Globally:," *asco publications,* 2019.

[37] A. F. Seddik, "Logistic Regression Model for Breast Cancer," *IEEE,* 2015.

[38] ∗. J. S. W. J. Y. G. L. Z. b. H.D. Chenga, "Automated breast cancer detection and classification using ultrasound images:," *Pattern Recognition,* pp. 299-317, 2010.

[39] M. Rupali Sood, "Ultrasound for Breast Cancer Detection Globally:," *ASCO Publications,* 2019.

[40] R. Guo, "ULTRASOUND IMAGING TECHNOLOGIES FOR BREAST CANCER DETECTION," *Ultrasound in medicine and biology,* 2018.

[41] z. sultana, "Early Breast Cancer Detection Utilizing Artificial Neural Network," *WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE,* 2021.

[42] A. B. Nassif, "Breast cancer detection using artificial intelligence techniques: A," *Artificial Intelligence in Medicine,* 2022.

[43] C. Zemmour, "Prediction of Early Breast Cancer Metastasis from," *Sage Journals,* 2015.

[44] N. Houssami, "Breast screening using 2D-mammography or integrating," *ScienceDirect,* pp. 1700-1807, 2014.

[45] A. McGuire, "Effects of Age on the Detection and Management of," *cancers,* 2015.

[46] M. C. Hye-Won Kang1., "A Mathematical Model for MicroRNA in Lung Cancer".