LEVERAGING MACHINE LEARNING AND DEEP LEARNING TO ENHANCE LEAN OPERATIONS IN HEALTHCARE: A FOCUS ON LUNG CANCER DETECTION

by

KEVIN DE LA ROSA, B.S.

THESIS
Presented to the Graduate Faculty of
The University of Texas at San Antonio
in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN ADVANCED MANUFACTURING AND ENTERPRISE ENGINEERING

COMMITTEE MEMBERS:
F. Frank Chen, Ph.D., Chair
HungDa Wan, Ph.D.
Omar Abbaas, Ph.D.

THE UNIVERSITY OF TEXAS AT SAN ANTONIO
Klesse College of Engineering and Integrated Design
Department of Mechanical Engineering
December 2023

DEDICATION

This thesis is dedicated to my family and loved ones. Thank you for providing me with constant inspiration and support throughout my academic journey.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. F. Frank Chen, for his guidance, patience, and encouragement throughout my research. His expertise and insights have been invaluable in shaping my work and fueling my passion for engineering and advanced manufacturing. I would also like to thank Dr. Mohammad Shahin and my committee members Dr. HungDa Wan and Dr. Omar Abbaas for their feedback and support. I am grateful for the opportunity to have attended The University of Texas at San Antonio and appreciate all the support I have received from the Department of Mechanical Engineering throughout my academic career.

December 2023

LEVERAGING MACHINE LEARNING AND DEEP LEARNING TO ENHANCE LEAN OPERATIONS IN HEALTHCARE: A FOCUS ON LUNG CANCER DETECTION

Kevin De La Rosa, M.S.
The University of Texas at San Antonio, 2023

Supervising Professor: F. Frank Chen, Ph.D.

As cancer ranks within the top five causes of death within the United States, the current cancer environment, applications of lean methodologies in the healthcare industry, and the implementation of artificial intelligence and machine learning to support cancer treatment, patient's experiences, and oncology operations is explored. Statistical analysis is then performed on a lung cancer patient dataset to understand the correlation the variables have to cancer diagnosis. Various artificial intelligence models such as Random Forest (RF), Convolutional Neural Networks (CNN), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron Neural Network (MLP-NN) are then applied to the dataset to evaluate model accuracy and identify if the application can improve oncology centers operational and treatment efficiency. XGBoost with and without Principal Component Analysis (PCA), Logistic Regression, Random Forest, and MLP-NN with and without PCA achieved an accuracy of 100%, with LR with PCA (98.93%), and CNN (96.27%) following. These high accuracies confirm the implementation of artificial intelligence within the healthcare organization can be successful in supporting diagnosis predictions and enhancing lean operations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Within the United States, cancer ranks within the top five causes of death following the leading cause of death, heart disease. Data collected from the Centers for Disease Control and Prevention shows over 600,000 cancer related deaths occurred in 2019 and 1.9 million new cancer related cases are projected for 2023. As there are multiple forms of cancer; breast, prostate, and lung are among the top commonly diagnosed with lung, colon, pancreatic, breast, and prostate as the top 5 leading causes of cancer deaths respectively. As there are multiple factors that result in cancer diagnosis and terminal illness, the common diagnosis is dependent on whether the patient is female or male. For females the leading diagnosis is breast cancer and for males it is prostate cancer. However, the leading cause of death for both male and female patients is lung cancer [1]. As cancer cases continue to increase, costs associated with cancer treatment continues to increase as well resulting in patients paying more for their treatments and oncology practices experiencing impact of rising costs resulting in practices striving to implement cost reduction strategies throughout their entire operations.

The complexity of treatment for the various types of cancer can result in high cost for both the patient and the health provider. Cost associated with cancer care includes multiple forms such as financial costs, costs with respect to loss of time, care provider and life adjustment costs, and emotional costs. Dependent on the form of cancer, stage, age of the patient, and diagnosis timeframe, the total amount of treatment costs varies. The Centers for Disease Control and Prevention (CDC) observed cost of treatment varied dependent on the form of cancer, the location within the patient's body, and revealed female breast cancer was the leading form of cancer with national patient expenses of 3.14 billion, along with prostate, colorectal, and lung cancer with over one billion in expenses each as well [2]. The stage of cancer and the different sections of the

lifecycle of treatment have various cost levels. It is shown that in the last year prior to death associated with cancer of a patient, treatment costs per patient is higher compared to the initial year after diagnosis and the average costs for continuous treatment throughout the treatment lifecycle. Patients with stage 0 compared to stage 4 may encounter different costs as well. Patients diagnosed with stage 0 may accumulate more costs associated with having to undergo surgeries while patients with stage 4 will spend more on chemotherapy. On average, up to 90 percent of costs are related to the planning and treatment within the first year of chemotherapy [3]. Dependent on the patient's stage of cancer at the time of diagnosis affects the total overall costs as well. If the patient is diagnosed with cancer while their cancer is at stage 0 there may be a spike in costs initially due to surgery with long-term treatment costs reducing significantly compared to patients diagnosed while their cancer is as at a higher stage. Patients are encountering high costs associated with cancer related pharmaceuticals, with higher cost associated with brand named drugs versus the generic brand. Location of the care provided may also affect the treatment costs for patients. As costs continue to rise, patients heavily rely on insurance coverage, care packages, and oncologists to provide the recommended treatment plan that balances cost and recovery effectiveness.

As health providers, cancer care centers, and oncology practices aim to provide high quality treatment for their patients, a financial challenge is encountered from performing daily operations, technology investments, employing specialist, and conducting research to effectively treat cancer. In addition to providing care for patients, doctors are becoming more involved with business operations and overall strategy to survive the increasing costs of care and to implement thoughtful practices to meet the needs of their patients and operational requirements. As cancer care costs is trending to surpass 300 billion within the next 5 years, physicians frequently find themselves

balancing the consideration of assets, liabilities, capital expenses, and operational expenses throughout their practice [4]. Healthcare organizations are also utilizing their frontline employees to continuously evaluate their operations to identify waste and recommend changes to improve the efficiency of their processes.

One of the largest expenses care providers must manage is pharmaceutical acquisition and associated operational costs such as proper storage. In certain oncology practices, cost associated with pharmaceutical purchasing and maintenance is higher than 50 percent of the overall operational costs requiring extensive inventory control, frequent inventory audits, purchase automation, supporting software and management systems, and potentially the required support from third party vendors [5]. Oncology practices must maintain awareness of changes in complex prior authorizations processes for medications to avoid delaying treatment to the patient and accruing unexpected costs.

While conducting financial management practices, the salaries of hiring specialist must be considered along with the training to operate new equipment, perform specific processes, and learn any changes to operations. Within an oncology practice, various types of equipment are used through the lifecycle of cancer treatment and includes devices to screen, help diagnose, treat, and monitor patients. One of the most common medical equipment devices found in an oncology practice is a linear accelerator. For radiation oncology, an average investment of 4 million is associated with linear accelerator purchases and the related equipment to operate the device [6]. To develop their cancer treatment, oncology practices invest in conducting cancer research and completing clinical trials. Immunotherapy is one of the most common types of cancer research being adopted.

The increase in cost trending across the industry has driven healthcare providers to invest in lean six sigma methodology and to embed new technologies such as artificial intelligence within their operations. As practices implement pilot programs in their environment, innovation of treatment, improvement of the operations and reduction of facility expenses, the addition of automation, and the utilization of latest forms of technology occur. Optimizing the environment improves the patient's overall experience, reduces various forms of waste, and has an indirect impact of improving sustainability.

CHAPTER TWO: LEAN HEALTHCARE

The incorporation of lean and six sigma methodologies can be applied across various industries to enable organizations to reduce waste, optimize, and decrease errors within their operations. The healthcare industry does not differ as healthcare providers use these methodologies as they strive to continuously improve their patient's experience, the treatment and care that is provided, and overall internal operations. Lean methods applied in healthcare across various case studies has resulted in a decrease in wait time, total time of stay, and patients leaving without seeing a doctor overall improving patient and employee satisfaction [7].

By understanding the long-term benefits of improving many small processes throughout their operations, the Baylor Scott and White McClinton Cancer Center were successfully able to improve their cancer care by incorporating lean strategies and utilizing the knowledge of their frontline workers. Utilizing lean tools such as value stream mapping, 5S, and kaizen, the medical center recognized improvements covering educational programs, department productivity, inventory, safety, communication, and patient care [8]. Lean methodology relies on continuous application and developing a workforce culture that naturally seeks improvements in addition to managerial awareness of lean strategies and their impact, all while incorporating monitoring of progression.

Research conducted at the King Hussein Cancer Center evaluated the addition of lean thinking to improve patient care. Value stream mapping and value-added/ nonvalue-added analysis was done to evaluate the patient's experience in an outpatient environment. It was discovered that even with a 14% increase in patients, the cancer center had trouble maintaining their patient's satisfaction as wait times increased, discomfort in packed waiting rooms existed, and resulting high cost that didn't relate to the treatment being received. By using value stream mapping and

adjusting their operations to increase value-added activities, the cancer center identified waste to eliminate and reduced their patients mean waiting time by up to 73.1%, enabled more time slots for treatment, and improved patient satisfaction score by approximately 34% [9].

Adding lean thinking into an organization requires all employees to believe in the process and understand that the negative impact of keeping processes stagnant is greater than the challenge of improving processes to meet patient demand and the quality the organization is striving to achieve. A quality-improvement initiative at the Smilow Cancer Hospital focused on developing an employee culture around lean thinking. To improve their inpatient and outpatient workflows, lean strategies such as Just-In-Time and the elimination of nonvalue-added activities were used to strive to reduce their patients wait time, increase the efficiency of their internal processes, and reduce the delivery time of medication. By implementing a new unified computerized prescriber order-entry system, making staffing adjustments to accommodate the new system, removing nonvalue-added activities in their workflows, and updating their standardization of their documentation, the medication delivery time was decreased by 47% [10]. This improvement is expected to grow along with additional improvements to occur as the employees continue using lean thinking to recommend changes.

Oral Cancer Surgery performed within the day hospital in the Maxillofacial Surgery Department encountered increased waiting times for patients resulting in unsatisfactory in patient's experience and stressful working conditions for the medical team. The use of lean methodology was used to improve to current situation, Value stream mapping and fishbone diagrams enabled the medical staff to improve their processes and adjust to patient demands. The fishbone diagrams supported the medical team in identifying improvement areas spanning across their internal systems, processes, staffing, and patients. This evaluation of cause and effects covered waiting

times for consultancy, clinical examinations and test results, evaluating complex bureaucratic procedures, business management and the utilization of information systems, and patient co-morbidity. Combining analysis with value stream mapping improvements, the department improved patient care and reduced the overall patient's length of stay at the hospital by 22.40% [11].

The Radiation Oncology Department at The University of Michigan Health System evaluated their environment utilizing current state mapping to identify process times, total lead time, and first-time quality. Becoming aware of their opportunities of improvement, a future state map was developed, and one-piece flow was implemented to reduce process steps and improve cross department collaboration. The standardization for medical record review process and terminology used by the clinical corrected resulted in accurate scheduling and treatment time estimations. The incorporation of lean methodology reduced overall variability and increased same day treatment from 43% to 95% [12].

At Aurora Health Care, the management initiated a learning path focusing on lean training of 800 employees covering what lean methodology is, how lean can improve processes, and how to complete a lean initiative within their organization. Following their training, the employees at Aurora Health Care completed a Gemba walk throughout their laboratory, pharmacy, and chemotherapy departments to understand their internal processes and identify waste. The completion of the Gemba walk resulted in corrections to improve communications, specimen analysis, and patient arrival times. Communication was also sent to the patients informing of the corrections made to their processes and the resulting outcome in patient wait time decreasing by 22% [13].

Striving to eliminate waste and to identify cost related to various internal processes, time-driven activity-based costing and lean methodology was utilized at a radiation oncology for breast cancer. This was accomplished by analyzing the patients care trajectory planning stage, identifying resource groups, and evaluating activities associated and aligning cost with the activities completed. The cost-breakdown of the various activities revealed waste was associated with duplicate data collection, remodifying treatment planning, physicians' resistance to change and adopting to new technology, duplicate manual and digital documentation, inconsistent digitization across departments, and multiple follow-ups requesting additional test data from prior treatments. Standardization and optimization of the internal processes relating to care planning and documentation sharing estimated a potential time reduction of 30 minutes per patient and approximately $20,000 in savings. Understanding the accumulation of costs and the associated time of the activities enabled the cancer hospital to plan for redefining their resource allocations, improving documentation consistency, communication across departments, and improvement of lean management decisions to reduce waste [14].

Following six-sigma methodologies to improve quality, DMAIC was utilized to identify delays in discharge within a neuro-oncology hospital. The first three phases of DMAIC, (define, measure, analyze) resulted in the development of qualitative interviews, process mapping, and root cause analysis. It was discovered, delays in discharge of patients resulted from communication gaps across multidisciplinary teams involving medical and rehabilitation teams. The communication errors included inconsistent discharge information provided to patients and their families resulting in lack of trust and quality of patient care. The main bottleneck of the discharge process was the patient's destination after discharge as the rehabilitation team and medical team had to come to consensus and identify if the patient qualified for rehabilitation. Striving for

continuous quality improvements in patient care, the DMAIC methodology enabled the internal teams to improve clarity amongst the teams by identifying team responsibility and involvement of discharge planning, the increase of meeting frequency, and the importance of specialist representation within meetings [15].

CHAPTER THREE: COMPUTATIONAL PATHOLOGY

To develop state of the art computational pathology systems that can successfully perform within a real-world environment, innovations in machine learning to train from data at the petabyte scale, developed datasets resembling clinically relevant real-world data, the collaboration of pathologist and computer scientist, and high-performance computing for efficient deep learning at scale is needed. Computational pathology within the healthcare industry has the potential to increase the efficiency of clinical workflow and the development of custom treatment plans for patients. The implementation of algorithms that can accurately operate using big data consumed by various processes within the medical environment can aid in identifying the nature of an illness, decrease errors in the selection of clinical codes, and provide effective insight on what patients are to expect throughout their treatment [16].

Artificial intelligence in breast histopathology image analysis can help decrease cost, increase resource utilization efficiency, and reduce errors in evaluating specimens made by humans due to the subtle discrepancy in diagnostic criteria tolerances, different philosophies on morphological criteria and the use of additional clinical information, and external factor variations such as emotional state, fatigue, and stress. With the use of public data, standardization of color normalization artificial intelligence has been used to support morphological assessments by inspecting degree of tubule/ gland formation, mitotic count, and nuclear pleomorphism [17]. With this degree of identification, the implementation of computational pathology may support further analysis, estimate survival predictions, and develop tailored treatment strategies.

The development of large datasets and computer algorithms have made it possible to discover further insights in common histopathology of cancer. The use of computational pathology along with the digitization of tissue on glass slides has transitioned manual tasks such as

segmentation and mitotic counting to become automated and has improved resource allocation of pathologists. Key use cases of computational pathology in cancer research and oncology include diagnostic tasks such as detecting tumor tissue on digitize whole slide images, the prediction of genetic alterations which aids in understanding the correlation of genomic alterations and the morphology in tissue slides, prognostication to determine chemotherapy strategies, and the prediction of treatment response to accurately select patients suitable for immunotherapy [18].

To mitigate the reliance on expensive and time-consuming manual annotations at the pixel-level on large datasets, frameworks from training classification models at large scale have been developed. This form of weakly supervised deep learning on whole slide images was implemented utilizing annotations at the slide-level from diagnosis determined from anatomic laboratory information systems and electronic health records. Multiple instance learning and slide-level annotations to determine if all tiles of a specimen are negative or if at least one tile is positive and includes a tumor portion of tissue has been used to develop and train deep learning systems. With the implementation of recurrent neural networks, ResNet34 models have been able to accurately classify clinical samples. This computational pathology method has enabled pathologist to exclude 65-75 percent of slides while retaining 100 percent sensitivity and proved the ability to train accurate classification models at scale with 10,000 slides needed for reliable performance [19].

Computational pathology has also been effective in supporting pathologist with identifying formations of breast cancer. Pathologists have been able to use artificial intelligence to evaluate whole slide images with average of 100,000 by 100,000 pixels and classify imaging alerts linked to disease processes. The use of multiple staining methods and different artificial intelligence approaches such as soft label convolutional networks, modified fully convolutional networks, and fully interactive automatic hierarchical registration models have enabled pathologist to overcome

bottlenecks occurring in their environment, increase the speed of sample analysis, and develop the capability to identify the grade and stage of a tumor within the dataset [20]. As datasets become more defined and the addition of medical records from the patient are connected, further insights can be identified with these models providing the patients with a clearer understanding and improves the process pathologist take throughout the treatment lifecycle.

Computational pathology-based descriptors have also been developed precisely for identifying the prognosis of lung cancer. A feature-driven local cell graph known as FLOCK, has been implemented in the analysis of nuclei spatial arrangement to identify different cancer types and determine prognosis values for patients. FLOCK can determine prognosis values by constructing local cell graphs and simultaneously evaluate nuclear properties that considers proximity which allows for interrogation of interactions between different groups of cell types and reveals more sub-visual information from the underlying tissue image. The development of quantitative histomorphometry features to capture interactions of local cell clusters with similar morphological properties enables medical staff to classify survival values among patients of early-stage non-small cell lung cancer with mean area under the curve values of 0.68 which was higher than other various methods tested [21].

The implementation of deep learning has also shown success in supporting pathologist in visually inspecting tissue properties from stained images as well as predicting intricate characteristics that are not commonly found by manual inspections conducted by pathologists. The use of deep learning classification of stained breast tumor tissue microarray images has been tested to identify properties such as tumor grade, histologic subtype, estrogen receptors, intrinsic breast cancer subtypes, and risk of reoccurrence scores. The deep learning method resulted in clinical biomarker predictions of 75 percent to 94 percent accuracy with incorrect classification common

in Luminal B tumors. The high accuracy values prove deep learning methods can be applied to support pathologist with additional information that can help determine if patients need further genomic testing or these methods can replace existing manual processes of visual inspections [22].

As computational pathology has shown a positive impact in the healthcare environment, future advancements rely on an increase in accurate data annotations and explainable algorithms. Pathologist have found using deep learning as an assistive tool to be beneficial to their work by decreasing analysis time, reducing human error, and its support in identifying factors that have a high impact on a patient's survival and the opportunity to evaluate which chemotherapy applications have the highest effect in treating cancer. Incorporating computational pathology and deep learning applications is important to determine both diagnoses early on and the high impact risk factors to increase successful treatment plans. The continuation of computational pathology within the healthcare environment relies on various strategies to analyze data, the use of annotations to reduce volume of data needed to train algorithms, and the addition of interpretable artificial intelligence to gain the full trust of pathologist [23].

CHAPTER FOUR: STATISTICAL DATA ANALYSIS

The dataset used in this research was obtained from 1000 cancer patients and includes various insights into their lifestyles [24]. This data included 23 lifestyle data variables from each patient associated with various variable types. Table 1 provides a list of the lifestyle items and the variable type. The data falls within 5 different variable types including Continuous Quantitative, Discrete Quantitative, Binary Categorical, Nominal Categorical, and Ordinal Categorical. For continuous quantitative variables the data includes gender, dust allergy, chronic lung disease, passive smoker, coughing of blood, wheezing, clubbing of fingernails, dry cough, and snoring. Nominal categorical variables included occupational hazards. Ordinal categorical variables included, generic risk, balanced diet, chest pain, fatigue, shortness of breath, and swallowing difficulty. Discrete quantitative variables included alcohol use, smoking, and frequent cold. Lastly, continuous quantitative variables included air pollution, obesity, and weight loss. The data collected from these variables were categorize in levels of severity ranging from 0 to 10, aside from Age, Gender, and outcome.

Table 1: Lifestyle items collected from patients with lung cancer

No.	Lifestyle Item	Variable Type
1	Age	Continuous Quantitative
2	Gender	Binary Categorical
3	Air Pollution	Continuous Quantitative
4	Alcohol use	Discrete Quantitative
5	Dust Allergy	Binary Categorical
6	Occupational Hazards	Nominal Categorical
7	Genetic Risk	Ordinal Categorical
8	Chronic Lung Disease	Binary Categorical
9	Balanced Diet	Ordinal Categorical
10	Obesity	Continuous Quantitative
11	Smoking	Discrete Quantitative
12	Passive Smoker	Binary Categorical
13	Chest Pain	Ordinal Categorical
14	Coughing of Blood	Binary Categorical
15	Fatigue	Ordinal Categorical
16	Weight Loss	Continuous Quantitative
17	Shortness of Breath	Ordinal Categorical
18	Wheezing	Binary Categorical
19	Swallowing Difficulty	Ordinal Categorical
20	Clubbing of Fingernails	Binary Categorical
21	Frequent Cold	Discrete Quantitative
22	Dry Cough	Binary Categorical
23	Snoring	Binary Categorical

Tables 2, 3, 4, and 5 display the descriptive statistics included within the dataset. It is shown that within the population 598 were male, 402 were female, shown in Figure 1, and the median age was 36, with a minimum age of 14 and maximum age of 73. Shown in Figure 2, the age bin with the largest population fell between ages 32-38, with a size of 297 patients. Table 3 and Figure 3 displays the descriptive statistics of the health-related variables included in the dataset. Out of the 1000 patients, dust allergy had the largest amount of level 7s reported for severity of dust allergy with 405 patients. With Genetic Risk, Chest Pain, Coughing of Blood also resulting in a

mode of 7. For Swallowing Difficulty severity, had 221 patients report a level 1. Table 4 and Figure 4, showcase the descriptive statistics on environmental variables such as Occupational Hazards and Air Pollution. Occupational Hazard's mode was level 7, with 365 patients, and Air Pollution's mode was level 6, with 326 patients. Lastly, Table 5 and Figure 5, showcase the descriptive statistics for habitual variables. It is shown that Obesity, had a mode of 7 with 356 patients reporting a level 7 for whether the patient is obese. Alcohol Use has a mode of 2, with 202 patients reporting a level 2 for alcohol use, however 188 patients also recorded a level 8 showing patients either didn't consume a lot of alcohol or the did consume a lot, with low percentages reported for levels 3-6. Similar trend followed for the remaining variables, as the histograms showcased peaks on level 2 and level 7.

Table 2: Descriptive Statistics on Gender and Age

Gender & Age	*Gender*	*Age*
Mean	1.402	37.174
Standard Error	0.015512467	0.379647015
Median	1	36
Mode	1	35
Standard Deviation	0.490547283	12.00549274
Sample Variance	0.240636637	144.1318559
Kurtosis	-1.843407143	0.059540224
Skewness	0.400354446	0.551095929
Range	1	59
Minimum	1	14
Maximum	2	73
Sum	1402	37174
Count	1000	1000

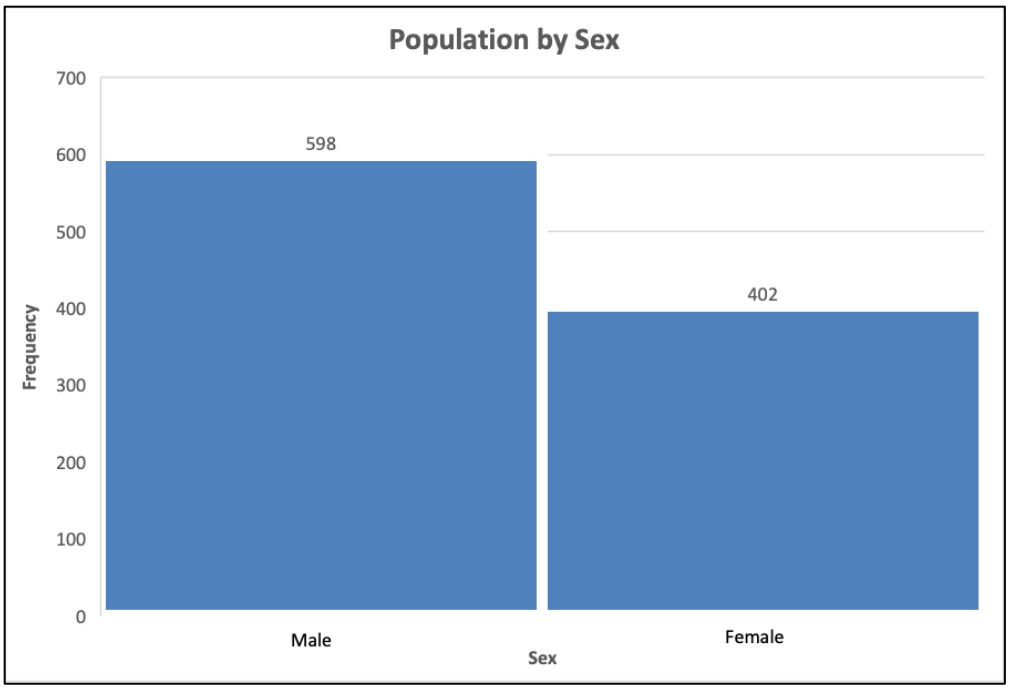

Figure 1: Population by Sex

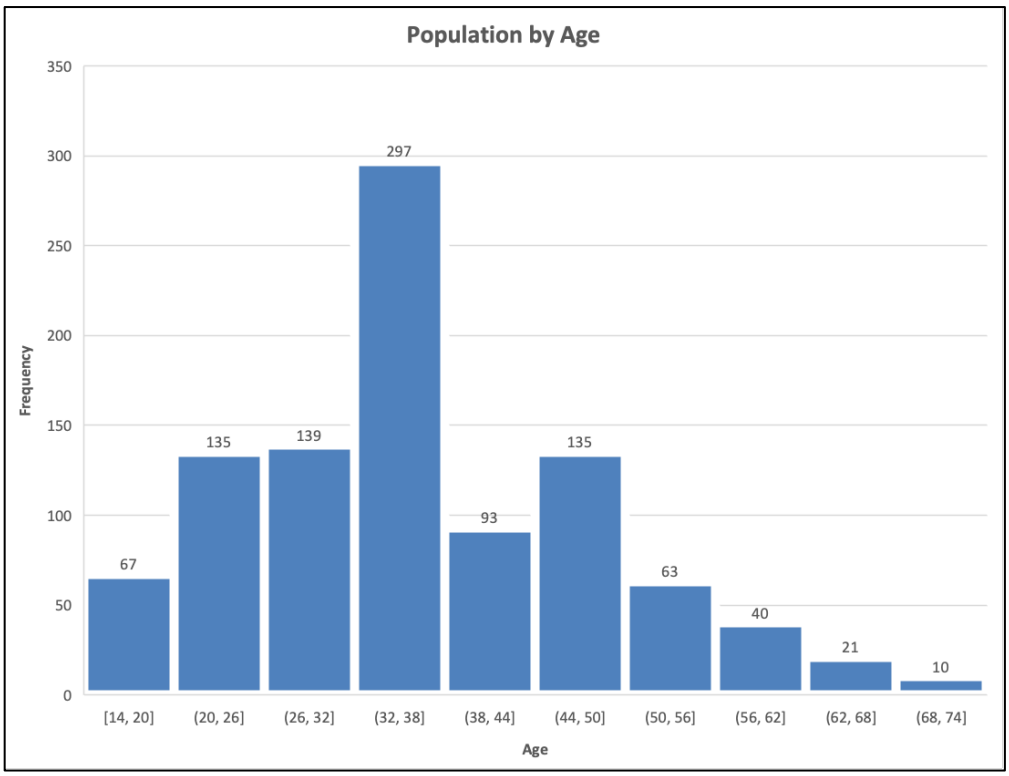

Figure 2: Population by Age

Table 3: Descriptive Statistics on Health variables

Health	*Genetic Risk*	*Chronic Lung Disease*	*Chest Pain*	*Coughing of Blood*	*Fatigue*	*Shortness of Breath*	*Wheezing*	*Swallowing Difficulty*	*Frequent Cold*	*Dry Cough*	*Snoring*	*Clubbing of Fingernails*	*Dust Allergy*
Mean	4.58	4.38	4.438	4.859	3.856	4.24	3.777	3.746	3.536	3.853	2.926	3.923	5.165
Standard Error	0.06726161	0.058455257	0.072106556	0.076778995	0.070981	0.072260789	0.064571204	0.071795812	0.057948788	0.064479055	0.046633665	0.075516712	0.062639434
Median	5	4	4	4	3	4	4	4	3	4	3	4	6
Mode	7	6	7	7	3	2	2	1	3	2	2	2	7
Standard Deviation	2.126998854	1.848517519	2.280209498	2.427964994	2.244616293	2.285086786	2.041920772	2.270382928	1.832501586	2.039006755	1.474685966	2.38804811	1.98083283
Sample Variance	4.524124124	3.417017017	5.199355355	5.895014014	5.038302302	5.221621622	4.16944044	5.154638639	3.358062062	4.157548549	2.174698699	5.702773774	3.923698699
Kurtosis	-1.596775844	-1.304259226	-1.359402592	-1.293398118	-0.208398746	-0.854903552	-1.184367557	-0.887465886	-0.942710894	-1.29081057	-0.551099266	-0.336646267	-0.861244256
Skewness	-0.126664724	-0.220465357	0.164707172	0.121997427	0.855629883	0.406383045	0.224154547	0.451176799	0.406448546	0.223835125	0.550045306	0.796563772	-0.644709196
Range	6	6	8	8	8	8	7	7	6	6	6	8	7
Minimum	1	1	1	1	1	1	1	1	1	1	1	1	1
Maximum	7	7	9	9	9	9	8	8	7	7	7	9	8
Sum	4580	4380	4438	4859	3856	4240	3777	3746	3536	3853	2926	3923	5165
Count	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

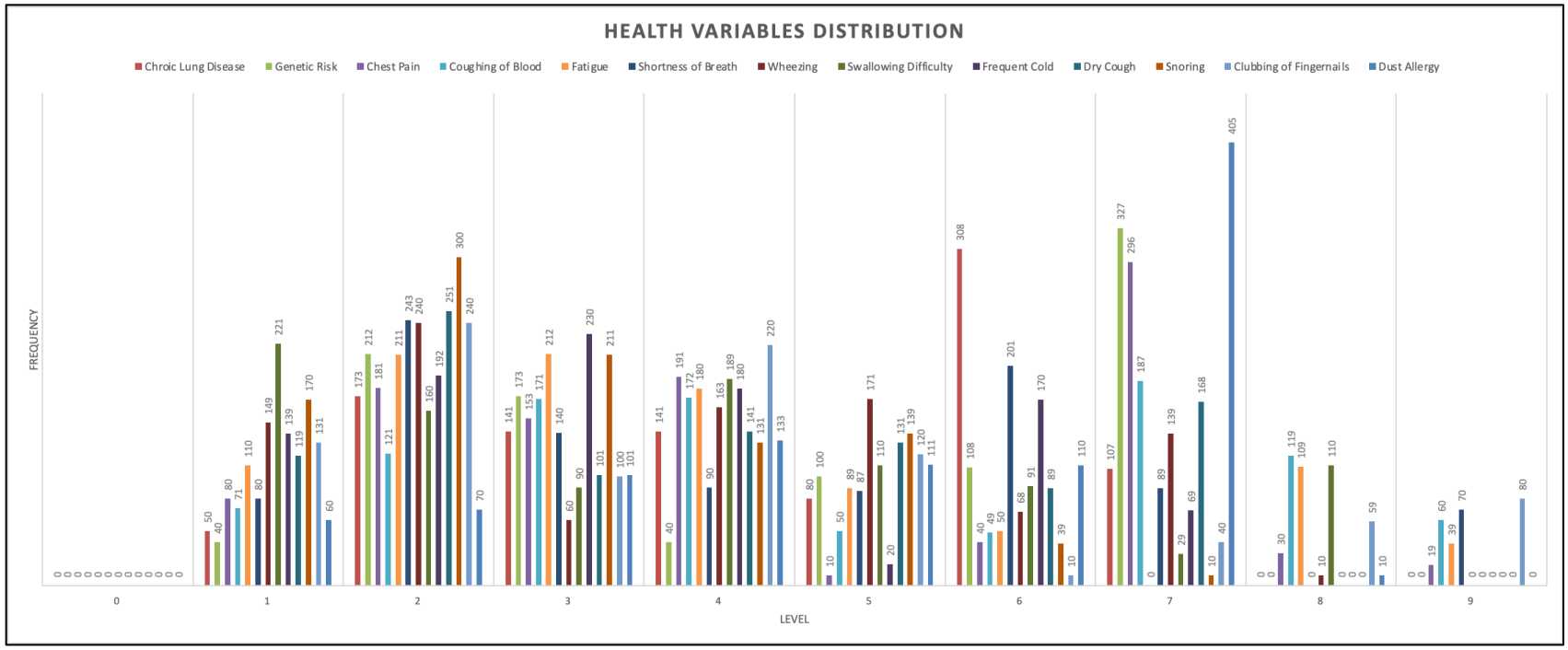

Figure 3: Health Variables Distribution

Table 4: Descriptive Statistics on Environment variables

Environment	*Air Pollution*	*Occupational Hazards*
Mean	3.84	4.84
Standard Error	0.064206873	0.066654654
Median	3	5
Mode	6	7
Standard Deviation	2.030399597	2.107805219
Sample Variance	4.122522523	4.442842843
Kurtosis	-1.386848847	-1.36440553
Skewness	0.125451607	-0.234504909
Range	7	7
Minimum	1	1
Maximum	8	8
Sum	3840	4840
Count	1000	1000

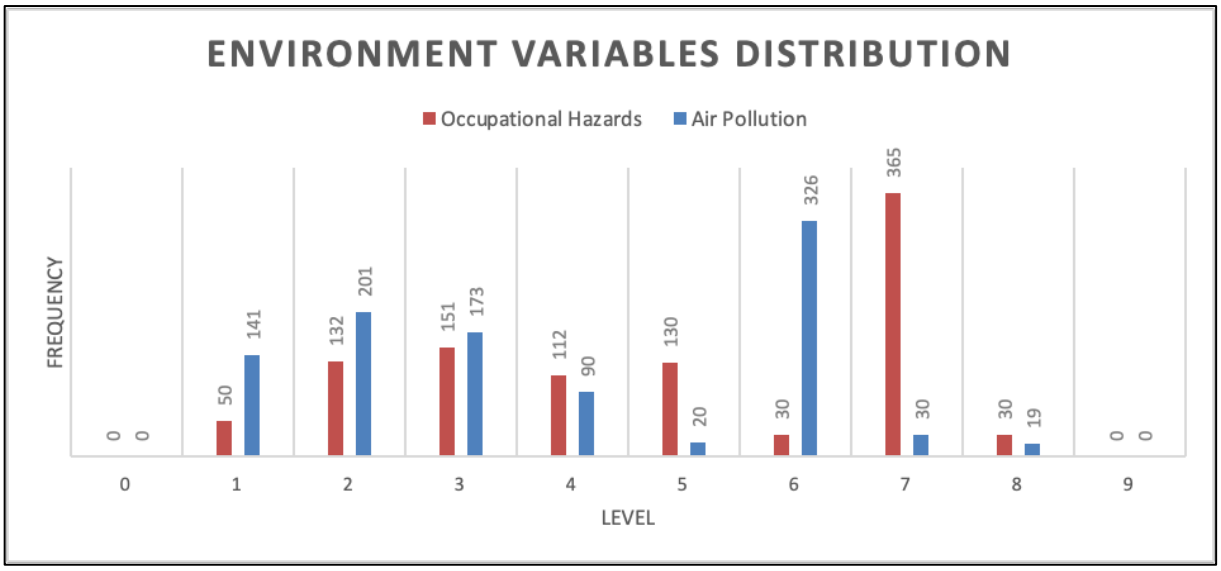

Figure 4: Environment Variables Distribution

Table 5: Descriptive Statistics on Habitual variables

Habits	*Balanced Diet*	*Alcohol use*	*Obesity*	*Smoking*	*Passive Smoker*	*Weight Loss*
Mean	4.491	4.563	4.465	3.948	4.195	3.855
Standard Error	0.067531322	0.082866748	0.06719591	0.078927343	0.073104852	0.069777101
Median	4	5	4	3	4	3
Mode	7	2	7	2	2	2
Standard Deviation	2.135527916	2.620476655	2.124921243	2.495901746	2.311778389	2.206545681
Sample Variance	4.560479479	6.866897898	4.51529029	6.229525526	5.344319319	4.868843844
Kurtosis	-1.641138365	-1.596021187	-1.476493147	-1.451150422	-1.328060847	-1.390635406
Skewness	-0.064495385	-0.016389972	0.028844654	0.381312131	0.411458822	0.355133591
Range	6	7	6	7	7	7
Minimum	1	1	1	1	1	1
Maximum	7	8	7	8	8	8
Sum	4491	4563	4465	3948	4195	3855
Count	1000	1000	1000	1000	1000	1000

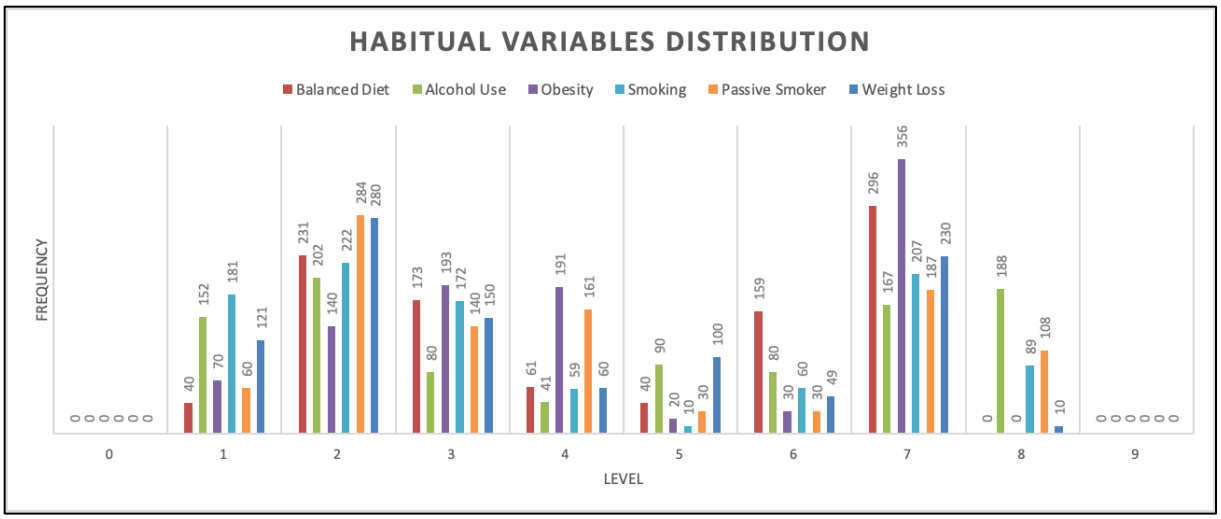

Figure 5: Habitual Variables Distribution

The impact of these variables is important and should be considered when developing artificial intelligence and machine learning models that predict whether a patient may have cancer or is susceptible to developing a cancerous disease. Developing an accurate model to precisely determine diagnosis can have an impact on the patient's treatment plan and overall success rate for curing the disease.

Age is an important factor to consider in cancer prediction, and even though patients can be diagnosed with cancer at any age, there are common trends and increases in cancer diagnosis relating to age groups. It is estimated from the American Cancer Society that the probability percentage of individuals developing an invasive cancer, based on the cancer statistics within the United States from 2017 to 2019, the older the individual the probability increases. The estimation includes the probability for the following age groups for people free of cancer at the beginning of the age interval: birth to age 49 (3.5 for male, 5.8 female), age 50-59 (6.2 for male, 6.4 for female), age 60-69 (13.8 for male, 10.4 for female), age 70-74 (34 for male, 27.2 for female) [25]. Based on the Word Cancer Day 2020: International Public Opinion Survey on Cancer, within the United States approximately 80% of cancers are diagnosed in individuals 55 years of age or greater [26].

Regardless of the type of cancer, risk increases, for example the incidence of colorectal cancer rate approximately doubles every subsequent 5-years until the age of 50 [27].

Gender is also a significant factor and should be included in analysis for predicting diagnosis, developing treatment plans, and cancer prevention. Although a complex variable, female genetics may interact differently to cancer development and treatment compared to male genetics. Studies have shown that dependent on the gender, the patient may have a higher risk in malignancy and a worse prognosis with various forms of cancer. Different genders also carry different genetics that can positively support the immune response reducing the rate of mortality from cancer. However, the implementation of gender as a variable for cancer diagnosis, treatment effectiveness, and survival rate is newly studied compared to environmental or habitual variables [28]. Hormones have shown to impact and affect cancer stem cells, tumor environments, immune system and metabolism with androgens more susceptible to cancer risk and mortality and estrogen as more preventative [29]. Evaluating the impact of gender is complex as many factors contribute such as societal norms that may expose patients to different occupational hazards, living conditions, and lifestyle habits.

Genetics and various patient health symptoms can have an impact of cancer as well. Genetic risk evaluations can be performed to obtain a probability of a genetic mutation, identify if any specific cancer related mutations are present, and estimate the chance of cancer developing throughout a patient's life. Predictive genetic testing can help patients identify if specific mutations have been inherited from a family line of cancer, if mutations present may develop other forms of cancer if the patient already has cancer, or raise awareness to other family members to encourage genetic testing [30]. Some genetic mutations do not have a large impact resulting in low genetic performance, small differentiating physical characteristics of the individual, or not performing the

intended genetic function. Genetic mutations with high impact such as pathogenic variants can affect functions of the entire body and lead to the development of cancer [31]. On average, 13% of women have a chance of developing breast cancer within their lifetime. With an inherited mutated or pathogenic variant of the BRCA1 or BRCA2 genes, the percentage can increase to 55-72%, for BRCA1, or 45-69%, for BRCA2, of women will develop breast cancer [32]. Identifying present pathogenic variants and genetic risk level can help predict if a highly correlated cancer will develop or if there is a possibility that other forms of cancer can occur.

Individuals with dust allergies can be either prone to or more resilient to developing cancer dependent on the type of the allergen. Although research is not definitive whether dust allergies have a direct correlation to cancer, an immune response may react differently whether allergens are present or not. Studies have shown high exposure to dust particles is capable of affecting the viability of cells [33]. The presence of allergens may also cause the immune response to hyper react and increase defenses against malignant cells resulting in an inconclusive understanding on the correlation [34].

As for Chronic Lung Disease, an environment prone to cancer development can occur. The various forms of lung disease can cause inflammation, tissue damage and scarring, mucus production, and a weakened immune system. Inflammation can lead to unwanted chemicals within the lung biome, scarring can disrupt the lung architecture, mucus production can trap harmful carcinogens, and a weaken immune system may not have the capability to defend against mutations [35]. These factors can lead to the growth of cancers cells resulting in lung cancer and even other forms of cancer [36].

Chest pain can be a symptom for various reasons and factors. In a study completed it was shown that chest pain was an initial symptom to the development of lung cancer [37]. In some

scenarios, chest pain can be an early sign of cancer and may be felt from pressure on blood vessels, nerves, or lymph nodes which can be impacted from tumor growth or cancer spreading throughout the chest area, ribs, or spine [38].

Coughing of blood is a common symptom that is reported as one of the early signs of lung cancer and can lead to a lower diagnosis time if discovered and reported [39]. This can occur when a tumor irritates or causes damage to bronchial lining, if blood vessels in the throat erode, or from oral cavity irritation all stemming from cancer occurrence. Bleeding within the lung airways can be a symptom of lung cancer which causes the coughing of blood to occur, in some cases the bleeding may be minimal (30 mL) and unrealized and other scenarios can be more severe (100-600 mL) [40]. Once reported, various forms of testing such as imaging, endoscopic, blood, or biopsy can be performed to verify if cancer is present [41].

Clubbing of fingernails can occur due to the decrease in oxygen levels within the blood and is most common from lung cancer and is a possible symptom of other forms of cancer such as liver, gastrointestinal, and Hodgkin lymphoma [42].Fatigue is another common symptom of cancer development and cancer treatment although it can be inter-related, vague, and difficult to identify what other symptoms are associated such as fatigue from the result of depression or anxiety. Differentiating from tiredness, fatigue typically is longer lasting, and interferes with everyday tasks. Anemia is a common symptom of cancer which reduces the red blood count and oxygen levels resulting in the feeling of fatigue [43]. Metabolic changes, sleep disturbances, immune response, and hormonal changes due to cancer formation can result in the feeling of fatigue.

Shortness of breath and wheezing may also be a result of cancer development within the lungs or from cancer affecting the chest area. Bronchial passages can become blocked due to the growth of cancerous tumors in the trachea and bronchi which restricts airway making it more

difficult to breath causing the shortness of breath experience [44]. The accumulation of fluid in the pleural space can result in breathlessness and the size of a pleural effusion has an impact on the survival duration and mortality probability [45]. A pulmonary embolism, or blood clots within the artery of the lungs, is another symptom of cancer development and can cause a sudden shortness of breath that increases with physical exertion [46]. The occurrence of a pulmonary embolism is high in patients with cancer and is six times more likely to occur in lung cancer patients compare to patients that are cancer free [47].

Swallowing Difficulty can be experienced from the result of tumor growth making is difficult for food to digest from the mouth to the stomach. In certain scenarios the esophageal lumen may narrowing increasing the resistance or even preventing food from progressing. The development of cancer may also result in scarring and fibrosis within the throat tissue, and muscle dysfunction resulting in swallowing difficulty, pain, or discomfort [48].

Some forms of cancer can alter the functions of the immune system blood cells causing cells to disrupt the immune systems operations to protect the body. Cancer can also destroy protective barriers, block natural drainage of mucus, and damage tissues increasing the opportunity for germs and infections [49]. This impact to the immune system can increase the frequency of a cold occurring. As a dry cough can occur as a symptom of various scenarios such as asthma, infections, and bronchitis, it is important to note the characteristics of a dry cough that may be related to cancer has a longer duration more than 8 weeks, interferes with sleep, and may accompany with coughing of blood, chest pain, and shortness of breath [50]. During the development of cancer, pleura, or lining of the lungs, can increase in thickness which then applies pressure to the lungs resulting in a dry cough to occur. Up to 50% of lung cancer patients may experience a persistent dry cough that can worsen in later stages of lung cancer [51]. In a study

completed, 18% of patients reported severe distress, 15% claimed sleep disturbance, and half of the population recognized treatment was needed for their cough [52]. Snoring is a symptom that is indirectly correlated to cancer diagnosis and snoring can evolve into obstructive sleep apnea, can result in chronic inflammation, reduce oxygen levels, effect hormonal changes, and reduce the immune system due to disturbance in sleep quality all which can lead to the development of cancer or the environment in which cancer cells can thrive. Snoring increased the risk of various forms of cancer including, head and neck, colorectal, and breast [53].

Environmental factors that patients are continuously exposed to can affect overall health and impact the development of cancer cells. A study has shown, air pollution that is inhaled can affect the alarm response system within the lungs resulting in inflammation and the activation of dormant cells that can carry cancer-causing mutations [54]. As air pollution increases, the risk of lung cancer and other forms of cancer can rise as well. Another study showed air pollution is estimated to cause 29% of lung cancer deaths globally [55]. Overall worldwide there is a positive correlation between air pollution with lung cancer incidence and mortality resulting in the large percentages of lung cancer deaths annually [56]. Occupational Hazards can have an impact on an individual's chance of developing cancer. Dependent on the work environment, an exposure to carcinogenic substances, radiation, respiratory hazards, and chemicals. An example that is commercially used is Asbestos, if inhaled this can cause respiratory illness that may lead to lung cancer [57]. Exposure to emissions and chemical substances as agriculture and public health workers are can compose of cancer-trigger factors such as pesticides which correlates to lung cancer [58].

Habits and lifestyle choices have an impact on the development of cancer, how the immune system responds, and the reaction to treatment. A balanced diet is an important factor and a

preventative to developing cancer. With a healthy and organic nutrition plan, a balanced diet can help reduce inflammation, improve the digestive system to remove toxins from the body, encourage new cell growth and prevent damage to existing cells and DNA. Avoiding processed food can reduce the exposure to harmful substances and including antioxidants can help prevent chronic diseases including cancer [59]. Obesity has shown to be highly correlate to cancer as hormonal changes can occur, insulin resistance decreases causing high levels of insulin in the blood, and a gut microbiome that enables the growth of cancer. A study showed obesity and excess body weight contribute to 9.6% of cancers in women and 4.7% in men within the United States [60].Weight loss can help improve overall health and quality of life, however if the weight loss is unexpected, unintentional, or rapid there is a probability for an underlying disease, infection, or the possibility of cancer development. An unexpected weight loss can result in an increased risk of pancreatic, gastro-oesophageal, lymphoma, hepatobiliary, lung, bowel and renal tract cancers [61]. The development of a tumor may apply pressure to areas of the digestive track, stomach, or throat areas resulting in the feeling of fullness reducing the amount of food intake, can prevent the proper absorption of nutrition, or cause uneasiness when trying to eat in which all can lead to weight loss.

Various studies showcase alcohol use as heavily correlated with cancer and is labeled as a known human carcinogen by the National Toxicology Program of the U.S. Department of Health and Human Services as clear patterns link alcohol consumption to head and neck, esophageal, liver, breast, and colorectal cancers through toxic chemicals, oxidation, resisting nutrient absorption, and impacting hormones [62]. As alcohol is consumed, it can be converted into acetaldehyde which can damage DNA and the alcohol's byproducts can damage the liver, causing inflammation and scarring that may lead to mutations in DNA and the development of cancer [63].

Smoking and exposure to tobacco smoke increases cancer risk and is a leading cause of cancer due to the various chemicals included that can harm and damage DNA which can lead to various forms of cancer including lung, larynx, mouth, esophagus, throat, bladder, kidney, liver, stomach, pancreas, colon, and more [64]. Of the 250 harmful chemicals within tobacco smoke, 69 chemicals are known to cause cancer and 36% of the 480,000 premature deaths are from cancer due to smoking and exposure to tobacco smoke [65]. In the U.S., 18% of all cancers can be prevented by managing body fat percentage, physical activity, nutrition, and alcohol consumption [66]. The World Health Organization states cancer risk can be reduced by implementing prevention practices such as avoiding the use of tobacco and alcohol, maintaining a healthy physical fitness companied by a healthy diet, minimizing exposure to occupational hazards regarding radiation and environmental pollution [67].

CHAPTER FIVE: METHODOLOGY

Machine learning is a subcategory of artificial intelligence that is vastly used throughout many industries and provides descriptive, predictive, or prescriptive functions to enable complex decisions to be made from data analysis. Machine learning algorithms fall into different categories including supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised machine learning models are developed with labeled data to train the model in predicting accurate results from unlabeled data. Unsupervised machine learning models operate by identify patterns within training data that is unlabeled to gather data points that are similar. Semi-supervised models are trained on a percentage of labeled data mixed with unlabeled data. Reinforcement learning models include the implementation of feedback based on the decisions the model is making to improve the accuracy of future outputs. Machine learning algorithms can help with complex scenarios such as including medical diagnosis from evaluating images, autonomous driving, natural language chatbots, anomaly detection, and object detection. With this supplementation of machine learning, caution must be considered due to bias and unintended results. However, with the development of explainable artificial intelligence, ethical artificial intelligence, and human-centered artificial intelligence machine learning is being implemented rapidly throughout many different applications and software used widely around the world [68].

The Random Forest algorithm is a method of classification and regression, supervised, machine learning model. This algorithm is constructed with multiple decision trees that develops a final output using the majority output of the various decision trees evaluated. For classification, the model provides the class that has appeared most throughout the multiple decision tree outputs. For regression applications, the mean or average solution across the multiple decision trees is the final output of the model. Random Forest models improve decision trees by reducing the risk of

overfitting, reducing the training time needed to develop an accurate model, and can maintain a highly accurate prediction even if a portion of the data is missing. The application of Random Forest models can scale to meet the demand of large datasets which makes it viable for real-world applications. A few applications of Random Forest methods include multiclass object detection, breast cancer prediction, remote sensing, and product recommendation [69].

Convolutional neural networks are a popular application of deep learning that is used for pattern recognition that reduces the required parameters in artificial neural networks allowing for larger datasets to be evaluated. With convolution neural networks accurate image classification can be achieved regardless of image positioning, and abstract image features can be identified. Convolution neural networks are constructed of multiple layers including convolutional, activation function, non-linearity, pooling, and fully connected layers. With the implementation of filtering, CNNs can recognize various forms of patterns including edges, circles, corners, and complex patterns like faces, textures, and animal legs. The convolutional layer is the core foundation and can be modified to manage to process of data analysis by adjusting strides to control overlap, padding like zero-padding can be used to reduce loss of information, reduce complexity, and manage output. The non-linear layer can be implemented to modify the output, a pooling layer can provide down-sampling to reduce the complexity. Applications of CNNs include image classification, computer vision, natural language processing, voice recognition, and data analysis. Applications within the medical field include image classification, segmentation, and localization. Various CNN architectures include multiple variations of the layers inside the CNN with Lenet and Alexnet as popular architectures that consist of 5 convolutional layers and 1 fully connected layer and 5 convolutional layers and 2 fully connected layers respectively [70].

Logistic regression is a supervised learning algorithm used to determine binary outputs or classification tasks like if an event occurred or not, or if a diagnosis is positive or negative. Logistic regression models perform analysis on probability that the input of an independent variable with various attributes falls into a specified category of an outcome variable. With logistic regression, if the independent variable size is large, a large data set with a recommended sample size of 400 and a minimum of 10 samples per independent variable is needed for sufficient accuracy. There are three types of logistic regression including binomial which consists of two types of outputs, multinomial consisting of 3 or more types of unordered outputs, and ordinal consisting of 3 or more types of ordered outputs. The logistic regression model consists of predictor coefficients to measure the impact of variation of the dependent variable, a logistic curve to represent the correlation between the independent and dependent variables, a logistic transformation to constrain an output range between 0 and 1, and a goodness-of-fit to evaluate the overall model [71]. Applications of logistic regression may include financial application decisions, medical variable correlations, natural language processing tasks, application user predictions, and other classification implementations.

Extreme gradient boosting (XGBoost) is an algorithm built on the foundation of decision tree algorithms and gradient boosting and is available as an open source. However, XGBoost offers a balance between bias and variance to reduce the chances of overfitting, offers regularization terms that can improve generalization for the algorithm, and has system optimizations resulting in an increase of scalability compared to gradient boosting which focus primarily on variance which may lead to overfitting. The model uses various parameters dependent on the design intention and includes general parameters, booster parameters, and task parameters. General parameters relate to the boosting method selected for the model, either linear or tree models. Booster parameters are

dependent on the booster chosen. Task parameters define the task at hand and the learning objective [72]. An ensemble methodology is also applied to correct errors within the model, by initially developing the model, calculating the error for each observation, then building new models to predict the errors, then adding the prediction from the model to the ensemble of models [73]. XGBoost has many applications and can be utilized for regression, classification, and ranking problems like ranking systems for user preference, click-through rate predictions, sentiment analysis, image classification.

A multi-layer perceptron neural network is an artificial intelligence model that consists of input, hidden, and output layers. The layers are formed with perceptrons that receive various inputs and weights subject to iterative adjustments into an activation function that is created with sigmoid, hyperbolic tangent, and rectifier activation functions to solve complex nonlinear predictions. The MLP-NN uses feed forward and backward propagation, as well as multiple layers and neurons that can be increased to achieve a higher accuracy and more predictive power [74]. However, increasing the size of the hidden layer may also increase the chance of the model overfitting. To avoid overfitting, tracking the error using validation data, iteration limiting, and limiting the complexity of the network can be done. MLP-NN have high predictive performance, can be utilized for classification and regression applications with inputs that have labels, and are appropriate for supervised learning and unsupervised clustering. The activation function of the output layer is dependent on the problem being solved, and whether a single output is needed for a regression solution, binary output used for classifications, or multiple outputs for multiclass classifications [75]. Some weakness of MLP-NN may include a lack of explanation capabilities, encountering issues with missing values, overfitting, and requiring heavy computation

requirements. Although these weaknesses occur, MLP-NN are still widely used across many industries for problem solving such as financial predictions and medical diagnosing.

CHAPTER SIX: RESULTS AND DISCUSSION

The following five models, XGBoost, Logistics Regression, Random Forest, Neural Network, and CNN, along with their counterpart including PCA, were applied to the lung cancer patient dataset and a confusion matrix was constructed to evaluate the performance of the various models. The results included the model's accuracy, recall, sensitivity, precision, specificity, F1 score, and G-mean1 values. To create a confusion matrix and calculate the model's performances, true positives, true negatives, false positives, and false negatives were recorded and used for the performance calculations. The accuracy is determined by testing the model's prediction with the provided dataset and taking the number of corrected predictions divided by the total number of the dataset or subtracting the error rate of predictions from 1, the highest accuracy that can be achieved is 1. The recall or sensitivity of the model can be calculated by taking the amount of correct positive predictions and dividing that by the amount of overall total number of positives, with 1 being the highest value. The precision is calculated by taking the amount of correct positive predictions and dividing that by the total amount of positive predictions, with the best precision equaling 1. Specificity is calculated by taking the amount of correct negative predictions and dividing that by the total amount of negatives. The F1 Score is calculated using the precision and recall and is the harmonic mean of precision and recall and can only be a high value if both precision and recall are high values, this metric can give a better measurement then using accuracy alone. The G-mean1 value is determined from the geometric mean of sensitivity and precision.

The results obtained for the confusion matrix are shown in Table 6, this includes the various algorithms tested, with and without PCA, the low, medium, and high classifications, and the amount of true positives, false positives, true negatives, and false negatives that occurred. It is observed XGboost without and with PCA, Random Forest, Logistic Regression, Neural Network

without and with PCA encountered zero false positives and zero false negatives. Random Forest with PCA, CNN without and with PCA, and Logistic Regression with PCA, encountered both false positives and false negatives in a least one of the classifications.

Table 6: Confusion Matrix

Confusion Matrix		TP	FP	TN	FN
XGBoost	Low	83	0	167	0
	Medium	83	0	167	0
	High	84	0	166	0
XGBoost with PCA	Low	83	0	167	0
	Medium	83	0	167	0
	High	84	0	166	0
Random Forrest	Low	73	0	177	0
	Medium	84	0	166	0
	High	93	0	157	0
Random Forrest with PCA	Low	73	0	177	0
	Medium	84	0	0	1
	High	92	1	157	0
CNN	Low	74	2	164	10
	Medium	69	12	169	0
	High	93	0	153	4
CNN with PCA	Low	70	6	170	7
	Medium	77	7	163	6
	High	90	3	157	3
Logistic Regression	Low	74	0	176	0
	Medium	83	0	167	0
	High	93	0	157	0
Logistic Regression with PCA	Low	71	3	175	1
	Medium	82	1	164	3
	High	93	0	157	0
Neural Network	Low	84	0	166	0
	Medium	71	0	179	0
	High	95	0	155	0
Neural Network with PCA	Low	84	0	166	0
	Medium	71	0	179	0
	High	95	0	155	0

Using the results of the confusion matrix, Table 7 showcases the accuracy, recall, and precision for the various algorithms. It was observed XGBoost without and with PCA, Logistic Regression, Random Forest, and Neural Network without and with PCA achieved accuracy, recall, and precision scores of 1. Regarding accuracy, shown in Figure 6, Random Forest with PCA scored

an accuracy of 0.996, with Logistic with PCA(0.984), CNN(0.96), and CNN with PCA(0.948) trailing. Regarding recall, shown in Figure 7, Random Forest with PCA scored an accuracy of 0.996, with Logistic with PCA(0.984), CNN with PCA(0.946), and CNN(0.942). Regarding precision, shown in Figure 8, Random Forest with PCA scored an accuracy of 0.996, with Logistic with PCA(0.984), with CNN(0.986), and CNN with PCA(0.976).

Table 7: Model Accuracy, Recall, and Precision

Prediction Model	Accuracy	Recall	Precision
XGBoost	1	1	1
XGBoost with PCA	1	1	1
Logistic Regression	1	1	1
Logistic with PCA	0.984	0.984	0.984
Random Forrest	1	1	1
Random Forrest with PCA	0.996	0.996	0.996
Neural Network	1	1	1
Neural Network with PCA	1	1	1
CNN	0.96	0.942	0.986
CNN with PCA	0.948	0.946	0.976

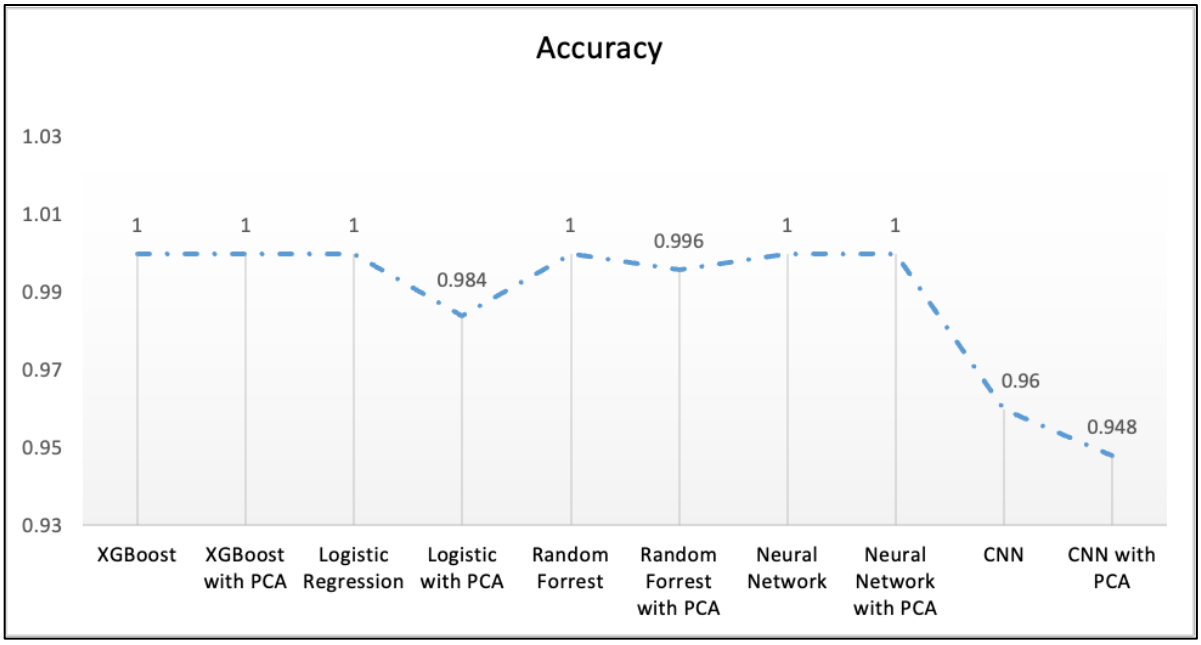

Figure 6: Model Accuracies

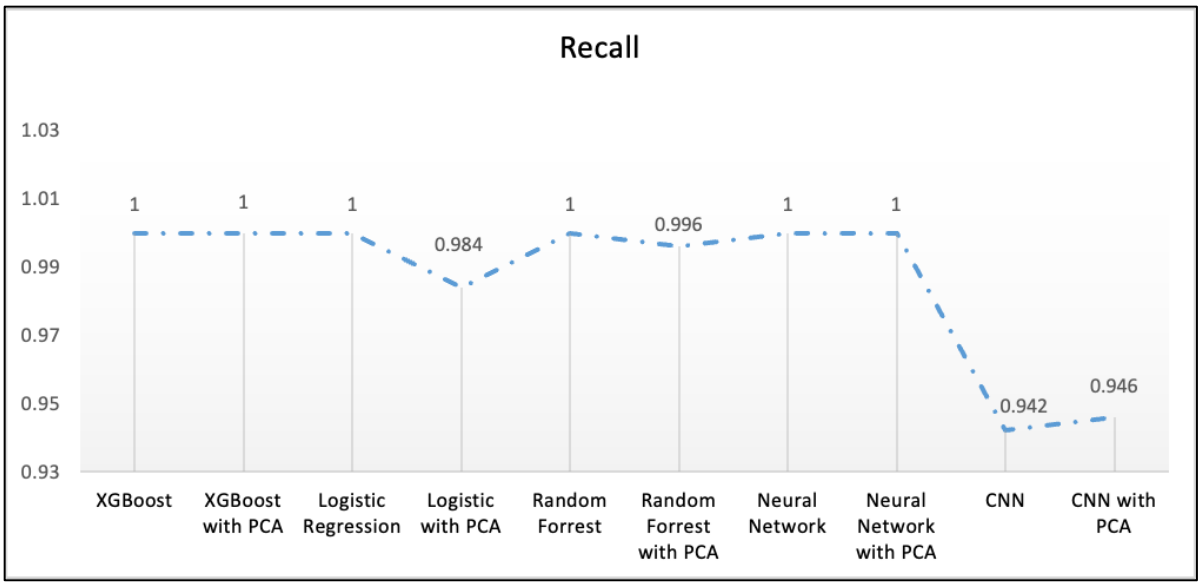

Figure 7: Model Recall

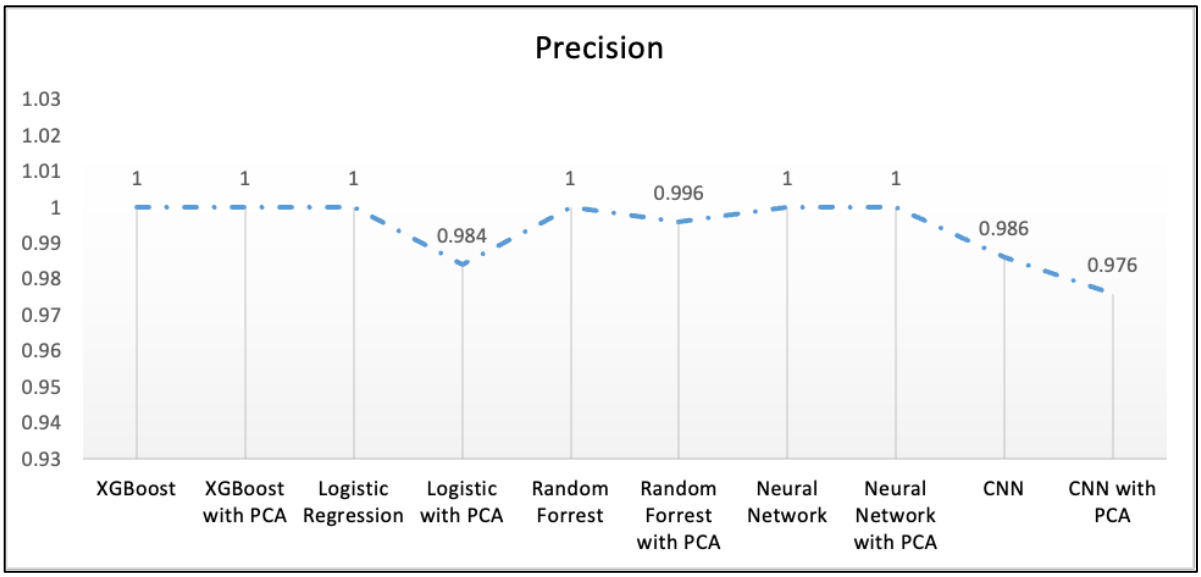

Figure 8: Model Precision

Table 8 showcases the F1 Score, Specificity, and G-mean1 results for the various models with RF with PCA, XGBoost with PCA, and MLP scoring 100% for F1 score, specificity, and G-mean1. For LR with PCA, the F1 score resulted in 98.29%, specificity(99.24%), and G-mean1(98.79%. Lastly, for CNN, the F1 score resulted in 94.13%, specificity(97.39%), and G-mean1(95.95%).

Table 8: Prediction Model Metrics

Prediction Model	Accuracy	Precision	Sensitivity	F1 score	Specificity	G-mean1
RF	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
XGBoost with PCA	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
CNN	96.27%	94.18%	94.66%	94.13%	97.39%	95.95%
LR with PCA	98.93%	98.25%	98.36%	98.29%	99.24%	98.79%
MLP	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Within the healthcare environment, accuracy of the model is significant as healthcare workers may rely on these models to forecast cancer development and identify cancer diagnosis to proactively test and treat patients. With accurate information, operational efficiency increases, and response plans can be activated earlier which can result in improving patient experience and reduce cost savings. Precision is important as well, as this states if the model capable of providing the correct classification. Having confidence in the model stating whether a patient has a low, medium, or high cancer prediction can enable healthcare teams to react accordingly. Recall is also significant to help healthcare teams avoid missing cancer diagnosis which can lead to a cancer further developing due to a delay in treatment. It is important to note that models implemented should have a high accuracy, precision, and recall ensuring the ideal results are provided so healthcare teams can act and support accordingly.

CONCLUSION

This paper explores the current cancer environment and its impact to the general population, oncology practices, and the cancer patients by reviewing the volume of cancer cases, complexities of treatment, and costs for both the treatment provider and patients. Lean healthcare is then reviewed to evaluate the lean six-sigma methodologies applied throughout the industry enabling organizations to reduce waste, optimize operations, reduce costs, and improve patient satisfaction. Computational pathologies are then explored to identify how these technologies are supporting healthcare organizations in accurate cancer diagnosis, image classification, and sample analysis thus enhancing their lean operations. Lung cancer patient data is used to evaluate variables leading to cancer to identify their correlation. To evaluate the accuracies of cancer prediction using artificial intelligence, Random Forest (RF), Convolutional Neural Networks (CNN), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron Neural Network (MLP-NN) models are tested using the lung cancer patient data. The testing of these algorithms resulted in high accuracies for prediction cancer diagnosis with RF with PCA, XGBoost with PCA, and MLP-NN achieving an accuracy of 100%, and LR with PCA (98.93%) and CNN (96.27%) following. These high accuracies confirm the implementation of artificial intelligence within the healthcare organization can be successful in supporting diagnosis predictions, classifications, and automation to enhance lean operations.

With regards to future research suggestions, the implementation of artificial intelligence throughout the entire healthcare organization and processes can be explored. The incorporation of this technology has the ability to increase efficiencies that can reduce operational waste and improve treatment experiences for patients. These models can also be evaluated against datasets that have output variables with larger ranges, such as, cancer levels ranging from 1-5 compared to

the three categories low, medium, and high provided in the lung cancer dataset. Various algorithm implementations in other industries can also be explored and evaluated to see if they can also be embedded in a healthcare environment to increase lean processes and procedures.

REFERENCES

[1] L. Eldrige, "Do You Know the Most Common Cancer In the U.S.?," Verywell Health. Accessed: Aug. 14, 2023. [Online]. Available: https://www.verywellhealth.com/what-is-the-most-common-cancer-in-the-us-2249408

[2] "The Cost of Cancer | Blogs | CDC." Accessed: Aug. 14, 2023. [Online]. Available: https://blogs.cdc.gov/cancer/2021/10/26/the-cost-of-cancer/

[3] P. J. Huckfeldt *et al.*, "Specialty Payment Model Opportunities and Assessment," *Rand Health Q.*, vol. 5, no. 1, p. 11, Jul. 2015.

[4] "Global Oncology Trends 2022 - IQVIA." Accessed: Aug. 14, 2023. [Online]. Available: https://www.iqvia.com/insights/the-iqvia-institute/reports/global-oncology-trends-2022

[5] A. E. Glode and M. B. May, "Rising Cost of Cancer Pharmaceuticals: Cost Issues and Interventions to Control Costs," *Pharmacother. J. Hum. Pharmacol. Drug Ther.*, vol. 37, no. 1, pp. 85–93, 2017, doi: 10.1002/phar.1867.

[6] B. Bourbeau, D. Harter, and E. Towle, "Results From the ASCO 2019 Survey of Oncology Practice Operations," *JCO Oncol. Pract.*, vol. 16, no. 5, pp. 253–262, May 2020, doi: 10.1200/OP.20.00009.

[7] D. Tlapa *et al.*, "Effects of Lean Healthcare on Patient Flow: A Systematic Review," *Value Health*, vol. 23, no. 2, pp. 260–273, Feb. 2020, doi: 10.1016/j.jval.2019.11.002.

[8] J. L. Watson, "Little Is Big: How 'Lean' Methodologies Can Continuously Improve Cancer Care,"*Oncol.Times*,vol.37,no.7,p.2,Apr.2015,doi:10.1097/01.COT.0000464264.62966.dd.

[9] A. Al Hroub *et al.*, "Improving the Workflow Efficiency of an Outpatient Pain Clinic at a Specialized Oncology Center by Implementing Lean Principles," *Asia-Pac. J. Oncol. Nurs.*, vol. 6, no. 4, pp. 381–388, Oct. 2019, doi: 10.4103/apjon.apjon_21_19.

[10] P. Sullivan, S. Soefje, D. Reinhart, C. McGeary, and E. D. Cabie, "Using lean methodology to improve productivity in a hospital oncology pharmacy," *Am. J. Health. Syst. Pharm.*, vol. 71, no. 17, pp. 1491–1499, Sep. 2014, doi: 10.2146/ajhp130436.

[11] A. Fiorillo, A. Sorrentino, A. Scala, V. Abbate, and O. G. Dell'aversana, "Improving performance of the hospitalization process by applying the principles of Lean Thinking," *TQM J.*, vol. 33, no. 7, pp. 253–271, Jan. 2021, doi: 10.1108/TQM-09-2020-0207.

[12] C. S. Kim, J. A. Hayman, J. E. Billi, K. Lash, and T. S. Lawrence, "The Application of Lean Thinking to the Care of Patients With Bone and Brain Metastasis With Radiation Therapy," *J. Oncol. Pract.*, vol. 3, no. 4, pp. 189–193, Jul. 2007, doi: 10.1200/JOP.0742002.

[13] D. Belter *et al.*, "Evaluation of Outpatient Oncology Services Using Lean Methodology," *Oncol. Nurs. Forum*, vol. 39, no. 2, pp. 136–40, Mar. 2012.

[14] V. Nabelsi and V. Plouffe, "Breast cancer treatment pathway improvement using time-driven activity-based costing," *Int. J. Health Plann. Manage.*, vol. 34, no. 4, pp. e1736–e1746, 2019, doi: 10.1002/hpm.2887.

[15] K. Rezk and C.-A. Miller, "Delays in Discharge in Neuro-Oncology: Using a Lean Six Sigma-Inspired Approach to Identify Internal Causes," *Can. Oncol. Nurs. J.*, vol. 26, no. 3, pp. 215–220, Jul. 2016, doi: 10.5737/23688076263215220.

[16] M. Cui and D. Y. Zhang, "Artificial intelligence and computational pathology," *Lab. Invest.*, vol. 101, no. 4, Art. no. 4, Apr. 2021, doi: 10.1038/s41374-020-00514-0.

[17] A. Duggento, A. Conti, A. Mauriello, M. Guerrisi, and N. Toschi, "Deep computational pathology in breast cancer," *Semin. Cancer Biol.*, vol. 72, pp. 226–237, Jul. 2021, doi: 10.1016/j.semcancer.2020.08.006.

[18] D. Cifci, G. P. Veldhuizen, S. Foersch, and J. N. Kather, "AI in Computational Pathology of Cancer: Improving Diagnostic Workflows and Clinical Outcomes?," *Annu. Rev. Cancer Biol.*, vol. 7, no. 1, pp. 57–71, 2023, doi: 10.1146/annurev-cancerbio-061521-092038.

[19] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nat. Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019, doi: 10.1038/s41591-019-0508-1.

[20] C.-W. Wang and H. Muzakky, "Computational Pathology for Breast Cancer and GynecologicCancer,"*Cancers*,vol.15,no.3,p.942,Feb.2023,doi:10.3390/cancers15030942.

[21] C. Lu *et al.*, "Feature-driven local cell graph (FLocK): New computational pathology-based descriptors for prognosis of lung cancer and HPV status of oropharyngeal cancers," *Med. Image Anal.*, vol. 68, p. 101903, Feb. 2021, doi: 10.1016/j.media.2020.101903.

[22] H. D. Couture *et al.*, "Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype," *Npj Breast Cancer*, vol. 4, no. 1, Art. no. 1, Sep. 2018, doi: 10.1038/s41523-018-0079-1.

[23] M. S. Iqbal, W. Ahmad, R. Alizadehsani, S. Hussain, and R. Rehman, "Breast Cancer Dataset, Classification and Detection Using Deep Learning," *Healthcare*, vol. 10, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/healthcare10122395.

[24] "Lung cancer data - dataset by cancerdatahp," data.world. Accessed: Nov. 27, 2023. [Online]. Available: https://data.world/cancerdatahp/lung-cancer-data

[25] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA. Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[26] “Cancer and ageing | UICC.” Accessed: Sep. 16, 2023. [Online]. Available: https://www.uicc.org/what-we-do/areas-focus/cancer-and-ageing

[27] W. Street, “Colorectal Cancer Facts & Figures 2020-2022”.

[28] A. Irelli, M. M. Sirufo, C. D’Ugo, L. Ginaldi, and M. De Martinis, “Sex and Gender Influences on Cancer Immunotherapy Response,” *Biomedicines*, vol. 8, no. 7, Art. no. 7, Jul. 2020, doi: 10.3390/biomedicines8070232.

[29] C. M. Lopes-Ramos, J. Quackenbush, and D. L. DeMeo, “Genome-Wide Sex and Gender Differences in Cancer,” *Front. Oncol.*, vol. 10, 2020, Accessed: Sep. 16, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2020.597788

[30] “What is Genetic Testing? Understanding Genetic Testing for Cancer.” Accessed: Oct. 28, 2023.[Online].Available: https://www.cancer.org/cancer/risk-prevention/genetics/genetic-testing-for-cancer-risk/understanding-genetic-testing-for-cancer.html

[31] “Genetic Mutations | Types of Mutations.” Accessed: Oct. 28, 2023. [Online]. Available: https://www.cancer.org/cancer/understanding-cancer/genes-and-cancer/gene-changes.html

[32] “BRCA Gene Mutations: Cancer Risk and Genetic Testing Fact Sheet - NCI.” Accessed: Oct.28,2023.[Online].Available:https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

[33] K. Ardon-Dryer, C. Mock, J. Reyes, and G. Lahav, “The effect of dust storm particles on single human lung cancer cells,” *Environ. Res.*, vol. 181, p. 108891, Feb. 2020, doi: 10.1016/j.envres.2019.108891.

[34] B. G. M. C. Carneiro *et al.*, “Clinical and immunological allergy assessment in cancer patients,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Sep. 2021, doi: 10.1038/s41598-021-97200-y.

[35] D. N. O’Dwyer, R. P. Dickson, and B. B. Moore, “The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease,” *J. Immunol.*, vol. 196, no. 12, pp. 4839–4847, Jun. 2016, doi: 10.4049/jimmunol.1600279.

[36] Y. Ai *et al.*, “Add fuel to the fire: Inflammation and immune response in lung cancer combined with COVID-19,” *Front. Immunol.*, vol. 14, 2023, Accessed: Oct. 28, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1174184

[37] P. M. Ellis and R. Vandermeer, “Delays in the diagnosis of lung cancer,” *J. Thorac. Dis.*, vol. 3, no. 3, Sep. 2011, doi: 10.3978/j.issn.2072-1439.2011.01.01.

[38] "Early Signs of Lung Cancer That May Save Your Life," Verywell Health. Accessed: Oct. 28, 2023. [Online]. Available: https://www.verywellhealth.com/early-signs-of-lung-cancer-5191947

[39] F. M. Walter *et al.*, "Symptoms and other factors associated with time to diagnosis and stage of lung cancer: a prospective cohort study," *Br. J. Cancer*, vol. 112, no. 1, Art. no. 1, Mar. 2015, doi: 10.1038/bjc.2015.30.

[40] P. Hu, G. Wang, H. Cao, H. Ma, P. Sui, and J. Du, "Haemoptysis as a prognostic factor in lung adenocarcinoma after curative resection," *Br. J. Cancer*, vol. 109, no. 6, pp. 1609–1617, Sep. 2013, doi: 10.1038/bjc.2013.485.

[41] "How to Detect Lung Cancer | Lung Cancer Tests." Accessed: Oct. 29, 2023. [Online]. Available:https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/how-diagnosed.html

[42] "Clubbing of the fingers or toes Information | Mount Sinai - New York," Mount Sinai HealthSystem.Accessed:Oct.31,2023.[Online].Available:https://www.mountsinai.org/health-library/symptoms/clubbing-of-the-fingers-or-toes

[43] "Anemia and Cancer," MD Anderson Cancer Center. Accessed: Oct. 29, 2023. [Online]. Available: https://www.mdanderson.org/patients-family/diagnosis-treatment/emotional-physical-effects/anemia-cancer.html

[44] "Treatment of Tracheal & Bronchial Tumors | Memorial Sloan Kettering Cancer Center." Accessed:Oct.29,2023.[Online].Available:https://www.mskcc.org/cancer-care/types/tracheal-diseases/diagnosis-treatment-msk/treatment-tracheal-bronchial-tumors

[45] S. Allen, "Relationship of Pleural Effusions with Outcomes in Cancer Patients," vol. 4, 2023.

[46] "Pulmonary Embolism - Symptoms and Causes | Penn Medicine." Accessed: Oct. 29, 2023.[Online].Available:https://www.pennmedicine.org/for-patients-and-visitors/patient-information/conditions-treated-a-to-z/pulmonary-embolus

[47] Y. Li, Y. Shang, W. Wang, S. Ning, and H. Chen, "Lung Cancer and Pulmonary Embolism: What Is the Relationship? A Review," *J. Cancer*, vol. 9, no. 17, pp. 3046–3057, Aug. 2018, doi: 10.7150/jca.26008.

[48] K. H. Ph.D, "Dysphagia in cancer patients: What to know," MD Anderson Cancer Center. Accessed:Oct.29,2023.[Online].Available:https://www.mdanderson.org/cancerwise/dysphagia-in-cancer-patients--what-to-know-causes-diagnosis-prevention-treatment.h00-159305412.html

[49] "Why People with Cancer Are More Likely to Get Infections." Accessed: Oct. 29, 2023. [Online]. Available: https://www.cancer.org/cancer/managing-cancer/side-effects/low-blood-counts/infections/why-people-with-cancer-are-at-risk.html

[50] "What Determines if a Cough is Related to Lung Cancer?," Healthline. Accessed: Oct. 30, 2023. [Online]. Available: https://www.healthline.com/health/lung-cancer/cough

[51] "Dry Cough Symptoms," Lung Cancer Center. Accessed: Oct. 30, 2023. [Online]. Available: https://www.lungcancercenter.com/lung-cancer/symptoms/dry-cough/

[52] A. Harle *et al.*, "A cross sectional study to determine the prevalence of cough and its impact in patients with lung cancer: a patient unmet need," *BMC Cancer*, vol. 20, no. 1, p. 9, Jan. 2020, doi: 10.1186/s12885-019-6451-1.

[53] W. Li *et al.*, "Self-reported sleep disorders and the risk of all cancer types: evidence from the Kailuan Cohort study," *Public Health*, vol. 223, pp. 209–216, Oct. 2023, doi: 10.1016/j.puhe.2023.08.007.

[54] E. Gourd, "New evidence that air pollution contributes substantially to lung cancer," *Lancet Oncol.*, vol. 23, no. 10, p. e448, Oct. 2022, doi: 10.1016/S1470-2045(22)00569-1.

[55] "Ambient air pollution." Accessed: Oct. 30, 2023. [Online]. Available: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/ambient-air-pollution

[56] M. C. Turner *et al.*, "Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations," *CA. Cancer J. Clin.*, vol. 70, no. 6, pp. 460–479, 2020, doi: 10.3322/caac.21632.

[57] "Asbestos | NIOSH | CDC." Accessed: Oct. 30, 2023. [Online]. Available: https://www.cdc.gov/niosh/topics/asbestos/default.html

[58] A. Shankar *et al.*, "Environmental and occupational determinants of lung cancer," *Transl. Lung Cancer Res.*, vol. 8, no. Suppl 1, p. S31, May 2019, doi: 10.21037/tlcr.2019.03.05.

[59] N. Ferrara, "7 steps to better nutrition habits for cancer survivors," Mayo Clinic Comprehensive Cancer Center Blog. Accessed: Oct. 30, 2023. [Online]. Available: https://cancerblog.mayoclinic.org/2021/12/30/7-steps-to-better-nutrition-habits-for-cancer-survivors/

[60] "Obesity and Cancer Fact Sheet - NCI." Accessed: Oct. 30, 2023. [Online]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/obesity/obesity-fact-sheet

[61] B. D. Nicholson, W. Hamilton, C. Koshiaris, J. L. Oke, F. D. R. Hobbs, and P. Aveyard, "The association between unexpected weight loss and cancer diagnosis in primary care: a

matched cohort analysis of 65,000 presentations," *Br. J. Cancer*, vol. 122, no. 12, Art. no. 12, Jun. 2020, doi: 10.1038/s41416-020-0829-3.

[62] "Alcohol and Cancer Risk Fact Sheet - NCI." Accessed: Oct. 31, 2023. [Online]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet

[63] "Alcohol Use and Cancer." Accessed: Oct. 31, 2023. [Online]. Available: https://www.cancer.org/cancer/risk-prevention/diet-physical-activity/alcohol-use-and-cancer.html

[64] "Risk Factors: Tobacco - NCI." Accessed: Oct. 31, 2023. [Online]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco

[65] "Harms of Cigarette Smoking and Health Benefits of Quitting - NCI." Accessed: Oct. 31, 2023.[Online].Available:https://www.cancer.gov/about-cancer/causes-prevention/risk/tobacco/cessation-fact-sheet

[66] "Diet and Physical Activity: What's the Cancer Connection?" Accessed: Oct. 30, 2023. [Online].Available:https://www.cancer.org/cancer/risk-prevention/diet-physical-activity/diet-and-physical-activity.html

[67] "Cancer." Accessed: Sep. 16, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[68] "Machine learning, explained | MIT Sloan." Accessed: Aug. 02, 2023. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[69] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[70] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, Aug. 2017, pp. 1–6. doi: 10.1109/ICEngTechnol.2017.8308186.

[71] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *J. Data Anal. Inf. Process.*, vol. 7, no. 4, Art. no. 4, Sep. 2019, doi: 10.4236/jdaip.2019.74012.

[72] "XGBoost Parameters — xgboost 2.0.0 documentation." Accessed: Sep. 13, 2023. [Online]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html

[73] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[74] M. Desai and M. Shah, “An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN),”*Clin.EHealth*,vol. 4, pp. 1–11, Jan. 2021, doi: 10.1016/j.ceh.2020.11.002.

[75] J.Brownlee,“Crash Course on Multi-Layer Perceptron Neural Networks,” MachineLearningMastery.com.Accessed:Sep.11,2023.[Online].Available:https://machinelearningmastery.com/neural-networks-crash-course/

VITA

Kevin De La Rosa is from El Paso, Texas. He studied Mechanical Engineering and earned a Bachelor of Science in Mechanical Engineering from The University of Texas at San Antonio. His future plans include pursuing a Ph.D. in the engineering discipline.