MSc Data Science Master's Thesis

Name: Sammi Kong Pek Chen

Student ID: 20044512

Research Title

"An exploration of the use of machine learning technologies in the early detection of cancer and appropriate treatment protocols through patient medical history."

Abstract

Cancer is a leading cause of deaths worldwide, with an estimated number of 19 million deaths in 2020. (Global Cancer Observatory, 2020.) This research project aims in contributing to awareness of the effects in implementing machine learning algorithms in healthcare diagnosis. Alongside the technological growth of machine learning algorithms, the healthcare industry has the potential to detect diseases early and determine treatment protocols with high accuracy. In this paper, we utilize quantitative research methods to present a machine learning algorithm and data visualization to support the hypothesis. The dataset was processed using Python on Jupyter Notebook on Mac OS for analysis and machine learning algorithm training. Our results indicate a positive correlation of early detection through machine learning methods. There were challenges collecting the appropriate dataset for analysis due to unavailability and privacy constraints. While there are many uncertainties surrounding ethical, security or inclusivity of machine learning models, the literature review and performance of the processing model is promising as an indication of future potential applications. In conclusion, there are existing artificial intelligence and machine learning technologies that have potential to be a major contributing factor in early cancer detection, treatment protocol administration and a higher rate of recovery.

Table of Contents

List of Tables and Figures

Chapter 1: Introduction

The advancement of machine learning technologies has been an exponential growth in the process radicalizing industries. Machine learning is a subset of artificial intelligence (Zhang, X.D., 2020.) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy (IBM Cloud Education, 2020.). The research will be focused on exploring the effectiveness of implementing machine learning algorithms to predict the risk of disease such as cancer and success rate in identification of appropriate treatment protocols. In this chapter, we will have an outline of the research motives and scope that will be achieved from August 2021 to December 2021.

1. Research Motives

Cancer is a leading cause of deaths worldwide, with an estimated number of 19 million deaths in 2020. (Global Cancer Observatory, 2020.) According to The Global Burden of Diseases, it is estimated that 9.56 million people died prematurely as a result of cancer in 2017. (Roser, M. et al., 2015.) If the rate of growth in undiagnosed cancer continues, the effects would be detrimental thus magnifying the urgent need for effective treatment strategies to combat the potential risk.

Alongside the technological growth of artificial intelligence systems, the application of machine learning algorithms within the healthcare industry has the potential in early detection of diseases and determination of appropriate treatment protocols with high accuracy. Even though the global pandemic which is COVID-19 is far from over, it has had devastating effects on patients with cancer, with huge numbers of missed diagnoses and delayed treatments due to a pressured healthcare system and safety concerns for hospital environments which led to reluctance to seek medical care. (The Lancet Oncology, 2021.) These new technologies have the potential to carry out important data analysis through an integrated platform so patients can remotely access a safer way of diagnosis and virtual treatment methods with their GP.

The project will focus on investigating the effects of artificial intelligence implementation within healthcare, more specifically effectiveness of implementing machine learning algorithms to predict the risk of diseases, early detection of cancer and success rate in identification of appropriate treatment protocols.

2. Research Aim and Objectives

2.1 Research Aim

This research project aims in contributing to awareness of the effects on human lifespan and improvement in our daily lives after implementing machine learning algorithms in healthcare diagnosis and treatment stages.

2.2 Research Hypothesis

The implementation of machine learning techniques have contributed to a higher rate of early disease diagnosis and higher rate of predicting effective treatment protocols for patients. There can be a generalized machine learning algorithm that processes large amounts of lab results, clinical characteristics and radiological exam results to predict outcomes of treatments and provide early disease detection.

2.3 Research Questions

1. Does the implementation of machine learning algorithms affect the efficacy of medical disease diagnosis in cancer?
2. What are the existing and new approaches in early cancer diagnosis and determination of treatment protocols for a patient?
3. What are the advantages and disadvantages of utilizing machine learning to analyze patient medical history and data?

2.4 Research Scope

The main focus of this project was the exploration of how the adoption of machine learning algorithms into early cancer diagnosis and treatment advice may contribute to a change in life expectancy of its affected demographic. Due to preliminary developmental constraints of machine learning algorithms and their limited adoption into healthcare systems, our analysis was made as a consideration of a continuously advancing technology. This research was limited in scope by the chosen sample dataset sizes, up to year 2021 and within the UK only.

2.5 Research Objectives

The objectives are to conduct data analysis and visualization of the relationship between utilizing machine learning algorithms in healthcare diagnosis and treatment as compared to lack of it such as the traditional doctor analysis and specialized prescription methods instead. We aim to analyze existing approaches in early cancer diagnosis and determination of appropriate treatment methods. This will be followed with an investigation into their advantages and disadvantages in order to establish a gap in the healthcare industry to finally determine the best machine learning technique and application in the research scope. We will be looking into the efficacies and efficiencies of adopting machine learning to analyze patient medical history and data and their challenges. We will also be exploring whether there are alternative (new) approaches to detect cancers at their early stages.

2.6 Dissertation Structure

The dissertation will be divided into six chapters, each of which plays a specific role to address the research topic in the academic research process. The six chapters of the dissertation are:

1. Introduction

   1.1. Background on cancer diagnosis

   1.2. Introduction

   1.3. Methods of Work

2. Literature Review

   2.1. Statistics on cancer and drivers in requirement for an integrated healthcare system

   2.2. Cancer detection and existing approaches

   2.3. Machine learning for cancer detection

   2.4. Challenges

3. Methodology

   3.1. Introduction

   3.2. Methods and techniques

4. Experimental methods

   4.1. Introduction

   4.2. Evaluation of dataset

5. Research outcomes and discussion

   5.1. Introduction

   5.2. Results

   5.3. Discussion

6. Conclusion

Chapter 2: Literature Review

1. Introduction

Based on the increase of fatal infectious diseases, the United Nations made a prediction of global death toll reaching a staggering high number of 10 million annually by 2050 (United Nations, 2019.), magnifying the urgent need for an accelerated detection method and effective treatment strategies to combat the growing figures. Whilst not being infectious, cancer is a leading cause of deaths worldwide, with an estimated number of 19 million deaths in 2020. (Global Cancer Observatory, 2020.) According to The Global Burden of Diseases, it is estimated that 9.56 million people died prematurely as a result of cancer in 2017. (Roser, M. et al., 2015. From Figure 1, the number of newly registered cancer patients has shown a steady increase annually from year 1971-2017. This indicates an increase in demand for early cancer detection methods and effective treatment strategies to combat the issue.

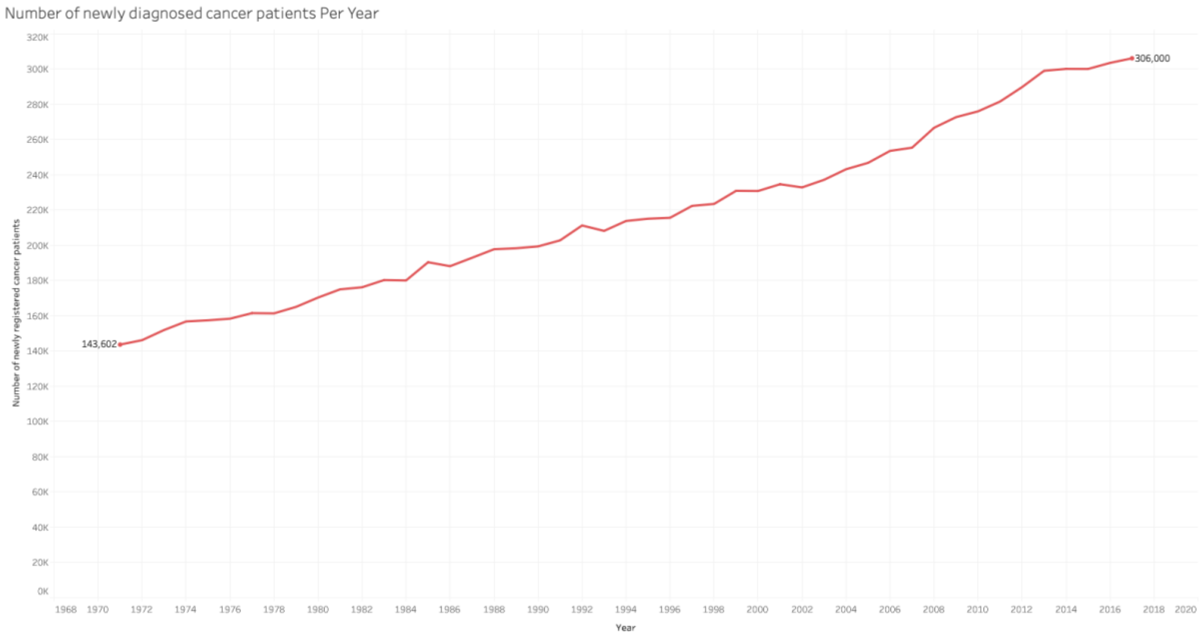
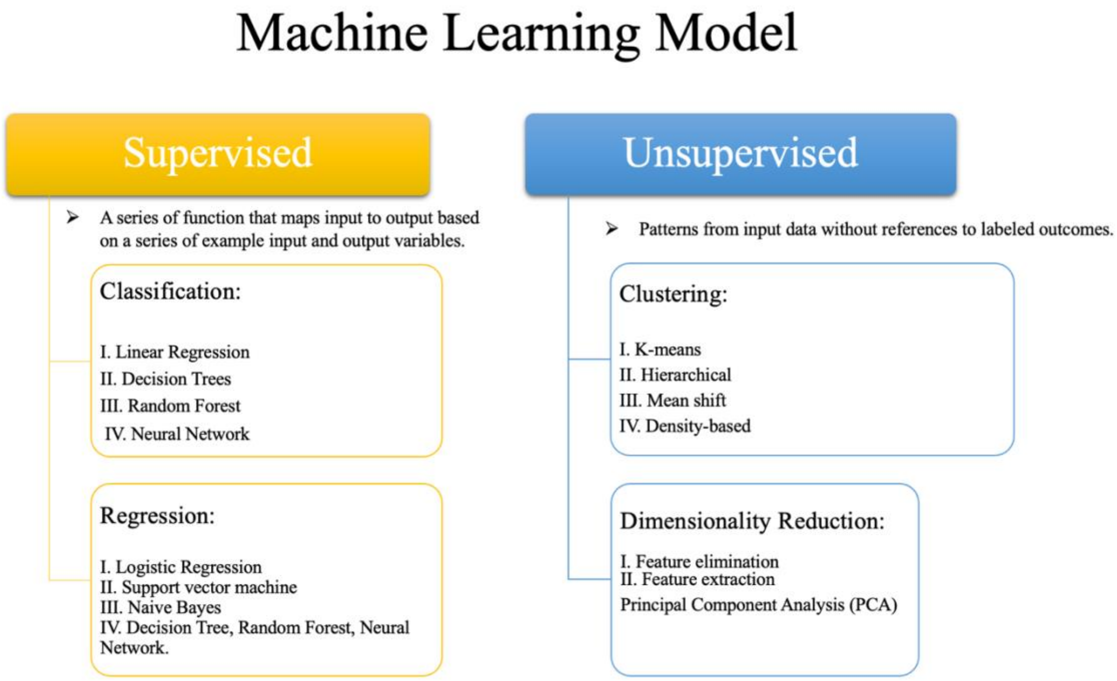


Figure 1: Number of newly diagnosed cancer patients between 1971-2017. The data obtained from the Office for National Statistics UK. The dataset includes cancer diagnoses, age standardized incidence rates for all types of cancer by age and sex.

The National Cancer Institute define a tumor as an abnormal mass of tissue that results when cells divide more than they should or do not die when they should. Tumors may be benign or malignant. Benign tumors may grow large but do not spread into, or invade, nearby tissues or other parts of the body but malignant tumors do. They can also spread to other parts of the body through the blood and lymph systems. (National Cancer Institute, 2021.) Some of the causes of cancer are attributable to key lifestyle and environmental risk factors. These include obesity, infections, excessive exposure to radiation or alcohol consumption. Smoking, a high-fat diet and working in an environment with a high risk exposure to toxic chemicals can affect the adult demographic. (Stanford Health Care, 2021.) Cancer is prevalent in occurring within family medical history, which means it can be genetically inherited. Once our immune system is compromised, its functional features of protecting our bodies from infection and disease is affected. One theory suggests that the cells in the bone marrow, the stem calls, become damaged or defective, so when they multiply into more cells, they are produced into abnormal cells or cancel cells. (Stanford Health Care, 2021.) While not all chemicals and substances from the environment are harmful, exposure to some of these chemicals may damage our DNA and be dangerous to our health, these chemicals can include asbestos, formaldehyde, radon, secondhand tobacco smoke, soot and wood dust. (Hantel, A., 2018.) The risk of developing cancer after coming into contact with carcinogens depends on the length of exposure and level of radiation. According to professionals, it is best to reduce or prevent any form of exposure at all however tough that may be in a world filled with some form of pollution or the other. Conveniently, there are smart meters in the form of artificial intelligence that could serve us well in measuring levels of radiation exposure in dangerous environments without risking the lives of any humans in the same situation. If the rate of undiagnosed cancer continues, the effects would be detrimental, magnifying the urgent need for effective treatment strategies to combat this.

Even though the global pandemic which is COVID-19 is far from over, it has had devastating effects on patients with cancer, with huge numbers of missed diagnoses and delayed treatments due to a pressured healthcare system and safety concerns for hospital environments which led to reluctance to seek medical care. Despite reassurance from officials that the UK's National Health Service (NHS) remained open for urgent care, a study estimated that as many as 45% of those with potential cancer symptoms did not contact their doctor through March to August 2020 in the first lockdown. (The Lancet Oncology, 2021.) Thus, the ongoing global pandemic COVID-19 has highlighted the accelerated need for an integrated system to handle digital and data infrastructures (NHS England, 2021.) from across the board to streamline patient health records for general practitioners under pressing time constraints. As part of the solution, there is a need for newer technologies such as fully integrating machine learning technologies with predetermined patient indicators to carry out the important data analysis through its system so patients can remotely access a safer way of diagnosis and virtual treatment methods with their GP which saves time, cost and avoids human errors.

Alongside the technological growth of artificial intelligence systems, the application of machine learning algorithms within the healthcare industry has shown potential in early detection of diseases and determination of appropriate treatment protocols with high accuracy. The broad spectrum of patient care covered by medical practitioners can be eased with the adoption of machine learning into healthcare systems. Machine learning involves setting pre-defined programming codes to learn from repeating written commands and storing the results as a way of recycling and continuously learn from the knowledge. Deep learning is a step further advanced than machine learning, where it is based on the structure of neural networks similar to a human brain which allows the program to identify patterns without pre-defined programming codes. There are 2 types of machine learning – supervised and unsupervised.



Machine Learning Model

Supervised

- A series of function that maps input to output based on a series of example input and output variables.

Classification:

I. Linear Regression
II. Decision Trees
III. Random Forest
IV. Neural Network

Regression:

I. Logistic Regression
II. Support vector machine
III. Naive Bayes
IV. Decision Tree, Random Forest, Neural Network.

Unsupervised

- Patterns from input data without references to labeled outcomes.

Clustering:

I. K-means
II. Hierarchical
III. Mean shift
IV. Density-based

Dimensionality Reduction:

I. Feature elimination
II. Feature extraction
Principal Component Analysis (PCA)

Figure 2: Types of Machine Learning Models and methods of function.

Supervised machine learning requires training the machine using labelled data to predict the outcomes of unforeseen data. (Johnson, D., 2021.) The process of building, scaling and deploying an accurate supervised machine learning model requires the results to be reproducible using other test data sources. Unsupervised machine learning models utilize unlabelled data to discover patterns without any supervision. This method can process more complex tasks to find features or patterns which can be useful for categorization.

Machine learning programs can identify patterns in image recognition over time and aid the human practitioner during each stage of diagnosis and prevention. A widely used application of this technology is prominent in biomedical imaging. Computer-aided detection models are utilized in imaging scans to predict the risk of breast cancer, classify lesions and predict treatment responses and clinical outcomes. As medical science advances, life expectancy raises which produces a growing demand for services, rise of costs and heavier workload on the workforce. In a report on world population growth, it states that by 2050, 1 in 4 people in

Europe and North America will be aged over 65, meaning more complex health issues will need handling. (World Population Prospects, 2019). This may cause healthcare systems to become unsustainable if it continues to be run at a manual level.

Another radical example of utilizing artificial intelligence in the healthcare and surgical environment are the manufacture of medical equipment with embedded systems. An embedded system is a combination of computer hardware and software designed for a specific function which in this case contain sensors and control mechanisms. Specialized medical equipment as such, much like industrial machines, are required to be designed as user-friendly as possible so that human health does not get jeopardized by preventable machine mistakes which means they often include a more complex operating system and graphical user interface (Lutkevich, B. 2020.) designed for ease of access and should be easily understood at any level.

In 2019, a total of 1.2 million people had operations with the aid of an American robot called Da Vinci. (Orchard, R., 2019.) It represents one example of newly human-controlled specialized surgical robots that have been able to deliver high patient outcomes such as sealing and stitching wounds at higher precision compared to a human practitioner. This era of technological advancement has brought in new business players keen to revolutionize surgical and treatment management methods in the industrial 4.0. This includes upstart British company CMR Surgical who launched their Versius robotic arm in October 2019 and has since been part of the success of a “COVID protected” robotic surgical centre. (Huddy, J.R., et al., 2021.) Huddy, J.R. et al had highlighted the strengths of using robotic surgical units during the era of the coronavirus pandemic, a highly transmissible respiratory disease. There were the increased health safety for both the patients and staff acknowledged by added hygiene steps such as contamination prevention measures and a shorter length of patient stay of 2 days compared to an average of 6 days pre-coronavirus as there were now dedicated staff members assigned to the dedicated surgical department. In terms of weaknesses, they identified the difficulties of staff shortages and isolation periods to ensure a level of safety for the cancer patients. The data they collected were quantitative, based on in-patient satisfaction surveys collected upon discharge. This research supports the hypothesis that a wider use of artificial intelligent systems in hospitals in the future could potentially ensure a high level of cleanliness and a lower risk of infection, provide faster recovery and decreased length of in-patient stay.

Gosrisirikul, C. et al. (2018) predicted the future of robotic companies based on various robotic system developments and highlighted the part the pioneering company Intuitive Surgical played in delaying competition by building high entry barriers using intellectual property protection, customer relationship maintenance, overcoming regulatory issues, and setting up training centers worldwide for almost the last decade. Reflecting upon their findings, since 2018 there has been plenty of growth from similar competitors in the robotic surgical device field. Restricted by entry barriers, competitors would only be able to research and develop their product with a low chance of early market release which could contribute to a higher quality end product delivery given their extended timeline. The research highlighted the importance of a haptic feedback feature in robotic surgical systems due to how it increases surgical awareness and provides additional security. The strengths in the article were in highlighting surgery robotic systems as an emerging technology were able to perform turn higher risk operations into low risk operations with greater accuracy and higher precision rates. The weakness found in this work is the lack of comparison of robotic surgical device performances in the same background application, each products benefit and disadvantage upon implementation within the surgical field listed in their findings. Hashizume, M. et al. (2004) produced a paper supporting a similar research identifying the problems faced with the development of robotic surgery for cancer. After the built these machines, there is training required for people to be skilled enough to understand their function, translate three dimensional images, create the desired computer graphics for user usability and utilize of visualization methods to produce appropriate simulations. In addition, they highlighted the lack of studies reviewing indications of robotic surgery for cancers. These are strong points that support the statement that the rate of technological advancement needs to be kept up with human understanding in order to produce any real impactful result in the healthcare sector.

An investigation into the growing use of natural language processing (NLP) in AI means that NLP allows machines to comprehend human communication and process as an interaction or service. It has radically changed the workforces in hospitality, industries and even administrative roles. Most online retail website utilize this technique with AI chatbots that redirects clients to the right services on the page, or is “emotionally intelligent”, able to analyze and speak to the client in real time based on predetermined keyword constraints. This has replaced the need for excessive amounts of customer service representatives to be on call, saves cost and time for the companies. Household assistants such as Alexa and Google Home devices function similarly, being able to control entire home electrical systems under a simple voice command. In terms of healthcare file keeping, NLP could assist medical physicians in automated filling in EHRs to relieve humans from conducting automated work, save time to consult with patients and improve work satisfaction. (V Soft Consulting, 2021.) NLP trained models could act as the first point of contact in clinics and hospitals before a human consultation is required. The basic understanding and processing of human language allows a ‘robot receptionist’ in theory to replace the need for front house staff to greet humans before being led straight to their specific needs practitioner or department.

2. Cancer Detection and Existing Approaches

The traditional approaches involved in cancer diagnosis include lab tests, imaging tests and running a biopsy. The least invasive methods are through lab tests which detect levels of certain substances in blood, urine or other bodily fluid samples. (NCI, 2021.) Blood tests measure the level of substances in your body, such as red blood cells, white blood cells and markers of inflammation. However, blood tests are not definitive. (Aijboye, T, 2021.) For imaging tests, circulating tumor markers can aid in identifying tumors, stages of cancer, assess the effectiveness of treatments, estimate prognosis and even detecting traces of residual cancer post-treatment.

Imaging Tests

Imaging tests produce pictures of areas inside your body for the doctor to identify whether a tumor is present. Among the many imaging tests there are computerized tomography (CT) scans, magnetic resonance imaging (MRI), nuclear scans, bone scans, positron emission tomography (PET) scans, ultrasounds and X-rays.

CT scans are done using an x-ray machine to create detailed 3D images (Mayo Clinic, 2020.) and an MRI uses powerful magnetic and radio waves to produce detailed images of the inside of the body in slices. (NHS UK, 2018.) A nuclear and bone scan requires a radioactive “tracer” material to be injected into your bloodstream then a machine measures the radioactivity in your body to produce the image. (Benisek, A., 2020.) A PET scan uses small amounts of radioactive materials called radiotracers or radiopharmaceuticals, a special camera and a computer to evaluate organ and tissue functions. By identifying changes at the cellular level, PET may detect the early onset of disease before other imaging tests can. (RadiologyInfo, 2021.) Ultrasound involves the use of high frequency sound waves to produces images known as sonograms. Ultrasounds are used to assess areas filled with fluid or to help diagnose cancers located in areas that do not show up clearly on X-rays. X-rays are fast, painless tests that use low doses of radiation to obtain images of parts of the body. In some cases, a contrast dye is used to make the imaging results show up clearer. (Ajiboye, T.,2021.)

Advantages and Disadvantages

The advantages in the existing approaches are a high accuracy in finding outcomes. The tests are conducted by specialists who are proficient in handling X-ray machines and run lab results. Screening can detect cancer at an early stage, ensuring a higher survival rate from early treatment methods. In England, almost all women diagnosed with breast cancer at the earliest stage survive their disease for at least 5 years. (Cancer Research UK, 2021.) There are many types of cancer screening programmes available to the public in the UK as early detection and treatment is key in improving the chance of survival.

There are risks involved with screening tests that need to be considered, some of which are misdiagnoses that can occur with medical equipment handling errors or false positive results. Long delayed wait times are a usual occurrence in hospitals, especially in poorer countries due to lack of infrastructure, or the lengthy processing times and combination of tests that are required to achieve results. The screening tests involved can be high cost due to the specification of machines and follow-up diagnostic tests. The imaging tests used to diagnose tumors expose test subjects to radiation, which in excess could contribute to the body producing cancerous cells. There should be safer ways of diagnosis without a potential health risk or compromising accuracy in results.

3. Machine learning for Cancer Detection

All the mentioned imaging test results in the existing approaches section require a specialist to fully function the devices to identify traces of cancer. However, with the help of machine learning algorithms in biomedical imaging that have been trained to identify pre-programmed patterns within image recognition, we can increase the rate of diagnosis effectively. As an example, computer-aided detection models are utilized in breast imaging to predict the risk of breast cancer, classify lesions and predict treatment responses and clinical outcomes. (Bitencourt, A., et al., 2021.) Among other potential machine learning developments in healthcare, there are many companies exploring methods to use this technology in telemedicine. This technology development aims to provide doctors with patient information during a telemedicine session, whilst capturing information during the virtual visit to assist with an increase in efficiency and workflow. (Thomas, M., 2021.)

Development of an effective computer-aided diagnosis (CAD) system is of great clinical importance and can increase the patient’s chance of survival. According to a research by El-Baz, A. et al. (2013), the success of a particular CAD system can be measured by its accuracy of diagnosis, speed and automation level. The research reviewed different CAD systems in application of lung cancer diagnosis. At the first stage, the model was trained to detect nodules into chest image segmentations using a pattern recognition technique called template matching. The filtered candidates are further processed for classification using a feature-based classifier. Due to the nature of research, constant changes are expected due to inhomogeneities in the lung region and pulmonary structures. Important factors such as the automation level, speed and ability to detect nodules of different shapes and sizes need to be considered to develop an effective solution for diagnosis. The lack of other existing CAD systems to compare the research outcomes also prevents efficient validation of the proposed approaches in screening for cancerous nodule for the time being.

In another research paper by Onan, A. (2015), they investigated a classification model based on fuzzy-rough nearest neighbour machine learning algorithm method to diagnose breast cancer. Based on unsupervised estimation and noise modelling, the fuzzy C-means clustering method has the ability to recognize large groupings of cancerous cells within the intensity distributions obtained from the PET images. These results may show tumor delineation with respect to the thresholding-based methods. Their evaluation metrics to measure the performance of the model were classification accuracy, sensitivity, specificity, F-measure, area under curve and Kappa statistics. The values obtained for sensitivity and area under curve were 1, and the specificity, F-measure and Kappa statistics returned a value of >0.99 with a classification accuracy of 99.71%. The research highlighted the importance in pre-processing databases for analysis to enhance efficiency in the final machine learning algorithm model. The research could potentially improve outcomes by including other datasets to run in comparison to test whether it is feasible to diagnose other types of cancer. However, the success of the research outcome represented by high performance rate for the machine learning algorithm in breast cancer diagnosis, makes it potentially suitable for wider use in future medical diagnosis or personalized healthcare screening system.

Another common application of machine learning in healthcare is deep learning and machine learning in radiomics. Radiomics involves mining of quantitative image features from standard-of-care medical imaging that enables data to be extracted and applied within clinical-decision support systems to improve diagnostic, prognostic, and predictive accuracy (Lambin, P. et al., 2017) and has grown to be largely involved in cancer research. Radiomics utilize quantitative image features based on intensity, shape, size or volume, and texture to determine tumor phenotypes and microenvironment. (Gillies, R.J., et al, 2015.)

In a research conducted by Murphy, D.R., et al. (2002), they investigated the effect of utilizing electronic health records (EHRs) for data mining and developed ‘trigger’ algorithms to detect potential high risk cancer patients that required urgent follow up consultations. Among their chosen dataset, there were 11.6% patients who were positive in showing subsequent signs of cancer which is a high proportion given that these patients would eventually worsen in health without any diagnosis. The strengths during this research were the careful consideration for what a ‘delay’ meant for both the process of training the algorithm to identify ‘triggers’ in the EHRs and the stages of which to alert the health providers to prevent information overload. This pre-definition could save cost, time and effort whilst avoiding backlog issues from arising. They identified many challenges in implementing the trigger algorithms widely, such as considering inclusivity or exclusivity criteria during evaluation, location and demographical adjustments, allowing a professional to make the final decision on contacting the patient for a follow-up after reading the alerts and further improvement on the data extraction and mining process was required. The methods of improvement for this research are by determining the main few general criteria that would return with a higher than 11.6% success rate of detection. This goal could be seen as ambitious, due to the nature of cancer research and mutation patterns in genes occurring that are not fully understood by specialists yet. Producing weightages of importance for each criteria evaluated could increase the accuracy of the data training model. One main criteria that should be included with a higher weightage is if cancer has occurred in the patient family history. Conclusively, the study can serve as a good basis for future research and development on trigger algorithms with higher accuracy and specific to the type of cancers.

The research by Manogaran, G., et al. (2017) utilized a Bayesian hidden Markov model with Gaussian Mixture clustering approach to similarly process DNA sequencing to diagnose early detection of cancer through genetic changes in comparison with other existing approaches. The main challenge addressed was the large DNA sequence data size, strengthening the need for a scalable machine learning approach to overcome the data mining difficulty. The performance of the proposed Bayesian HMM based change detection algorithm produced an 80% accuracy rate, slightly lower than other existing approaches mentioned such a Pruned Exact Linear Time (PELT) algorithm, binary segmentation algorithm and segment neighbourhoods so further quality control to improve the model is required.

A machine learning predictive model was developed to predict the response of cancer cell lines to drug treatment based on genomic features of the cell lines and chemical properties of the drugs in a research by Menden, M.P., et al (2013). The aim of the research was to produce a functional computational framework in identification of new drug repositioning opportunities through in silico models that simplify the drug-cell screening process. In future applications of personalized medicine, specific genomic traits of patients can be linked into the framework to predict the success rate of treatment protocols to determine the best outcomes. This approach of treatment could significantly improve the rate of recovery compared to current trial-and-error methods that do not guarantee success and comes with the risk of side effects. In order to produce an effective well-trained machine learning predictive model as attempted by Menden, it is common to face challenges adhering to timeline constraints due to many training tests or cross-validation tests to achieve maximum results. This needs to be taken in consideration when building future models or improving current models in use.

4. Challenges

In the early 2000's, the proposition of artificial intelligence being incorporated into the healthcare system has been a controversial topic within the public community due to a lack of understanding on the potential implications of disruption in our lives. The discussion on ownership, usage and security of acquired health data are currently among the challenges that could cause a major effect in data protection and hindering the implementation of machine learning techniques in the advancement of cancer detection. Machine learning that involve deep learning are commonly used in image analysis which can identify potentially cancerous lesions in radiology image results. The common challenges in the adoption in daily clinical practice were regulations, integration and standardization problems, however currently ethical issues are prevalent in convincing people to allow the implementation itself and will prove to be a continuous issue to consider.

Automation has the potential to free humanity from the fetters of repetitive, physically demanding and often, unpleasant work. (McKinsey & Company, 2016.) It means humans are then allowed the freedom to work on more pressing matters and harness their creativity. Automated artificial intelligent technologies allows streamlined and revolutionized future healthcare options, potentially soon providing personalized and real-time treatment for all of us. There are many advantages listed with examples throughout the literature review but there still remains fear into how radically our lives would change with the growing implementation of machine learning systems in all industries. One argument would be the drastic changes employment will experience as a whole in the economy. Robots and early AI have become proficient enough to displace vast tracts of predictable physical work and data process tasks. Hence, tasks that involve big data or the performance of physical activities and operating machinery in predictable patterns will be replaced by new digital technologies. According to a report by Deloitte, as we begin to enter this Fourth Industrial Revolution, it is estimated that by 2025, 35% of roles could be entirely automated using the technology we have available to us now. (Deloitte, 2021.) This means that a third of current work opportunities can be replaced by automated machines in less than 5 years. These replaceable occupations do not include work that is requires adaptability, creative intelligence or social intelligence. To prepare for the future dynamic of work, government initiative is required to work with private sectors to grow with the pace of technology to ensure a high living standard, a better labor market, promote a constant learning mindset and providing assistance to communities that are heavily impacted. (Muro, M. et al., 2019.) A growing number of new jobs to manage data, new technologies and human resourcing will surface as machine learning advances. The government plays a large role of working alongside private sectors to keep up with the pace of technology. Efforts to promote training and accelerated learning, education and certifications on operating these new technologies are key to deliver quality artificial intelligence in healthcare. The initiative will contribute to making digital skills development more financially accessible, mitigating the harsh impacts of automation on the lifestyle of lower income communities. Training in skills such as basic digital literacy, the fundamentals of genomics, and machine learning methods (Spatharou, A. et al., 2020) need to become mainstream as an upskill option for all practitioners to keep up with the digitalized era.

In terms of implementation into processing healthcare databases, a major argument is that the system wide private information sharing would be an issue once machine learning technologies are adopted widely when disease diagnosis is run in place of the traditional preliminary health checks with human general practitioners. Cancer care coordination and symptom management requires close communication between imaging, pathology, genetic and treatment departments. (Moser, E.C., et al, 2020.) Experts need to share opinions upon diagnosis and best treatment methods through data sharing. The various types of health data are subject to strict security, legal and regulatory requirements which can be a limiting factor in the storage of private health records. Blockchains are tamper evident and tamper resistant digital ledger implemented in a distributed fashion (i.e., without a central repository) and usually without a central authority (i.e., a bank, company or government). At their basic level, they enable a community of users to record transactions in a shared ledger within that community, such that under normal operation of the blockchain network no transaction can be changed once published. (Yaga, D., et al., 2018.) In the effort to tackle the issue on security, blockchain can be used to store and distribute AI models to provide an audit trail, hence enhancing data security. (IBM, 2021.) Hence, the basic stages of data security could be solved by using the blockchain technology for privacy protection. In the paper by McGhin, T. et al.(2019), they addressed how blockchain can be adopted in healthcare applications due to its many benefits. The listed application benefits were fraud detection, introduction of smart contracts and identity confirmation. They managed to identify the gap in the healthcare industry where this technology could be of best service to the community. The tracked end-to-end transaction process allows both parties to have a fair exchange and ensure security of information while preventing fraud by encoding unique numbers for easy verification.

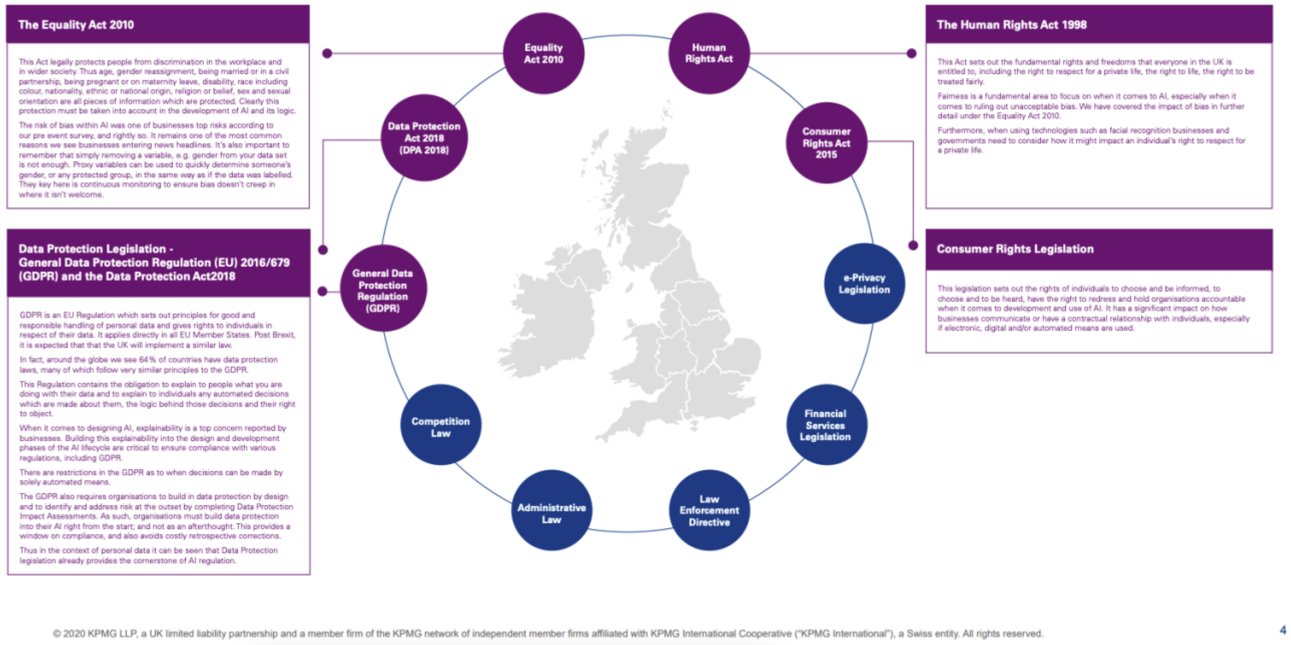
In the effort to fight cancer, many foundations are participating in scientific research projects. However, research grants to advance machine learning technologies within the healthcare industry take a long time to be approved through governing bodies and the research stages itself has a high attrition rate. It is common to extend project timelines to achieve a conclusion or figure out a solution (if any). Prominent global challenges during the stages of drug development for disease control also include lengthy development timelines, attrition rates and production stagnation rates. The delay in this initial developmental stage causes a backlog that could well affect thousands and into millions of lives. Current delays in cancer diagnosis cost lives, as the earlier it is detected, the better the patient outcomes in life expectancy. Early detection and diagnosis in cancer treatment involve data mining and machine learning techniques that combine comprehensive health records to evaluate the risk factors of patients in early cancer diagnosis. Certain attributes such as family history, blood pressure levels, heart pulse rates, oxygen saturation levels all contribute to specific probabilities to produce the total percentage of risk and flag warnings in a patient's records for an expert to further evaluate and carry out confirmation tests.

Data cleansing is a challenging key first step for conducting any data analysis or data mining. It is essential in improving the quality and reliability of all data. (McKelvey, N. et al., 2016.) Initially sourced data for machine learning models to process could characterize to massive sample dataset sizes and high dimensionalities. Any pre-filtered datasets could introduce computational and statistical challenges resulting in scalability and storage issues, measurement errors, incidental endogeneity, noise accumulation and spurious correlation. (Fan, J. et al, 2014). There are tools which are available to convert and transform data in the market, but are expensive and can be ineffective. It is vital that more healthcare staff are educated in data handling to understand on a fundamental level what types of data they are pre-processing to reduce wasted time in between stages of analysis by machines.

Machine learning techniques can effectively aid the early detection and outcomes of cancer treatments through pattern recognition and classification and contribute to a higher survival rate of patients with existing health history records. IBM Watson as a technology was intended to be in application of solving healthcare problems when it launched as an artificially intelligent technology back in 2011, expanding as a company and making several major acquisitions and increasing their headcount of data scientists. (IBM, 2021.) They faced difficulties in achieving realistic ROI targets and goals such as fighting against cancer and other chronic diseases. The identified determinant for the technology's failure was the lack of intended application for the use of their end product. From this example, we can understand the significance of predefining goals and outcomes to ensure goal alignment throughout any type of work effort.

Factors such as conditioned bias in politics, economical and commercially ingrained medical practice norms remain an obstacle when it comes to implementation of machine learning tools within the adoption in healthcare. According to an article by Panch, T. et al. (2019), most healthcare organizations lack the data infrastructure required to collect sufficient data to optimally train algorithms to fit the local population or practice patterns, and that rigorous evaluation and re-calibration must continue after implementation to track patient demographics and practice patterns which can change overtime. This challenge can be overcome if the trained algorithms and developers collate data from all demographics to ensure a higher accuracy for provisional testing and implementation to ensure the end product suited all demographics and types of application. Most importantly, the source of data is required to be updated constantly to keep up to date with new possible patterns to train the models.

The UK's digital technology sector had an estimated turnover of £170 billion in 2015. (Free, R. et al, 2021.) The current laws in place in the UK as the leading country in AI technology and data revolution are the responsibility of the Department for Digital, Culture, Media & Sport (DCMS) and Department for Business, Energy, and Industrial Strategy (BEIS). The general data protection regulation (GDPR) has addressed many concerns on the basis of handling personal data and privacy. However, there is currently no AI specific legislation due to the fact that the law surrounding the topic have to be technology agnostic to apply for future growth in the industry. There are a number of laws which businesses are required to comply with when developing and using AI technologies. There are the Equality Act 2010, The Human Rights Act, Consumer Rights Act 2015 and Data Protection Legislation. (KPMG, 2020.)



Equality Act 2010
Human Rights Act
Data Protection Act 2018 (DPA 2018)
Consumer Rights Act 2015
General Data Protection Regulation (GDPR)
e-Privacy Legislation
Competition Law
Financial Services Legislation
Administrative Law
Law Enforcement Directive

The Equality Act 2010

This Act legally protects people from discrimination in the workplace and in wider society. Thus age, gender reassignment, being married or in a civil partnership, being pregnant or on maternity leave, disability, race including colour, nationality, ethnic or national origin, religion or belief, sex and sexual orientation are all pieces of information which are protected. Clearly this protection must be taken into account in the development of AI and its logic.

The risk of bias within AI was one of businesses top risks according to our pre event survey, and rightly so. It remains one of the most common reasons we see businesses entering news headlines. It's also important to remember that simply removing a variable, e.g. gender from your data set is not enough. Proxy variables can be used to quickly determine someone's gender, or any protected group, in the same way as if the data was labelled. They key here is continuous monitoring to ensure bias doesn't creep in where it isn't welcome.

Data Protection Legislation -
General Data Protection Regulation (EU) 2016/679 (GDPR) and the Data Protection Act2018

GDPR is an EU Regulation which sets out principles for good and responsible handling of personal data and gives rights to individuals in respect of their data. It applies directly in all EU Member States. Post Brexit, it is expected that that the UK will implement a similar law.

In fact, around the globe we see 64% of countries have data protection laws, many of which follow very similar principles to the GDPR.

This Regulation contains the obligation to explain to people what you are doing with their data and to explain to individuals any automated decisions which are made about them, the logic behind those decisions and their right to object.

When it comes to designing AI, explainability is a top concern reported by businesses. Building this explainability into the design and development phases of the AI lifecycle are critical to ensure compliance with various regulations, including GDPR.

There are restrictions in the GDPR as to when decisions can be made by solely automated means.

The GDPR also requires organisations to build in data protection by design and to identify and address risk at the outset by completing Data Protection Impact Assessments. As such, organisations must build data protection into their AI right from the start; and not as an afterthought. This provides a window on compliance, and also avoids costly retrospective corrections.

Thus in the context of personal data it can be seen that Data Protection legislation already provides the cornerstone of AI regulation.

The Human Rights Act 1998

This Act sets out the fundamental rights and freedoms that everyone in the UK is entitled to, including the right to respect for a private life, the right to life, the right to be treated fairly.

Fairness is a fundamental area to focus on when it comes to AI, especially when it comes to ruling out unacceptable bias. We have covered the impact of bias in further detail under the Equality Act 2010.

Furthermore, when using technologies such as facial recognition businesses and governments need to consider how it might impact an individual's right to respect for a private life.

Consumer Rights Legislation

This legislation sets out the rights of individuals to choose and be informed, to choose and to be heard, have the right to redress and hold organisations accountable when it comes to development and use of AI. It has a significant impact on how businesses communicate or have a contractual relationship with individuals, especially if electronic, digital and/or automated means are used.

© 2020 KPMG LLP, a UK limited liability partnership and a member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative ("KPMG International"), a Swiss entity. All rights reserved.

4

Figure 3: Current laws in the UK for businesses to comply with to develop and use artificial intelligent technologies. (KPMG, 2020.)

The argument is that implementing AI solutions will bring difficulty in terms of sustainable change. Whether it be a lack of funding or lack of cooperation from different demographics during data collection, most healthcare organizations lack the data infrastructure in terms of hardware and software it needs to store, process and transmit healthcare data or properly train machine learning algorithms to conduct localized testing methods for a specific population or consider preferential bias patterns within the community. Businesses are interested in scaling solutions fast, contradicting a health practitioner's desire to build clinical evidence of quality and effectiveness. (Spatharou, A. et al., 2020) There needs to be a middle ground for both parties to work comprehensively together and build an effective system to combat correlating problems. The 3 major cloud technology companies that offer services such as machine learning, data analytics, cloud native development, application migration to their clients are

Microsoft, Google and Amazon Web Services. In the third quarter of 2021, Amazon Web Services controlled 32% of the entire cloud infrastructure services market share, Microsoft Azure with 21% market share and followed by Google Cloud with an 8% market share. (Mlitz, 2021.) As cloud computing technology advances, the requirement for companies to store their exponentially growing amount of data contributes to higher performance rate expectations in contrast to a reduction in cost as software architectural companies compete to deliver desired outcomes. Especially in the private sector, only a handful of large technology companies with significant market power and strong commercial interests in other sectors provide this service which could result in a conflict of interest.

Chapter 3: Methodology

1. Introduction

In this chapter, we will be exploring the methods of research design, and justification of chosen methods. We will be exploring the use of machine learning algorithms in the early detection of cancer through feature selection in patient medical records. This report utilizes data from publicly available open resources Kaggle and depositories on Github to support the chosen quantitative data analysis techniques. Secondary resources were used to collect the datasets required to analyze how survival rates perform annually as ML technological growth in healthcare systems are implemented. The dataset was processed using Python on Jupyter Notebook on Mac OS for further analysis and machine learning algorithm training. There were challenges collecting the appropriate dataset for analysis and training which will be discussed in the discussion section.

2. Methods and techniques

Statistical analysis is the process of understanding how variables in a dataset relate to each other and how those relationships depend on other variables. Visualization can be a core component of this process because, when data are visualized properly, the human visual system can see trends and patterns that indicate a relationship. (Waskom, M., 2012.)

The main research methodology is quantitative research by presenting the machine learning algorithm to run data analysis to predict early detection of cancer and data visualization methods using Tableau to compare and estimate projections of survival rates of cancer patients alongside the existence and advancement of machine learning technologies. A descriptive and correlational analysis was conducted in the Python language to support the hypothesis that machine learning algorithms has the potential identify patterns for early detection of at a high accuracy rate. Upon background research on current implementation of different machine learning algorithms in the literature review, the justification for conducting our analysis through machine learning was the ability to process large amounts of data and produce visual graphs to identify patterns.

In our analysis, we compared reports on survival rates of cancer patients through case studies and statistics collected from the Office for National Statistic UK which can be found on https://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/all-cancers-combined#heading-One. The data was originally produced to compare cancer incidence, mortality, survival and risk factors by cancer type. The dataset used for testing the machine learning algorithm was obtained from https://data.world/cancerdatahp/lung-cancer-data to detect evidence of early diagnosis for lung cancer. We conducted data analysis to investigate the effect AI has contributed to the healthcare industry and extending survival rates, whether machine learning algorithms could detect diseases in the early stages through patient data and analyze machine learning methods to predict diseases accurately.

Chapter 4: Experimental Methods

1. Introduction

In this section, each of the experimental methods are explained in detail. We analyzed the datasets to identify trends and patterns in which visualize the correlation of combined attributes (input variables) that effectively detect cancer as an output. This study evaluates the effectiveness of the machine learning algorithm in analyzing factors that lead to early detection of cancer.

2. Evaluation of dataset

The dataset was collected from https://data.world/cancerdatahp/lung-cancer-data. It consists of 1000 records. Figure 4 shows the fundamental steps for data mining to produce the best research outcomes:

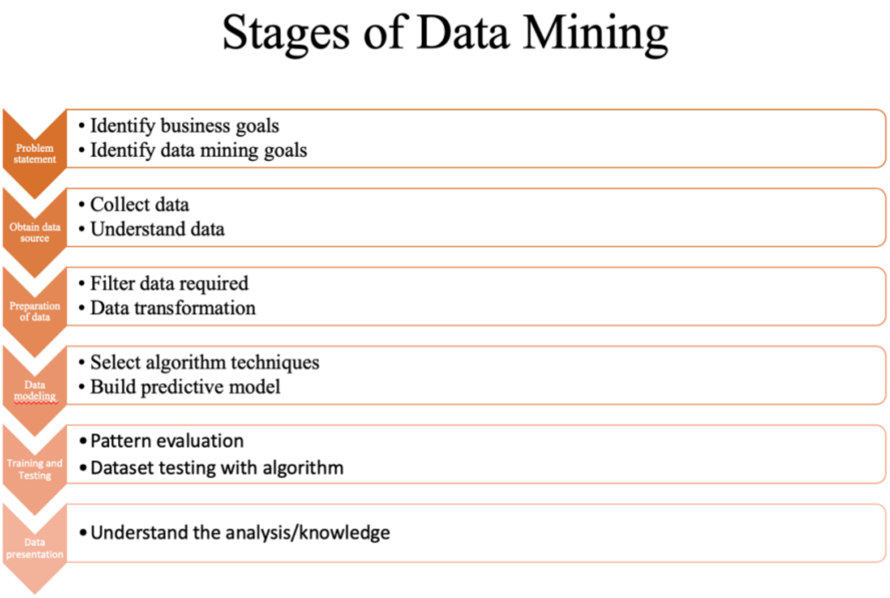


Figure 4: Stages of data mining involved for data analysis.

Data mining describes the method used to extract information from the large datasets to identify patterns through machine learning, statistics and database systems. The attributes in the dataset were checked to ensure no null values were included using Python language, to return a ‘clean’ data set. Firstly, the dataset is analyzed by previewing the first 5 rows using head(). The machine learning test algorithm was run against the chosen dataset to detect whether the feature selection of health issues can aid in early cancer detection and its treatment outcomes. In training the ML algorithm, feature selection is key for an efficient and effective research output. (Dash, M et al., 1997.) Feature selection is the process of reducing the number of input variables when developing a predictive model. (Brownlee, J., 2019.) Filter-based feature selection methods use statistical measures to score the correlation of dependence between input variables that can be filtered to choose the most relevant features. It is typically required to carry out data cleaning and filtering from the original dataset before being able to perform data visualization due to excessive large amounts of data that can populate the chosen datasets. Huge dimensionality and noise are common characteristics in medical databases.

After initial analysis, the machine learning model was split into 80% training model and 20% test set for accuracy testing after training. Datasets that contain many variables require feature scaling to transform the values into a similar scale for evaluation. StandardScaler function was implemented building the code to distribute the data to have a mean value of 0 and standard deviation of 1 to create a common scale for interpretation. We then tested 3 different models to compare prediction accuracies to improve measured accuracy.

Chapter 5: Research Outcomes and Discussion

1. Introduction

In the experimental analysis, we will be reviewing our results and forming a discussion on our analysis.

2. Results

Figure 5 show the average survival rates of adult cancer patients has increased from 50% to 70.4% in year 1971 up to year 2011. This statistic shows great progress for healthcare throughout the years, magnifying the importance of technological advancements in science.

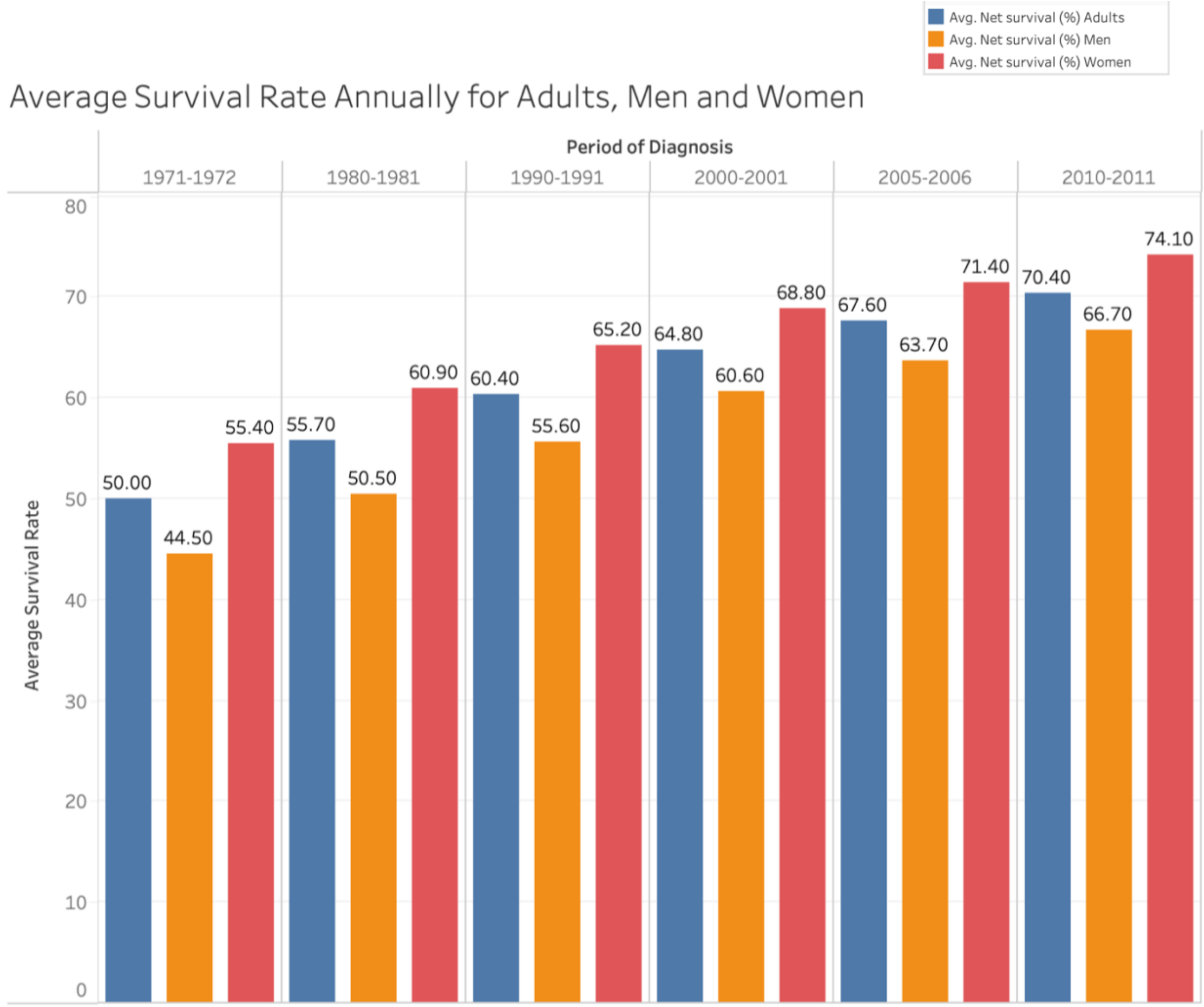


Figure 5: The average survival rates for adults, men and women annually from year 1971-2011. The dataset is sourced from the Office of National Statistics UK.

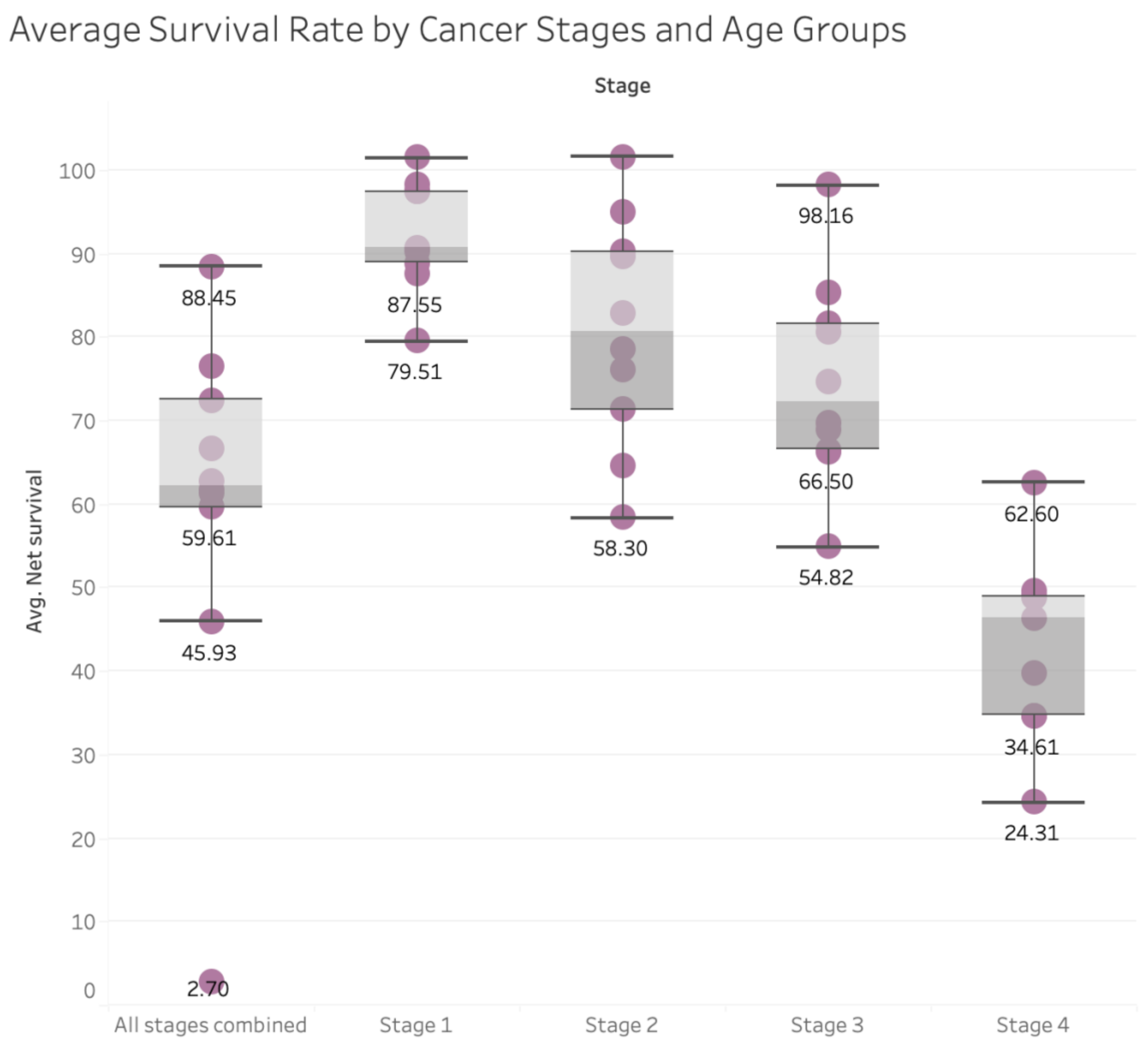
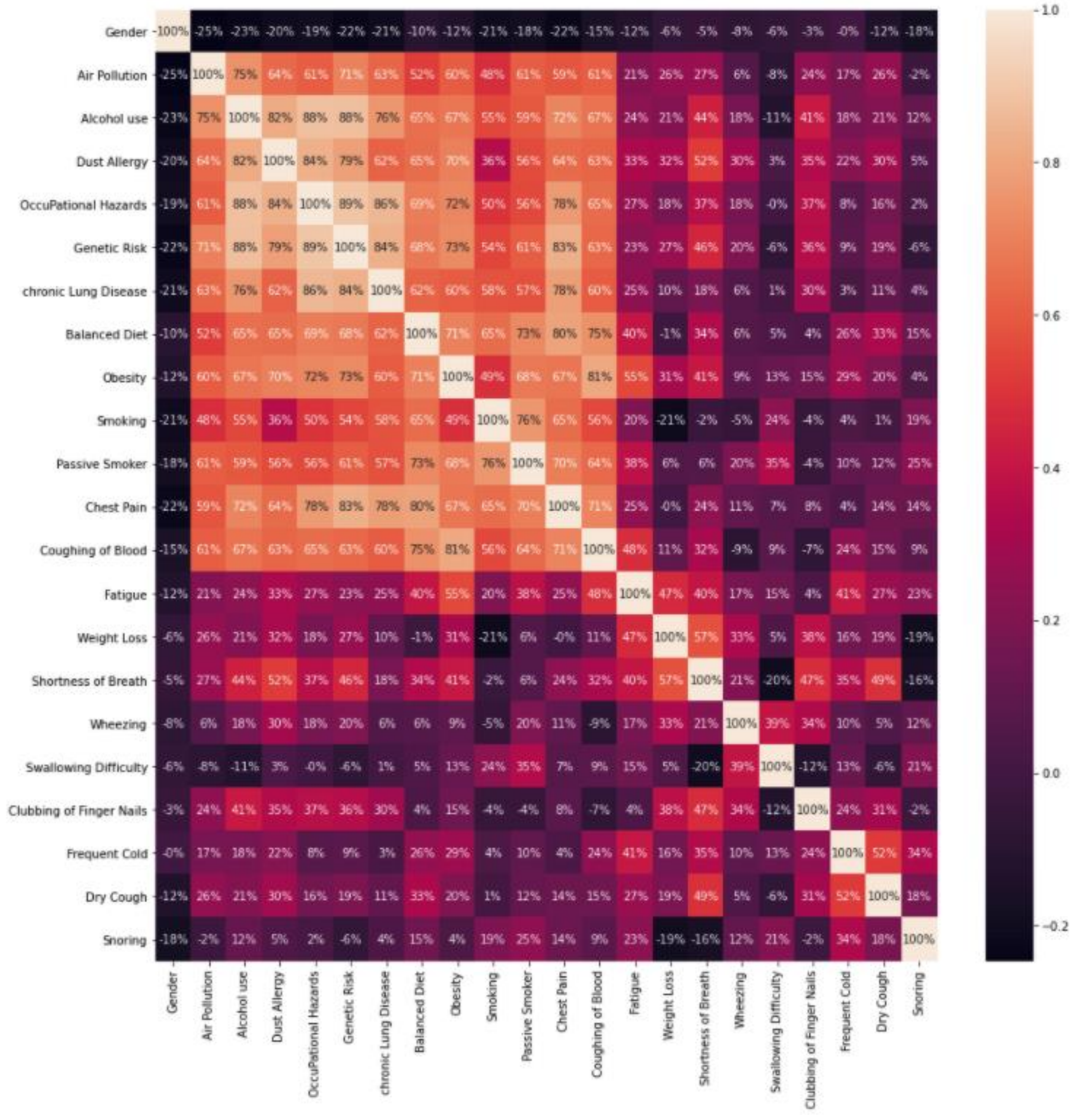


Figure 6: The average survival rates according to cancer stages and age groups.

Figure 6 demonstrates the average survival rate of cancer patients by stage and age groups. Survival rates in stage 1 are the highest at an average of 87.55%, followed by stage 2 with an average of 73%, stage 3 at 66.50% and stage 4 is lowest at 34.61%. We can produce an inference that the chance of survival is higher at early stages of cancer development and early diagnosis is vital for the success rate of treatment.

Our cancer dataset consisted of 24 features, including age, gender, air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced 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, From figure 7, we can observe through the heatmap which variables correlate with each other to produce a better understanding of our data analysis.



	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced 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
Gender	100%	-25%	-23%	-20%	-19%	-22%	-21%	-10%	-12%	-21%	-18%	-22%	-15%	-12%	-6%	-5%	-8%	-6%	-3%	-0%	-12%	-18%
Air Pollution	-25%	100%	75%	64%	61%	71%	63%	52%	60%	48%	61%	59%	61%	21%	26%	27%	6%	-8%	24%	17%	26%	-2%
Alcohol use	-23%	75%	100%	82%	88%	88%	76%	65%	67%	55%	59%	72%	67%	24%	21%	44%	18%	-11%	41%	18%	21%	12%
Dust Allergy	-20%	64%	82%	100%	84%	79%	62%	65%	70%	36%	56%	64%	63%	33%	32%	52%	30%	3%	35%	22%	30%	5%
OccuPational Hazards	-19%	61%	88%	84%	100%	89%	86%	69%	72%	50%	56%	78%	65%	27%	18%	37%	18%	-0%	37%	8%	16%	2%
Genetic Risk	-22%	71%	88%	79%	89%	100%	84%	68%	73%	54%	61%	83%	63%	23%	27%	46%	20%	-6%	36%	9%	19%	-6%
chronic Lung Disease	-21%	63%	76%	62%	86%	84%	100%	62%	60%	58%	57%	78%	60%	25%	10%	18%	6%	1%	30%	3%	11%	4%
Balanced Diet	-10%	52%	65%	65%	69%	68%	62%	100%	71%	65%	73%	80%	75%	40%	-1%	34%	6%	5%	4%	26%	33%	15%
Obesity	-12%	60%	67%	70%	72%	73%	60%	71%	100%	49%	68%	67%	81%	55%	31%	41%	9%	13%	15%	29%	20%	4%
Smoking	-21%	48%	55%	36%	50%	54%	58%	65%	49%	100%	76%	65%	56%	20%	-21%	-2%	-5%	24%	-4%	4%	1%	19%
Passive Smoker	-18%	61%	59%	56%	56%	61%	57%	73%	68%	76%	100%	70%	64%	38%	6%	6%	20%	35%	-4%	10%	12%	25%
Chest Pain	-22%	59%	72%	64%	78%	83%	78%	80%	67%	65%	70%	100%	71%	25%	-0%	24%	11%	7%	8%	4%	14%	14%
Coughing of Blood	-15%	61%	67%	63%	65%	63%	60%	75%	81%	56%	64%	71%	100%	48%	11%	32%	-9%	9%	-7%	24%	15%	9%
Fatigue	-12%	21%	24%	33%	27%	23%	25%	40%	55%	20%	38%	25%	48%	100%	47%	40%	17%	15%	4%	41%	27%	23%
Weight Loss	-6%	26%	21%	32%	18%	27%	10%	-1%	31%	-21%	6%	-0%	11%	47%	100%	57%	33%	5%	38%	16%	19%	-19%
Shortness of Breath	-5%	27%	44%	52%	37%	46%	18%	34%	41%	-2%	6%	24%	32%	40%	57%	100%	21%	-20%	47%	35%	49%	-16%
Wheezing	-8%	6%	18%	30%	18%	20%	6%	6%	9%	-5%	20%	11%	-9%	17%	33%	21%	100%	39%	34%	10%	5%	12%
Swallowing Difficulty	-6%	-8%	-11%	3%	-0%	-6%	1%	5%	13%	24%	35%	7%	9%	15%	5%	-20%	39%	100%	-12%	13%	-6%	21%
Clubbing of Finger Nails	-3%	24%	41%	35%	37%	36%	30%	4%	15%	-4%	-4%	8%	-7%	4%	38%	47%	34%	-12%	100%	24%	31%	-2%
Frequent Cold	-0%	17%	18%	22%	8%	9%	3%	26%	29%	4%	10%	4%	24%	41%	16%	35%	10%	13%	24%	100%	52%	34%
Dry Cough	-12%	26%	21%	30%	16%	19%	11%	33%	20%	1%	12%	14%	15%	27%	19%	49%	5%	-6%	31%	52%	100%	18%
Snoring	-18%	-2%	12%	5%	2%	-6%	4%	15%	4%	19%	25%	14%	9%	23%	-19%	-16%	12%	21%	-2%	34%	18%	100%

Figure 7: Heatmap comparing correlation between variables in the dataset

Gender shows negative correlation to any other variable in the figure, showing that it may be irrelevant as a feature to predict cancer using the algorithm. The feature with the highest correlations are genetic risk and occupational hazards. Features with high correlations with other also include alcohol use with occupational hazards, genetic risk, chronic lung cancer, dust allergy and chest pain at above 80%.

Figure 8 is a visualization graph to show the range and counts of genetic risk in the dataset.

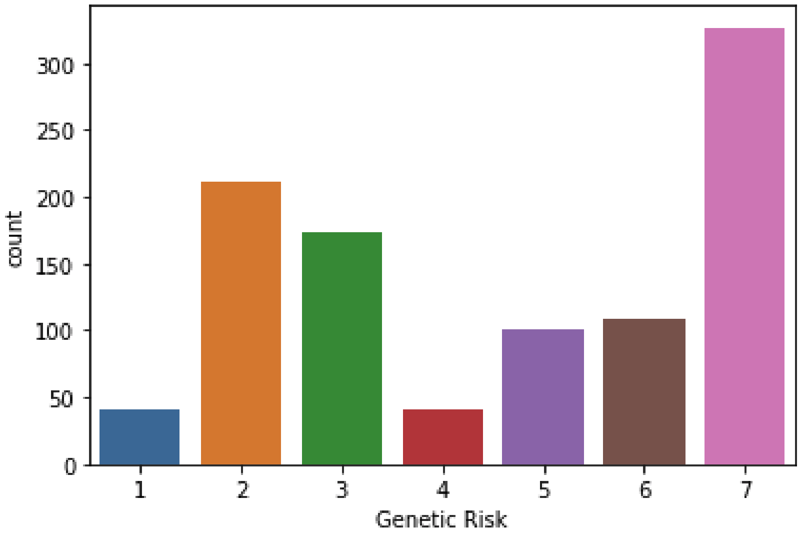


Figure 8: The count of genetic risk grading in the cancer patient dataset.

The figure shows that the highest count of genetic risk lies in the highest level, at level 7. The machine learning analysis should arguably be conducted on patients with the highest risk of cancer to facilitate early detection of cancer.

We tested 3 different models to compare prediction accuracies to improve measured accuracy. The machine learning models chosen for testing were a decision tree model, K nearest neighbours model, and logistic regression model. K nearest neighbours model determines neighbourhoods based on numeric features, which is what the dataset metrics are based upon whereas decision trees predict a class for a given input vector but the features could be numeric or nominal. The KNN model is an instance based learning model that requires constant data input to be tested against to produce an output. The decision tree model processes data in groups, storing trained observation models in memory. The model training accuracies produced were 85% for the decision tree, 99.25% for the KNN model and 84% for the logistic regression model. There could be an error made in the model training and testing, however the high accuracies show that machine learning algorithms could successfully detect cancer using patient data.

3. Discussion

From the data analysis produced, we can support the initial hypothesis that the implementation of machine learning technology in early detection stages has the potential to positively impact a significant increase in survival rates of cancer patients. The main research questions are discussed as follows:

3.1.Does the implementation of machine learning algorithms affect the efficacy of medical disease diagnosis in cancer?

Based on the analysis predictive models, the highest accuracy among the Logistic Regression mode, Decision Tree Model, K nearest neighbours algorithm and Random Forest was the KNN Model at 99.25%. This model could be used as a simple predicting tool, assumingly duplicable using other source of input variable to produce a prediction model at a high accuracy rate. Hence, machine learning algorithms can effectively detect cancer in the early stages and improve survival rates of patients.

3.2. What are the existing and new approaches in early cancer diagnosis and determination of treatment protocols for a patient?

Based on literature review, the current approaches for cancer diagnosis are based on annual checkups or showing symptoms. Imaging tests such as a computerized tomography (CT) scan, bone scan and magnetic resonance imaging (MRI) are common approaches to obtain results for tumor detection in the body. It is common for a patient to obtain results too late, after tumor growth in a part of the body has affected way of life or shown side effects. After detection, the practitioner will approach treatment protocols by chances of survival or success rate of treatment methods. The patient often has to accept the side effects they are willing to compromise on or the potential risk when deciding on treatment options.

The recent innovations in early cancer detection are implementation of machine learning algorithms that process big data to produce statistics on the likelihood of cancer based on patient medical records. This method saves time, cost and effort of practitioners or data analysts to manually process large amounts of data and prevents backlog. From the machine learning algorithm model tested in this research, the KNN model had an accuracy rate of 99.25% which is impactful in predicting traces of cancer in future applications. Other innovative treatment approaches not covered in this research analysis are nanoparticles, natural antioxidants, targeted therapy, gene therapy, extracellular vesicles (EVs), radiomics and thermal ablation magnetic hyperthermia show potential in other research papers. (Pucci, C., et al, 2019.)

3.3. What are the advantages and disadvantages of utilizing machine learning to analyze patient medical history and data?

The advantages of machine learning models are the high speed of conducting analysis compared to a human practitioner. 90% of data was collected in the last 2 years, which means more data will be exponentially created in the coming years. Businesses and healthcare organizations can only keep up with this by training a machine learning algorithm to store, process and analyze the future data to facilitate future research. Ideally, AI methods would only complement human decision-makers, find complex data patterns, prioritize experiments, find better targets, improve modeling, choose appropriate patients for clinical trials, extract insights, and even predict clinical results.

The biggest roadblock on implementing machine learning in processing human medical records are privacy concerns. In many instances, people are carelessly giving consent to data sharing without reading the fine print on disclosure of information on digital platforms. The complex growth of machine learning techniques has been exponential and beyond understanding of regulatory bodies, strongly suggesting AI monitoring is required to ensure privacy of patients are still intact to avoid personal data being sold to parties that could exploit the information for their benefit.

Chapter 6: Conclusion

1. Introduction

In conclusion, there are existing artificial intelligence and machine learning technologies that have potential to be a major contributing factor in early cancer detection, treatment protocol administration and a higher rate of recovery.

2. Research Limitations

One of the challenges faced during this research was the availability of appropriate health datasets to the public for analysis. Upon conducting this research, it is discovered that the datasets available were limited and lacked latest figure updates. Cancer research has a history of having a large occurrence of backlogs due to large uncertainties around the disease and late registrations. The datasets obtained were up to the year 2018 for registrations and 2011 for survival statistics, proving that the backlog of registrations are an existing problem to this day. Machine learning algorithms face challenges in reproducibility due to limited vital results, hindering progress within real world applications. As a solution, organizations should work together in order to transform the combination of health research outcomes into meaningful algorithms.

3. Future Research

There are many laws and regulations that need to be considered when implementing machine learning into processing healthcare data, ethical methods that need to be reviewed before, during and after development of the systems. Transparency on the use of algorithmic-assisted decision making and AI monitoring needs to be made public information. The digitalization of healthcare requires strengthening data quality, governance, security and interoperability. (Sparathou, A. et al, 2020.) It is also crucial that once machine learning algorithms are widely accepted and utilized in the healthcare system to be clarified whether it will be regulated as a product or to merely be used as a tool to support decision making and to introduce a consistent regulatory approach to be monitored. (McKinsey, 2020.)

Machine learning algorithms are able to identify patterns when testing against sample datasets. The potential applications in managing cloud software systems of the healthcare industry will be commonplace, enabling the data sharing across hospital branches or clinics possible, given there has been agreement for permission by patients and parties involved. Early cancer diagnosis and treatment will be a quick process, streamlining patient consultations and treatments by hospital staff. As a result, human lives will improve and extend our lifespans effectively.

Glossary

Artificial intelligence (AI): An area of study in the field of computer science. (Kok, J.N., et al., 2021.) The theory and development of computer systems to be able to perform tasks that normally require human intelligence, such as visual perception, decision-making, and translation between languages. (Oxford Languages, 2021.)

Machine learning: is a subset of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Deep learning: is a step further advanced than machine learning, where it is based on the structure of neural networks similar to a human brain which allows the program to identify patterns without pre-defined programming codes.

Blockchain: are tamper evident and tamper resistant digital ledger implemented in a distributed fashion (i.e., without a central repository) and usually without a central authority (i.e., a bank, company or government). At their basic level, they enable a community of users to record transactions in a shared ledger within that community, such that under normal operation of the blockchain network no transaction can be changed once published.

Feature Scaling: the process of normalizing the range of features in a dataset.

Appendices

References

1. Zhang, X.D. (2020). Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence* (pp. 223-440). Springer, Singapore.

2. IBM Cloud Education. (2020). Machine Learning. [online]. Available from: https://www.ibm.com/uk-en/cloud/learn/machine-learning [Accessed on 13 October 2021].

3. Global Cancer Observatory. (2020). Estimated number of new cases in 2020, both sexes, all ages. [online]. Available from https://gco.iarc.fr/today/home [Accessed on 1 November 2021].

4. Roser, M., Ritchie, H. (2015). Cancer. [online]. Available from: https://ourworldindata.org/cancer#citation [Accessed on 14 October 2021].

5. World Population Prospects. (2019). Global Issues: Ageing. [online]. Available from: https://www.un.org/en/global-issues/ageing [Accessed on 14 October 2021].

6. Murphy, D.R., Laxmisan, A., Reis, B.A., Thomas, E.J., Esquivel, A., Forjuoh, S.N., Parikh, R., Khan, M.M., Singh, H. (2013). Electronic health record-based triggers to detect potential delays in cancer diagnosis. [online]. Available from: https://qualitysafety.bmj.com/content/23/1/8 [Accessed on 29 October 2021].

7. The Lancet Oncology. (2021). COVID-19 and cancer: 1 year on. Vol 22 Issue 4 P4114. [online]. Available from: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(21)00148-0/fulltext [Accessed on 28 October 2021].

8. United Nations. (2019). UN, global health agencies sound alarm on drug-resistant infections; new recommendations to reduce ‘staggering number’ of future deaths. https://news.un.org/en/story/2019/04/1037471 [Accessed on 20 August 2021].

9. NHS England. (2021). Our 2019/20 Annual Report. [online]. Available from: https://www.england.nhs.uk/wp-content/uploads/2021/01/nhs-england-annual-report-2019-20-full.pdf [Accessed on 1 October 2021].

10. Johnson, D. (2021). Supervised vs Unsupervised Learning: Key Differences. [online]. Available from: https://www.guru99.com/supervised-vs-unsupervised-learning.html [Accessed on 3 December 2021].

11. Lutkevich, B. (2020). Embedded system. [online]. Available from: https://internetofthingsagenda.techtarget.com/definition/embedded-system [Accessed on 9 October 2021].

12. Orchard, R. (2019). On the cutting edge: how robots are revolutionizing surgery. [online]. Available from: https://www.slow-journalism.com/stories/on-the-cutting-edge-how-robots-are-revolutionising-surgery [Accessed on 9 October 2021].

13. Huddy, J.R., Crockett, M., Nizar, A.S., et al. Experiences of a “COVID protected” robotic surgical centre for colorectal and urological cancer in the COVID-19 pandemic. Journal of Robotic Surgery (2021). [online]. Available from: https://link.springer.com/article/10.1007/s11701-021-01199-3 [Accessed 10 October 2021].

14. Gosrisirikul, C., Chang, K.D., Raheem, A.A., Rha, K.H. (2018). New era of robotic surgical systems. [online]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/ases.12660?saml_referrer [Accessed 10 October 2021].

15. Hashizume, M., Tsugawa, K. (2004). Robotic surgery and Cancer: the Present State, Problems and Future Vision, Japanese Journal of Clinical Oncology, Volume 34, Issue 5, Pages 227-237. [online]. Available from: https://doi.org/10.1093/jjco/hyh053 [Accessed 10 October 2021].

16. McKinsey & Company. (2016). Robots and AI eating into job market for predicatable and routine work. [online]. Available from: https://www.consultancy.uk/news/12433/robots-and-ai-eating-into-job-market-for-predictable-and-routine-work [Accessed on 1 October 2021].

17. Deloitte. (2017). Automation is here to stay…but what about your workforce? [online]. Available from: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-fsi-automation-here-to-stay.pdf [Accessed on 27 October 2021].

18. Muro, M., Maxim, R., Whiton, J. Hathaway, I. (2019). Automation and Artificial Intelligence: How machines are affecting people and places. [online]. Available from: https://www.brookings.edu/wp-content/uploads/2019/01/2019.01_BrookingsMetro_Automation-AI_Report_Muro-Maxim-Whiton-FINAL-version.pdf [Accessed on 27 October 2021].

19. National Cancer Institute (2019). How Cancer Is Diagnosed. [online]. Available from: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis [Accessed on 1 October 2021].

20. Stanford Health Care (2021). What Is Cancer? [online]. Available from: https://stanfordhealthcare.org/medical-conditions/cancer/cancer/cancer-causes.html [Accessed on 1 October 2021].

21. Hantel, A. (2018). Environmental factors that cause cancer. [online]. Available from: https://www.eehealth.org/blog/2018/04/environmental-factors-that-cause-cancer/ [Accessed on 1 October 2021].

22. V Soft Consulting. (2021). The Robotic Future of Artificial Intelligence and Natural Language Processing(NLP). [online]. Available from: https://blog.vsoftconsulting.com/blog/the-robotic-future-of-artificial-intelligence-and-natural-language-processing-nlp [Accessed on 1 October 2021].

23. National Cancer Institute (2013). Understanding Laboratory Tests. [online]. Available from: https://www.cancer.gov/about-cancer/diagnosis-staging/understanding-lab-tests-fact-sheet [Accessed on 3 August 2021].

24. Ajiboye, T. (2021). How Cancer is Diagnosed. [online]. Available from: https://www.verywellhealth.com/cancer-diagnosis-4689149 [Accessed on 3 August 2021].

25. Mayo Clinic. (2020). CT Scan. [online]. Available from: https://www.mayoclinic.org/tests-procedures/ct-scan/about/pac-20393675 [Accessed on 1 October 2021].

26. NHS UK. (2018). MRI Scan Overview. [online]. Available from: https://www.nhs.uk/conditions/mri-scan/ [Accessed on 1 October 2021].

27. Benisek, A. (2020). What Are Bone Scans for Cancer? [online]. Available from: https://www.webmd.com/cancer/bone_scan_cancer [Accessed on 1 October 2021].

28. RadiologyInfo. (2021). Positron Emission Tomography – Computed Tomography (PET/CT). [online]. Available from: https://www.radiologyinfo.org/en/info/pet [Accessed on 1 October 2021].

29. Thomas, M. (2021). 16 Examples of a Healthcare Revolution Using Machine Learning. [online]. Available from: https://builtin.com/artificial-intelligence/machine-learning-healthcare [Accessed on 10 December 2021].

30. El-Baz, A., Beache, G.M., Gimel'farb, G., Suzuki, K., Okada, K., Elnakib, A., Soliman, A., Abdollahi, B. (2013). Computer-Aided Diagnosis Syst4ems for Lung Cancer: Challenges and Methodologies, *International Journal of Biomedical Imaging*, vol.2013, Article ID 942353, 46 pages.

31. Lambin, P., Leijenaar, R., Deist, T. et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14, 749-762.

32. Gillies, R.J., Kinahan, P.E., Hricak, H. (2015). Radiomics: Images Are More than Pictures, They Are Data. *Radiology 278:2*, pg 563-577.

33. Bitencourt, A., Naranjo, I.D., Gullo, R.L., Saccarelli, C.R., Pinker, K. (2021). AI-enhanced breast imaging: Where are we and where are we heading? [online]. Available from: https://www.ejradiology.com/article/S0720-048X(21)00363-6/fulltext [Accessed on 20 October 2021].

34. Onan, A. (2015). A fuzzy-rough nearest neighbour classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications, Volume 42, Issue 20* (pp6844-6852).

[online]. Available from: https://doi.org/10.1016/j.eswa.2015.05.006. [Accessed on 1 28 October 2021].

35. Menden, M.P., Iorio, F., Garnett, M., McDermott, U., Benes, C.H., Ballester, P.J., Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. [online]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061318 [Accessed on 20 October 2021].

36. Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P.M., Sundarasekar, R., Hsu, C. (2017). Machine Learning Based Big Data Processing Framework for Cancer Diagnsosis Using Hidden Markov Model and GM Clustering, *Wireless Personal Communications*, 102, 2099-2116.

37. Demircioğlu, A. (2021). Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights Imaging* **12,** 172. https://doi.org/10.1186/s13244-021-01115-1

38. Spatharou, A., Hieronimus, S., Jenkins, J. (2020). Transforming healthcare with AI: The impact on the workforce and oorganizations. [online]. Available from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/transforming-healthcare-with-ai [Accessed on 1 August 2021].

39. Moser, E.C., Narayan, G. (2020). Improving breast cancer care coordination and symptom management by using AI driven predictive toolkits, The Breast, Volume 50, pg 25-29. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S0960977619312135 [Accessed on 1 August 2021].

40. Yaga, D., Mell, P., Roby, N., Scarfone, K. (2018). Blockchain Technology Overview. https://doi.org/10.6028/NIST.IR.8202

41. IBM. (2021). Blockchain and artificial intelligence. [online]. Available from: https://www.ibm.com/topics/blockchain-ai [Accessed on 4 September 2021].

42. McGhin, T., Choo, K.K.R., Liu C.Z., He, D. (2019). Blockchain in healthcare applications: Research challenges and opportunities. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S1084804519300864 [Accessed on 25 September 2021].

43. Taulli, T. (2021). IBM Watson: Why Is Healthcare AI So Tough? [online]. Available from: https://www.forbes.com/sites/tomtaulli/2021/02/27/ibm-watson-why-is-healthcare-ai-so-tough/?sh=55a5512a5375 [Accessed on 10 December 2021].

44. Panch, T., Mattie, H., Celi, L.A. (2019). The “inconvenient truth” about AI in healthcare. [online]. Available from: https://www.nature.com/articles/s41746-019-0155-4 [Accessed on 12 December 2021].

45. Free, R., Curtis, H., Zapisetskaya, B., Kerrigan, C. (2021). AI, Machine Learning & Big Data Laws and Regulations 2021 | United Kingdom. [online]. Available from:

https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/united-kingdom [Accessed on 12 December 2021].

46. Mlitz, K. (2021). Vendor market share in cloud infrastructure services market worldwide 2017-2021. [online]. Available from: https://www.statista.com/statistics/967365/worldwide-cloud-infrastructure-services-market-share-vendor/ [Accessed on 10 December 2021].

47. McKelvey, N., Curran, K., Toland, L. (2016). The Challenges of Data Cleansing with Data Warehouses. *Effective Big Data Management and Opportunities for Implementation. Chapter 5. IGI Global.*

48. Fan, J. Han, F., Liu, H. (2014). Challenges of Big Data Analysis. National Science Review, Vol 1, Issue 2, pages 293-314. [online]. Available from: https://academic.oup.com/nsr/article/1/2/293/1397586?login=true [Accessed on 19 October 2021].

49. Waskom, M. (2012). Visualizing statistical relationships. [online]. Available from: https://seaborn.pydata.org/tutorial/relational.html [Accessed on 4 September 2021].

50. Dash, M., Liu, H. Feature selection for classification, Intelligent Data Analysis, Volume 1, Issues 1–4, 1997, Pages 131-156, ISSN 1088-467X, https://doi.org/10.1016/S1088-467X(97)00008-5.

51. Brownlee, J. (2020). How to Choose a Feature Selection Method for Machine Learning. [online]. Available from: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ [Accessed on 10 September 2021].

52. Brownlee, J. (2019). Your First Machine Learning Project in Python Step-By-Step. Python Machine Learning. [online]. Available from: https://machinelearningmastery.com/machine-learning-in-python-step-by-step/ [Accessed on 10 September 2021].

53. Cancer Research UK. (2014). Cancer survival statistics for all cancers combined, Cancer Research UK. [online]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/all-cancers-combined#heading-One [Accessed on 12 August 2021].

54. Davenport, T., Kalakota, R. (2019). The potential for artificial intelligence in healthcare, Future Healthcare Journal, pg 94-98. [online]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/ [Accessed on 12 August 2021].

55. Chakravarty, K., Antontsev, V., Bundey, Y. Varshney, J. (2021). Driving Success in personalized medicine through AI-enabled computational modeling, Drug Discovery Today, Volume 26, Issue 6, pg 1459-1465. [online] Avaiable from: https://www.sciencedirect.com/science/article/pii/S1359644621000738#fig0005 [Accessed on 12 August 2021].

56. Liu, Z., Roberts, R.A., Lal-Nag, M., Chen, X., Huang, R., Tong, W. (2021). AI-based language models powering drug discovery and development, Drug Discovery Today, Volume 26, Issue 11, pg 2593-2607. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S1359644621002816 [Accessed on 10 December 2021].

57. Schmidt-Erfurth, U., Reiter, G.S., Riedl, S., Seebock, P.. Vogl, W., Blodi, B.A., Domalpally, A., Fawzi, A., Jia, Y., Sarraf, D., Bogunovix, H. (2021). AI-based monitoring of retinal fluid in disease activity and under therapy, Progress in Retinal and Eye Research. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S1350946221000331 [Accessed on 10 December 2021].

58. He, S., Leanse, L.G., Feng, Y. (2021). Artificial intelligence and machine learning assisted drug delivery for effective treatment of infectious diseases, Advanced Drug Delivery Reviews, Volume 178. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S0169409X2100315X [Accessed on 10 December 2021].

59. Chiang, K., Huang, C., Hsieh, L., Chang, K. (2020). A propositional AI system for supporting epilepsy diagnosis based on the 2017 epilepsy classification: Illustrated by Dravet syndrome, Epilepsy & Behavior, Volume 106. [online]. Available from: https://www.sciencedirect.com/science/article/pii/S1525505020302006 [Accessed on 10 December 2021].

60. Office for National Statistics. (2019). Dataset – Cancer survival in England – adults diagnosed. [online]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed [Accessed on 20 September 2021].

61. Pucci, C., Martinelli, C., Ciofani, G. (2019). Innovative approaches for cancer treatment: current perspectives and new challenges, Ecancermedicalscience. [online]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6753017/ [Accessed on 20 December 2021].

62. Kok, J.N., Boers, E.J.W., Kosters, W.A., Putten, P.v.d. (2021). Artificial Intelligence: Definition, Trends, Techniques, and Cases. [online]. Available from: https://www.eolss.net/Sample-Chapters/C15/E6-44.pdf [Accessed on 1 September 2021].

63. Leite, M.L., Costa, L.S.L., Cuhna, V.A., Kreniski, V., Filho, M.O.B., Cunha, N.B., Costa, F.F. (2021). Artificial intelligence and the future of life sciences, Drug Discovery Today.

64. Chong, J. (2020). What is Feature Scaling & Why is it Important in Machine Learning? MinMaxScaler vs StandardScaler vs RobustScaler. [online]. Available from: https://towardsdatascience.com/what-is-feature-scaling-why-is-it-important-in-machine-learning-2854ae877048 [Accessed on 10 December 2021].

65. Oxford Languages. (2021).

66. Buczak, A.L., Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrussion Detection, IEEE Communications Surveys & Tutorials, vol. 18, no.2, pp. 1153-1176.

67. Bhatt, A. (2021). Artificial intelligence in managing clinical trial design and conduct: Man and machine still on the learning curve?, Perspect Clin Res. [online]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8011519/ [Accessed on 19 November 2021].