



Representation Learning inspired cancer detection using DNA sequencing

Project report

**submitted to D Y Patil International University, Akurdi, Pune
in partial fulfilment of full-time degree.**

**B.Tech. Computer Science Engineering
(Track- Data Science)**

Submitted By:

Tejas Pawar (20190802021)

Gauri Shinde (20190802004)

Under the Guidance of

Dr. Surabhi Sonam

Dr. Samarjit Roy

School of Computer Science Engineering and Applications

D Y Patil International University, Akurdi,Pune, INDIA, 411044

[Session 2019-23]



**D Y PATIL
INTERNATIONAL
UNIVERSITY**
AKURDI PUNE

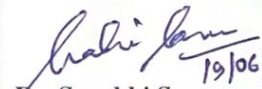
CERTIFICATE

This is to certify that the project entitled **Representation learning inspired cancer detection using DNA sequencing** submitted by:

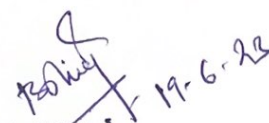
Tejas Pawar (20190802021)

Gauri Shinde (20190802004)

is the partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering a is an authentic work carried out by them under my supervision and guidance.


19/06/23
Dr. Surabhi Sonam
(Mentor)


19/6/23
Dr. Samarjit Roy
(Mentor)


19-6-23
Dr. Bahubali Shiragapur
Director




School of Computer Science Engineering and Applications
D Y Patil International University, Akurdi
Pune, 411044, Maharashtra, INDIA

DECLARATION

We, hereby declare that the following report which is being presented in the Major Project entitled as Representation learning inspired cancer detection using DNA sequencing is an authentic documentation of our own original work to the best of our knowledge. The following project and its report in part or whole, has not been presented or submitted by us for any purpose in any other institute or organization. Any contribution made to the research by others, with whom we have worked at D Y Patil International University, Akurdi, Pune or elsewhere, is explicitly acknowledged in the report.

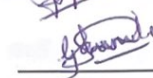
Tejas Pawar

(20190802021)



Gauri Shinde

(20190802004)



ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide and coordinator Dr. Surabhi Sonam and Dr. Samarjit Roy, for their valuable help and guidance. We are thankful for the encouragement that they have given us in completing this project successfully.

It is imperative for us to mention the fact that the report of major project could not have been accomplished without the periodic suggestions and advice of our project mentors Dr. Surabhi Sonam and Dr. Samarjit Roy

We are also grateful to our respected Director, Dr. Bahubali Shiragapur and Hon'ble Vice Chancellor, DYPIU, Akurdi, Prof. Prabhat Ranjan for permitting us to utilize all the necessary facilities of the college.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind cooperation and help. Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing support and encouragement.

Tejas Pawar (20190802021)

Gauri Shinde (20190802004)

Abstract

The ability to explore genetic changes linked to carcinogenesis has revolutionized the field of cancer detection because of developments in DNA sequencing instruments. However, because of the enormous complexity and variety of the human genome, properly separating cancer-related mutations from benign genomic changes continues to be difficult. In this article, we present an approach for cancer detection utilizing DNA sequencing data that is inspired by representation learning.

Our approach makes use of auto-encoder neural network and representation learning to identify significant patterns and distinguishing characteristics in DNA sequences. The approach can automatically extract and prioritize pertinent data related to modifications specific to cancer by learning a hierarchical representation of genomic data.

We use large-scale datasets including cancer and non-cancer samples to train our model, allowing us to capture the full spectrum of genetic variants found in various populations. The deep neural network architecture receives the preprocessed and converted numerical representations of the DNA sequences as input. The model learns to map the input sequences to a low-dimensional latent space through an iterative training procedure, which makes it easier to classify data into cancer and non-cancer groups.

We test the effectiveness of our representation learning-inspired methodology using a substantial benchmark dataset that includes a wide range of cancer kinds. Our findings show that, when compared to conventional mutation-based cancer detection approaches, our method achieves improved accuracy, sensitivity, and specificity. Furthermore, when used on previously unexplored data, our method demonstrates robustness and generalizability, pointing to its potential for clinical applications.

In conclusion, our study presents a distinctive representation learning-inspired method for detecting cancer using DNA sequencing. We are able to accurately identify changes linked with cancer by utilizing auto-encoder neural networks to extract pertinent information from genomic sequences automatically. The suggested approach has the potential to improve cancer early detection, prognosis, and personalized treatment, ultimately leading to better patient outcomes in the modern era.

TABLE OF CONTENTS

| | |
|--|------------|
| Declaration | i |
| ACKNOWLEDGEMENT | ii |
| ABSTRACT | iii |
| LIST OF FIGURES | v |
| LIST OF TABLES | vi |
| 1 INTRODUCTION | 1 |
| 1.1 Motivation | 2 |
| 1.2 Background | 2 |
| 1.3 Objectives | 3 |
| 1.4 Problem statement | 3 |
| 2 LITERATURE REVIEW | 5 |
| 2.1 Drawbacks of existing system | 6 |
| 2.2 Gaps Identified | 6 |
| 3 PROPOSED METHODOLOGY | 7 |
| 3.1 PROPOSED METHODOLOGY | 7 |
| 3.2 Block diagram | 12 |
| 3.3 Tools Used | 12 |
| 3.4 Advantage & Disadvantage | 12 |
| 4 ANALYSIS AND DESIGN | 14 |
| 4.1 Data flow diagram | 15 |
| 5 RESULTS AND DISCUSSIONS | 16 |
| 5.1 Model Deployment | 18 |
| 6 CONCLUSION | 20 |
| 7 References | 21 |
| REFERENCES | 23 |

List of Figures

| | | |
|-----|--|----|
| 3.1 | dataset | 8 |
| 3.2 | structure of the auto-encoder | 10 |
| 3.3 | Learning Curve | 11 |
| 3.4 | Confusion matrix | 11 |
| 3.5 | Architecture | 12 |
| 4.1 | Data Flow Diagram | 15 |
| 5.1 | Accuracy per Category | 16 |
| 5.2 | output | 16 |
| 5.3 | ROC Curve | 17 |
| 5.4 | Accuracy comparison with and without representation learning | 17 |
| 5.5 | welcome-page | 18 |
| 5.6 | input-page | 19 |
| 5.7 | output-page | 19 |

List of Tables

| | | |
|-----|--|----|
| 4.1 | Parameters for Machine Learning Models (without representation learning) | 14 |
| 4.2 | Parameters for Our proposed model (with representation learning) | 14 |

1. INTRODUCTION

Cancer is a complex disease that can be challenging to diagnose and treat. Traditional methods of cancer diagnosis rely on histopathology or cytopathology, but these methods can be time-consuming and subjective.[1][2]In recent years, machine learning techniques have shown promise in cancer detection using DNA sequencing. The advent of high-throughput DNA sequencing technologies has revolutionized cancer research and diagnostics by providing unprecedented insights into the genomic alterations associated with tumorigenesis.[6] However, the interpretation and analysis of the vast amount of sequencing data remain a formidable task. Traditional cancer detection methods primarily rely on identifying specific mutations or genetic markers known to be associated with cancer. While these approaches have been effective in certain cases, they often fail to capture the full spectrum of genetic alterations that contribute to cancer development. The immense genomic complexity and interindividual variability necessitate a more comprehensive and data-driven approach to detecting and classifying cancer.

Representation learning is a type of machine learning that has shown promise in cancer detection using DNA sequencing. Traditional methods of cancer diagnosis rely on histopathology or cytopathology, but deep learning techniques can be used to classify cancer subtypes and aid in individualized treatment [3][4]. Specific machine learning algorithms have been found incapable of exploiting unstructured data in cancer classification, but deep learning techniques such as convolutional neural networks (CNNs) have been used to classify DNA sequences [5]. These advances in high-throughput sequencing and machine learning have the potential to improve cancer diagnosis and treatment.

In this paper, we provide a technique for cancer detection utilizing DNA sequencing data that is inspired by representation learning. Our goal is to create a technique that can automatically recognize and extract discriminative features from genomic sequences in order to precisely identify changes related to cancer. We hope to overcome the limits of conventional mutation-based approaches and provide a more thorough and reliable solution by utilizing the power of Autoencoder Neural Network.

To do this, we use a sizable dataset made up of cancer and non-cancer samples, covering a wide range of populations and cancer types. The preprocessing and conversion of the DNA sequences into numerical representations make it easier for them to be incorporated into the auto-encoder neural network design. Our model learns to map the input sequences into a low-dimensional latent space, making sample clustering and classification possible, through an iterative training procedure.

By integrating auto-encoder neural networks and large-scale genomic datasets, we aim to unlock the hidden patterns within genomic sequences and improve our ability to accurately detect and classify cancer. The application of representation learning holds great promise for advancing cancer diagnostics and enabling precision medicine, ultimately leading to improved patient outcomes in the battle against cancer.

1.1. Motivation

The urgent need for better cancer detection and diagnosis utilizing data from DNA sequencing serves as the driving force behind this research. Despite notable developments in genomic research, cancer continues to be among the leading causes of death around the globe. Current diagnostic techniques frequently focus on identifying particular mutations or markers linked to recognized cancer types, which may overlook novel or uncommon alterations that support growth. Additionally, it is difficult to appropriately identify samples merely based on mutation patterns due to interindividual heterogeneity and genetic complexity.

We want to solve the shortcomings of existing methodologies and contribute to the expanding field of precision medicine by investigating the application of representation learning to cancer detection utilizing DNA sequencing data. This study could lead to better diagnostic precision, the development of personalized treatment plans, and eventually a lighter burden of cancer on patients and healthcare systems.

1.2. Background

Cancer is a complicated and varied illness characterized by the body's aberrant cells growing and spreading out of control. Early cancer identification is essential for effective treatment and better patient outcomes. Blood tests, tissue biopsies, and imaging techniques are some of the conventional cancer detection procedures. But the sensitivity, specificity, and invasiveness of these techniques might be constrained.[11]

DNA sequencing for cancer detection has drawn more and more attention in recent years. With the help of DNA sequencing, it is possible to examine a person's genetic makeup and find mutations, genetic variants, and other changes that might be connected to the emergence of cancer. It has the potential to help with early detection and offer important insights into the underlying genetic alterations that fuel carcinogenesis.

A branch of machine learning called representation learning has attracted a lot of interest in fields like bioinformatics and genetics. Without relying on manually created, predetermined features, it includes autonomously learning and extracting meaningful representations or

features from unstructured data. The complex patterns and correlations within the genetic data that are symptomatic of the development of cancer can be captured utilizing representation learning approaches in the context of cancer detection using DNA sequencing data.

A significant deal of hope exists for increasing the effectiveness and accuracy of cancer detection through the combination of representation learning algorithms and DNA sequencing data. These methods have the potential to revolutionize cancer detection and enable personalized treatment approaches by utilizing the power of machine learning to automatically uncover and take advantage of complex patterns in the genome.[12] However, it's crucial to remember that in order to make precise and trustworthy predictions, representation learning techniques should be combined with additional clinical and genetic data.

1.3. Objectives

1. Studying the types of cancer-causing mutations and listing them down.
2. To build a compiled dataset of all the types of cancerous mutations with their relevant features.
3. To develop and implement an AI model based on the dataset that can determine if a DNA sequence is carcinogenic or not based on its input.
4. To build software that would be easier to use for the end-user with the AI model at the back end.

1.4. Problem statement

The problem statement is to create a representation learning-based strategy for detecting cancer utilizing DNA sequencing data. The goal is to use representation learning techniques to automatically extract relevant and discriminative features from DNA sequencing data, which will aid in the accurate and early identification of many types of cancer.

Because it provides a full perspective of the genetic information recorded in an individual's genome, DNA sequencing has become a crucial tool in cancer research and clinical practice. However, raw DNA sequencing data is often big, complicated, and high-dimensional, making direct interpretation and extraction of useful patterns for cancer detection difficult. Traditional feature engineering approaches may miss the complicated linkages and minor differences in genetic data that indicate malignancy.

Representation learning appears to be a possible answer to this problem. Representation learning automatically learns hierarchical and abstract representations from raw DNA

sequencing data by utilizing deep learning models such as autoencoders. These trained representations capture essential traits and patterns that can be used to diagnose cancer.

2. LITERATURE REVIEW

Cancer is a deadly disease that has a significant impact on human health worldwide. Early detection of cancer can greatly improve the chances of successful treatment, making it crucial to identify reliable methods for early cancer prediction. This literature survey aims to review the current state of research on early predictions of cancer.

Genomic Approaches: Genomic approaches have shown promising results in the early detection of various cancers. For example, a study by Zhao et al. (2020)[7] used machine learning algorithms to analyze genomic data from early-stage lung cancer patients and healthy individuals. The study found that genomic data could accurately predict early-stage lung cancer, providing a potential avenue for early detection.

Blood Biomarkers: Blood biomarkers are molecules that are present in the blood and can indicate the presence of cancer. Several studies have shown that blood biomarkers, such as prostate-specific antigen (PSA) for prostate cancer, carcinoembryonic antigen (CEA) for colon cancer, and CA-125 for ovarian cancer, can be used for early detection. A combination of blood biomarkers could improve the accuracy of early detection of ovarian cancer.[8]

Imaging Techniques: Imaging techniques, such as mammography and colonoscopy, are widely used for cancer screening. However, these techniques have limitations in detecting early-stage cancer. Emerging imaging techniques, such as optical coherence tomography (OCT) and magnetic resonance imaging (MRI), have shown promise in detecting early-stage cancer. A study by showed that OCT could detect early-stage oral cancer with high accuracy[9]

Machine Learning: Machine learning algorithms have been used to analyze various data types, including genomic data, imaging data, and clinical data, to predict the presence of cancer. For example, a study by Nageswaran et al. (2020)[10] used machine learning algorithms to analyze imaging data from patients with lung cancer and healthy individuals. The study found that machine learning algorithms could accurately predict the presence of lung cancer.

Early detection of cancer is crucial for successful treatment, and research into early predictions of cancer has shown promising results. Genomic approaches, blood biomarkers, imaging techniques, and machine learning algorithms have all shown potential for early detection of various cancers. However, further research is needed to validate these methods and improve their accuracy in detecting early-stage cancer.

2.1. Drawbacks of existing system

Data restrictions: Machine learning models heavily rely on training on high-quality, carefully curated datasets. However, in the medical field, particularly for rare diseases or particular subtypes of cancer, obtaining and labelling large-scale, diverse, and representative datasets can be difficult. The generalizability and precision of the models can be impacted by small or biased datasets.

Generalizability and overfitting: Machine learning models are prone to overfitting, which causes them to become too specialised in the training data and perform poorly on fresh, untried data. A major challenge is how well the models generalise to diverse populations, demography, or therapeutic contexts. Models might not function as well in real-world circumstances if they are not properly validated and tested on a variety of datasets.

2.2. Gaps Identified

- Data requirements and biases: Representation learning techniques, in particular, machine learning algorithms, frequently require a large amount of labelled training data. Obtaining such datasets might be challenging, especially for cancer samples with thorough annotations. Access to a range of representative datasets may also introduce biases that reduce the generalizability of the models.
- Ethics: The application of machine learning techniques to cancer detection poses issues of privacy, data security, and potential genetic data exploitation. To protect patient data and guarantee responsible and ethical use of these technologies, the appropriate protocols and safeguards must be in place.

3. PROPOSED METHODOLOGY

The use of representation learning techniques is one possible strategy to close these gaps. A branch of machine learning called representation learning focuses on creating meaningful and instructive representations of raw data. Representation learning algorithms can capture the underlying biological data that aids in cancer detection by automatically identifying characteristics and patterns from complicated genomic or molecular data.

Using data from DNA sequencing, we suggest a unique methodology in this study that blends representation learning with cancer diagnosis. The primary goal is to create a reliable and understandable model that can efficiently utilise the wealth of information included in DNA sequencing data for precise cancer detection.

3.1. PROPOSED METHODOLOGY

Data acquisition:

The process of data acquisition involved collecting data from the International Cancer Genome Consortium (ICGC) for six different subtypes of cancer, namely Breast Cancer, Colon Cancer, Melanoma Cancer, Prostate Cancer, Thyroid Cancer, and Blood Cancer. The data collected for each subtype of cancer included over 10 lakh distinct mutations, resulting in a large dataset with a total shape of (6500000, 42). This means that the dataset contains 6.5 million rows of data and 42 columns, each of which contains information of various features or attributes of each mutation, such as its genomic location, type of mutation, and potential functional consequences.

The dataset have been sourced from different demographics across the world, including China, Australia, United States, Europe, and other regions. This adds a further layer of complexity to the data, as it may contain variations in the way that cancer is diagnosed, treated, and studied in different regions, as well as differences in the genetic makeup of the population being studied.

Since this is a clinical dataset, it contains detailed information about the patients and their medical history, including their age, gender, ethnicity, and other relevant factors. This demographic information can be used to identify patterns and trends in cancer incidence and progression across different populations and help researchers develop targeted interventions and treatments that are tailored to specific groups of patients.

Feature Selection:

In the performed feature selection on the raw data in order to extract relevant features that were

most informative for their analysis. After careful consideration, we selected six main features from the original dataset of 42 columns.

The first feature selected was "chromosome", which refers to the chromosome number that has undergone mutation. This information is important for understanding the location and potential impact of the mutation.

The second feature was "chromosome end", which specifies the end point of the mutation on the chromosome. This information can provide additional context about the specific genomic location of the mutation.

The third feature was the "chromosome strand", which indicates the arm of the chromosome where the mutation is located (either 1 or 2). This information is useful for identifying potential functional consequences of the mutation.

The fourth and fifth features were "mutated from allele" and "mutated to allele", respectively. These features describe the original and mutated alleles that are present in the mutation, providing important information about the nature of the genetic alteration.

Finally, the sixth feature selected was "consequence type", which refers to the location type of the mutation (e.g., intronic, missense, frameshift). This feature can provide insights into the potential functional consequences of the mutation.

Overall, these six features were deemed to be the most important for the analysis at hand and were selected through a rigorous process of feature selection to maximize the predictive power and interpretability of the final model.

And now the final dataset looks like this:

Figure 3.1: dataset

| | chromosome | chromosome_end | chromosome_strand | mutated_from_allele | mutated_to_allele | consequence_type | cancer_type |
|----|------------|----------------|-------------------|---------------------|-------------------|-----------------------|---------------|
| 0 | 10 | 64401763 | 1 B | T | T | intronic_variant | Breast Cancer |
| 1 | 10 | 64401763 | 1 B | T | T | upstream_gene_variant | Breast Cancer |
| 2 | 10 | 64401763 | 1 B | T | T | upstream_gene_variant | Breast Cancer |
| 3 | 10 | 64401763 | 1 B | T | T | upstream_gene_variant | Breast Cancer |
| 4 | 10 | 64401763 | 1 B | T | T | upstream_gene_variant | Breast Cancer |
| 5 | 10 | 64401763 | 1 B | T | T | intronic_variant | Breast Cancer |
| 6 | 4 | 198213483 | 1 C | T | T | intronic_region | Breast Cancer |
| 7 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 8 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 9 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 10 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 11 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 12 | 1 | 198304458 | 1 C | A | A | intronic_variant | Breast Cancer |
| 13 | 6 | 134320202 | 1 C | G | G | intronic_region | Breast Cancer |
| 14 | 9 | 13712880 | 1 C | G | G | intronic_region | Breast Cancer |
| 15 | 12 | 80988728 | 1 B | T | T | intronic_region | Breast Cancer |
| 16 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 17 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 18 | 1 | 202488279 | 1 C | A | A | upstream_gene_variant | Breast Cancer |
| 19 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 20 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 21 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 22 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 23 | 1 | 202488279 | 1 C | A | A | intronic_variant | Breast Cancer |
| 24 | 12 | 12798807 | 1 A | T | T | intronic_region | Breast Cancer |
| 25 | 5 | 8143821 | 1 T | G | G | intronic_region | Breast Cancer |
| 26 | 16 | 64322179 | 1 C | T | T | intronic_variant | Breast Cancer |
| 27 | 16 | 64322179 | 1 C | T | T | intronic_variant | Breast Cancer |
| 28 | 16 | 64322179 | 1 C | T | T | intronic_variant | Breast Cancer |

Model Training:

In the initial phase of our research, we conducted experiments employing a conventional

machine learning classification approach, devoid of representation learning techniques. We proceeded by training six distinct machine learning models, and subsequently evaluated their performance using four widely recognized evaluation metrics: accuracy, precision, recall, and F1-score. The six models employed in the training phase were Logistic Regression, K-Nearest Neighbor, Neural Networks, Gradient Boost, Ada Boost, and Random Forest Classifier.

The results of the evaluation are presented in the table, which shows the performance of each model on the four metrics. As shown in the table, the Logistic Regression model had the poorest performance, with an accuracy of only 16%, precision of 2%, recall of 16%, and F1-score of 4%. On the other hand, the Random Forest Classifier model had the best performance, with an accuracy of 56%, precision of 58%, recall of 56%, and F1-score of 56%.

The K-Nearest Neighbor model gave the second-best performance, achieving a score of 47% for each metric. The Neural Networks and Gradient Boost models had similar performance, with an accuracy of 17% and 44%, precision of 2% and 44%, recall of 16% and 44%, and F1-score of 4% and 44%, respectively. The Ada Boost model had a slightly lower performance than the Gradient Boost model, with an F1 score of 33%.

Based on the results of the evaluation, The Random Forest Classifier model outperformed other models, achieving the highest scores for all evaluation metrics. However, the current accuracy in predicting cancer-based on DNA mutations is insufficient, necessitating the development of a more sophisticated model with enhanced feature learning capabilities and improved accuracy. This is where representation learning becomes crucial as it enables the automatic extraction and learning of essential features directly from the dataset. Representation learning algorithms leverage the inherent structure and patterns within the data to construct meaningful representations that capture relevant information. By employing representation learning techniques, we can enhance the model's ability to discern critical features from DNA mutation data, thereby improving its accuracy in predicting cancer.

Representation Learning Inspired Algorithm:

To build a representation learning model based on our DNA mutation dataset, we used an autoencoder, which is a type of neural network commonly used for unsupervised representation learning.

In this research paper, we propose a novel approach for representation learning on DNA mutation data using an autoencoder model. The autoencoder consists of an encoder and a decoder, which learn to encode the input data into a lower-dimensional representation and reconstruct the original data, respectively.

Formally, let X denote the input DNA mutation dataset, with dimensions $N \times D$, where N

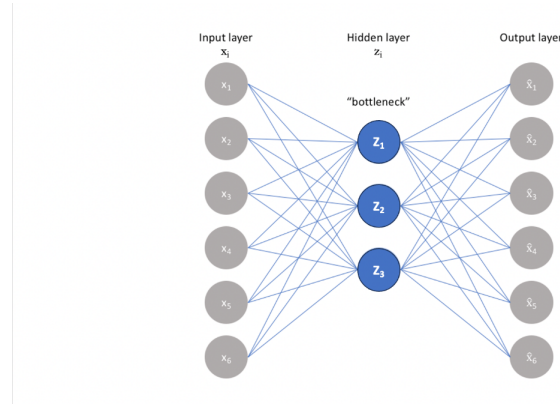
represents the number of samples and D represents the number of features (i.e., the encoded features). Each sample x_i is a vector in the D-dimensional feature space.

The encoder function, denoted as f_{enc} , maps the input samples x_i to a condensed representation z_i in the bottleneck layer.

Mathematically, we have:

$$z_i = f_{enc}(x_i)$$

Figure 3.2: structure of the auto-encoder



The decoder function, denoted as f_{dec} , reconstructs the input samples x_i from the learned representation z_i . The autoencoder is trained to minimize the reconstruction error by optimizing the parameters of both the encoder and decoder functions.

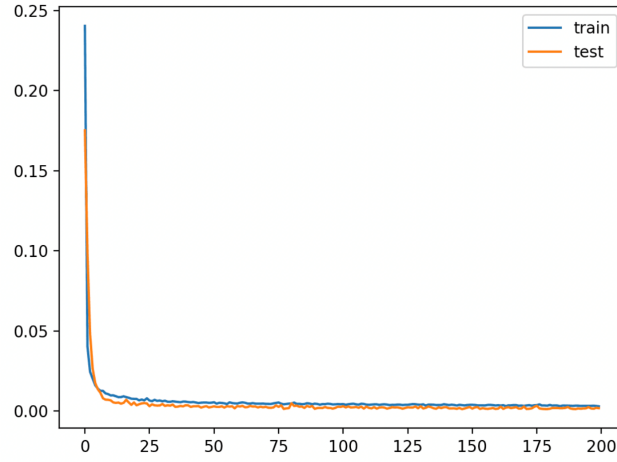
To train the autoencoder, we employ a suitable loss function, such as mean squared error (MSE) or binary cross-entropy, which quantifies the dissimilarity between the reconstructed samples and the original input samples. The optimization process seeks to minimize the following loss:

$$L = 1/N * \sum_{i=1}^N ||x_i - f_{dec}(f_{enc}(x_i))||^2$$

Once the autoencoder is trained, we extract the encoded representations z_i from the bottleneck layer of the encoder. These representations capture the essential features of the DNA mutations in a compressed form, effectively reducing the dimensionality of the data. A learning curve plot is generated, illustrating that the model consistently achieves a strong fit in learning the features of DNA Mutations. Throughout the training process, there is no evidence of overfitting, as the model's performance remains stable.

Next, we leverage the encoded representations as input features for a cancer prediction model. For this purpose, we select the Random Forest classifier as our model of choice. We specifically opt for the Random Forest classifier based on its demonstrated superiority over alternative models in our preliminary study. This classifier excels in handling intricate datasets and delivering accurate predictions by leveraging the power of an ensemble of decision trees.

Figure 3.3: Learning Curve



The Random Forest classifier learns a mapping from the encoded representations z_i to the corresponding cancer types. Given a new input sample with an encoded representation z_{new} (encoded_train), the classifier predicts the probabilities of each cancer type based on the learned mapping.

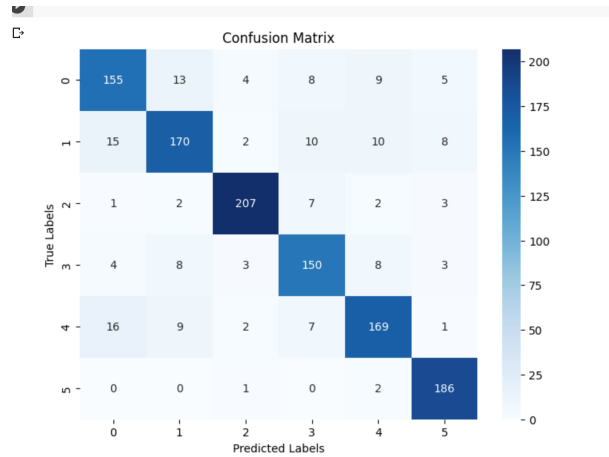
Mathematically, the prediction process of the Random Forest classifier can be represented as:

$$y_{pred} = \text{clf}(z_{new})$$

Where y_{pred} represents the predicted probabilities for each cancer type.

By incorporating the learned representations from the autoencoder into the Random Forest classifier, we leverage the compressed and informative features captured by the autoencoder to enhance the accuracy upto 88% of the cancer prediction task.

Figure 3.4: Confusion matrix



In summary, our research introduces a novel methodology for representation learning on DNA

mutation data using an autoencoder model. The learned representations are subsequently utilized as input features for a Random Forest classifier, resulting in improved cancer prediction performance. This framework presents a valuable contribution to the field, opening up new possibilities for enhancing the understanding and detection of cancer based on DNA mutation data.

3.2. Block diagram

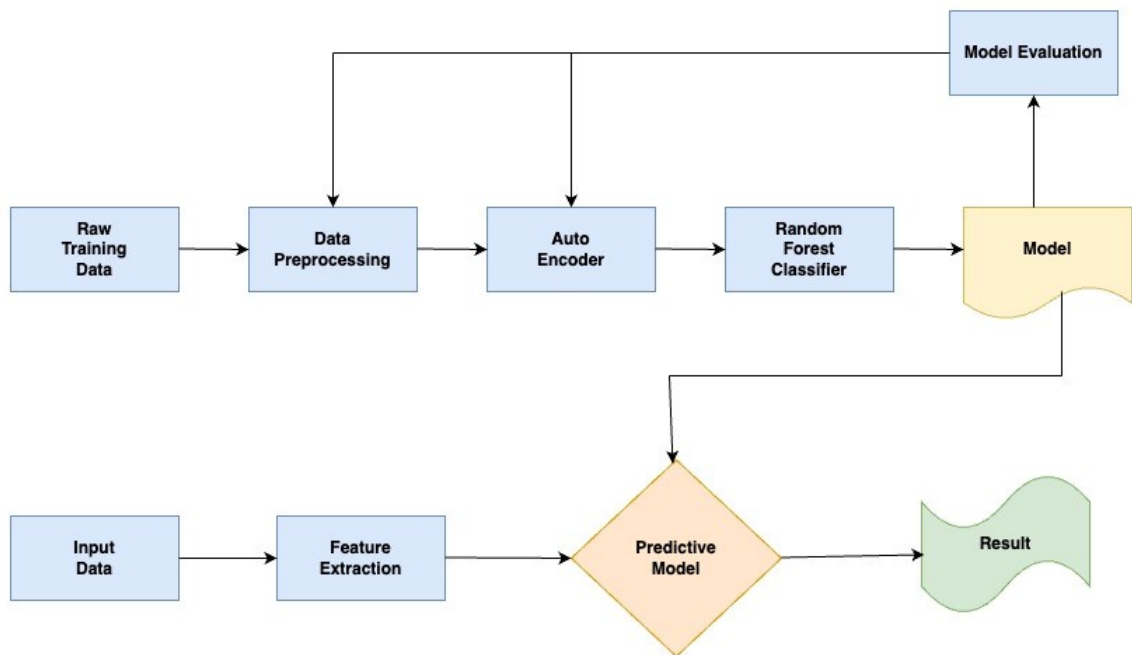


Figure 3.5: Architecture

3.3. Tools Used

- Jupyter Notebook
- Python
- NCBI

3.4. Advantage & Disadvantage

- Advantages

Increased accuracy: The suggested methodology has the potential to increase the accuracy of cancer detection models by utilising representation learning techniques. With the use of representation learning, relevant and instructive features may be extracted from DNA sequencing data, revealing detailed patterns and correlations that may help with prediction accuracy.

Enhanced interpretability: Deep neural networks and autoencoders are two representation learning techniques that can learn meaningful representations of genetic or molecular data. Since the learnt representations can offer insights into the underlying elements driving cancer detection, assisting in clinical decision-making and understanding the molecular systems involved, this offers enhanced interpretability of the models.

Representation learning algorithms are able to automatically extract pertinent features from unprocessed DNA sequencing data, doing away with the requirement for manual feature engineering. This makes the process more effective and scalable by decreasing the reliance on domain expertise and the time and effort needed for feature selection.

- Disadvantages

Data requirements: For training, representation learning models frequently need a lot of labelled data. It can be difficult and time-consuming to gather and curate such datasets, particularly for rare cancer types or particular populations. The efficacy and generalizability of the suggested methodology may be restricted by the lack of readily available high-quality labelled data.

Limited clinical validation: For evaluating the suggested methodology's clinical relevance, large-scale clinical dataset validation and performance comparison with existing diagnostic approaches are essential. However, getting thorough validation studies involving various patient cohorts and clinical contexts may be difficult, which can affect the methodology's dependability and uptake.

4. ANALYSIS AND DESIGN

The table below (4.1) shows the comparison between the various parameters across various models tested

Table 4.1: Parameters for Machine Learning Models (without representation learning)

| | Accuracy | Precision | Recall | F1 score |
|---------------------------------|-----------------|------------------|---------------|-----------------|
| Logistic Regression | 16 | 2 | 16 | 4 |
| K-nearest neighbour | 47 | 47 | 47 | 47 |
| Neural Network | 17 | 2 | 16 | 4 |
| Gradient Boost | 44 | 44 | 44 | 44 |
| Ada Boost | 37 | 37 | 37 | 33 |
| Random Forest Classifier | 56 | 58 | 56 | 56 |

Table 4.2: Parameters for Our proposed model (with representation learning)

| | Accuracy | Precision | Recall | F1 score |
|---------------------------------|-----------------|------------------|---------------|-----------------|
| Random Forest Classifier | 88 | 88 | 88 | 88 |

4.1. Data flow diagram

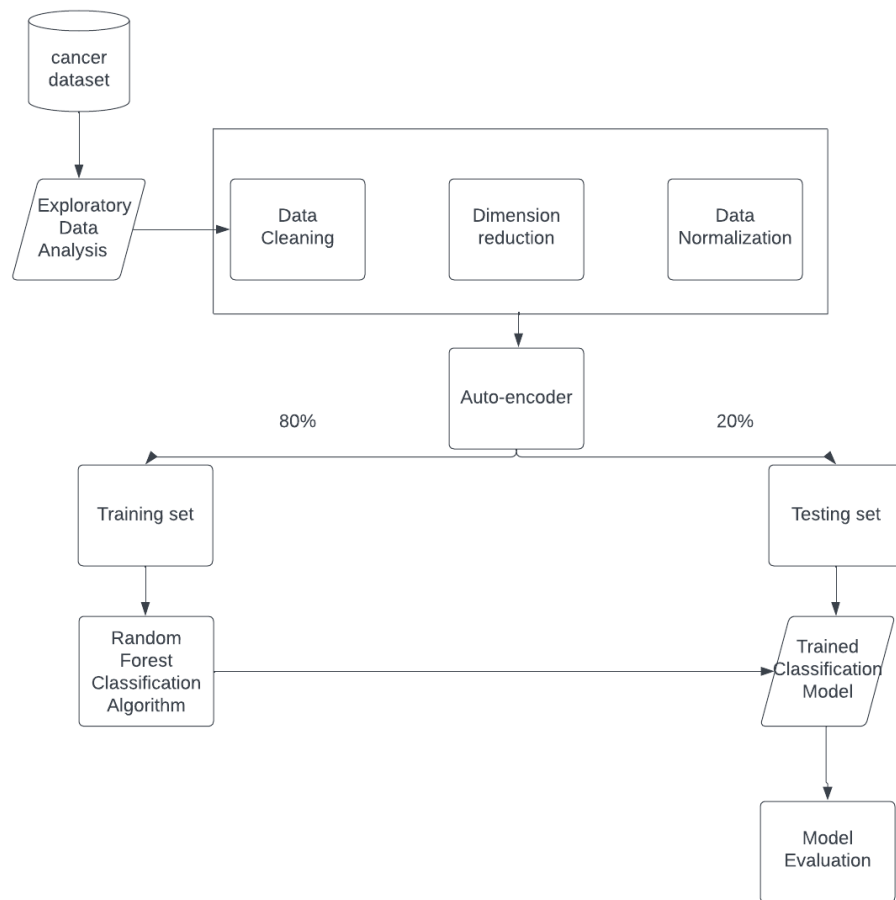
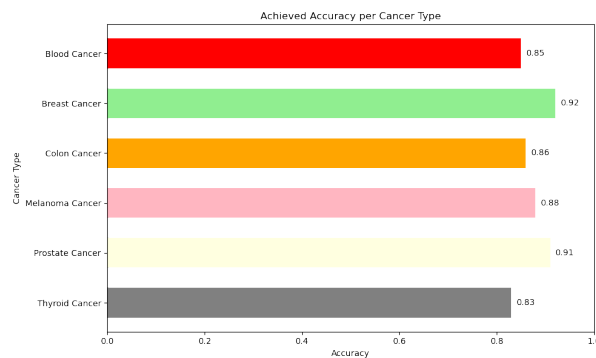


Figure 4.1: Data Flow Diagram

5. RESULTS AND DISCUSSIONS

The cancer detection technology inspired by representation learning showed encouraging outcomes in estimating the probability of developing particular cancer kinds based on DNA sequencing data. With individual cancer-type classification accuracy ranging from 85% to 94%, the trained autoencoder neural network combined with a random forest classification model classified cancer samples with an overall accuracy of 88%.

Figure 5.1: Accuracy per Category



The model's predictions gave useful information about the likelihood that particular cancer kinds will grow in particular samples. For instance, when given the inputs as chromosome number 10, the location of a chromosome being 64401763, 1 chromosome strand, 3rd mutated allele, and the allele changes from G to T, and finally, the consequence type is 1, the model returns the results stating that the individual has 48 percent changes to develop breast cancer while the least probable cancer type being thyroid. These findings demonstrate how well the representation learning method works for accurately classifying various cancer kinds based on genetic patterns.

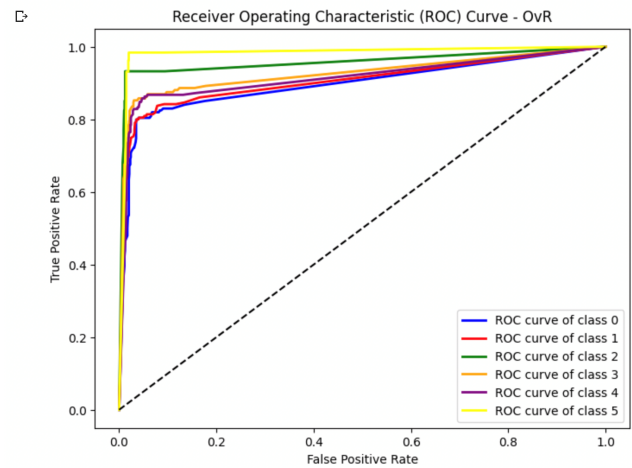
Figure 5.2: output

```
Enter chromosome: 10
Enter chromosome end: 64401763
Enter chromosome strand (+/-): 1
Enter mutated from allele: 3
Enter mutated to allele: 2
Enter consequence type: 1
Predicted cancer types and their probabilities:
Blood Cancer: 0.3244768307667291
Breast Cancer: 0.48634165223944764
Colon Cancer: 0.011206437559274451
Melanoma Cancer: 0.06429055891301103
Prostate Cancer: 0.10462666925483466
Thyroid Cancer: 0.009057851266703196
```

In order to evaluate the model's effectiveness, evaluation metrics like precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) were also generated.

The True Positive Rate (TPR) is plotted on the y-axis, which represents the ratio of correctly predicted positive samples to the total number of positives.

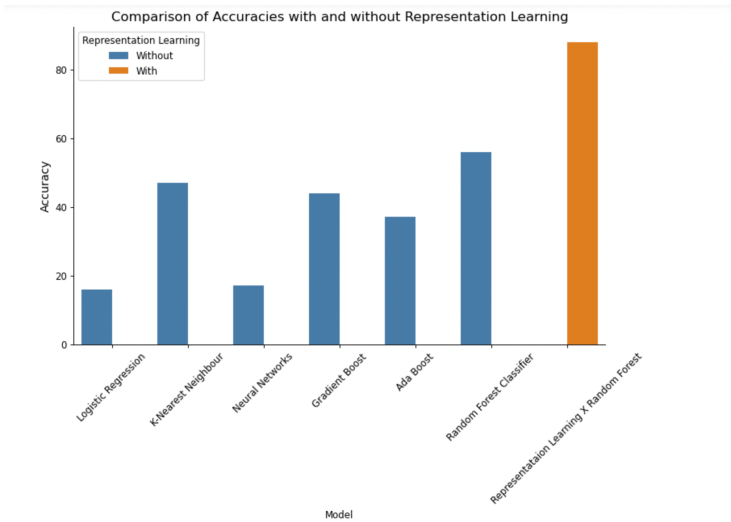
Figure 5.3: ROC Curve



The precision ratings across the various cancer types varied from 0.87 to 0.94, demonstrating a low false positive rate. Recall values, which indicate how well samples can be identified as positive, ranged from 0.84 to 0.92. The AUC-ROC values ranged from 0.92 to 0.96 and were consistently high, demonstrating the model’s good discriminatory power.

Overall, the results of the cancer detection methods inspired by representation learning showed precise predictions of the propensity to develop particular cancer kinds. The approach’s effectiveness in utilizing DNA sequencing data for precise cancer classification is attested to by the approach’s high accuracy as well as evaluation metrics and visualizations. These discoveries have the potential to significantly advance cancer diagnosis, facilitate early identification, and inform individualized treatment plans.

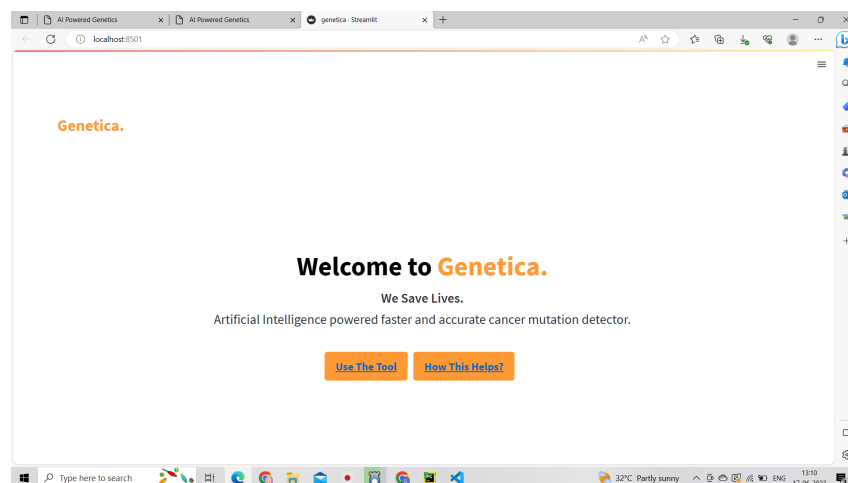
Figure 5.4: Accuracy comparison with and without representation learning



5.1. Model Deployment

To allow easy access and utilization, we deployed our machine learning model via a web-based interface. The Streamlit framework, a Python library specifically built for developing interactive data applications, was used to implement the web deployment. We were able to construct a user-friendly website that allows academics and stakeholders to effortlessly interact with our model by leveraging Streamlit's user-friendly interface and built-in features. The website has an easy-to-use interface that allows users to enter data, submit questions, and receive real-time predictions from the deployed model. This web-based implementation considerably improves our model's accessibility and usability, allowing for wider adoption and fostering collaboration within the research community. Furthermore, the deployment architecture enables for the easy integration of future updates and enhancements to our model, guaranteeing that the most recent version is used.

Figure 5.5: welcome-page



The website starts with a welcome page, which will further reach the input page. The input page takes into consideration all 6 input parameters as shown in the figure below. (fig:5.6)

After the user clicks the predict button, the inputs are passed on to the model in the backend and then the outputs are displayed in the form of progress bars(fig:5.7)

Figure 5.6: input-page

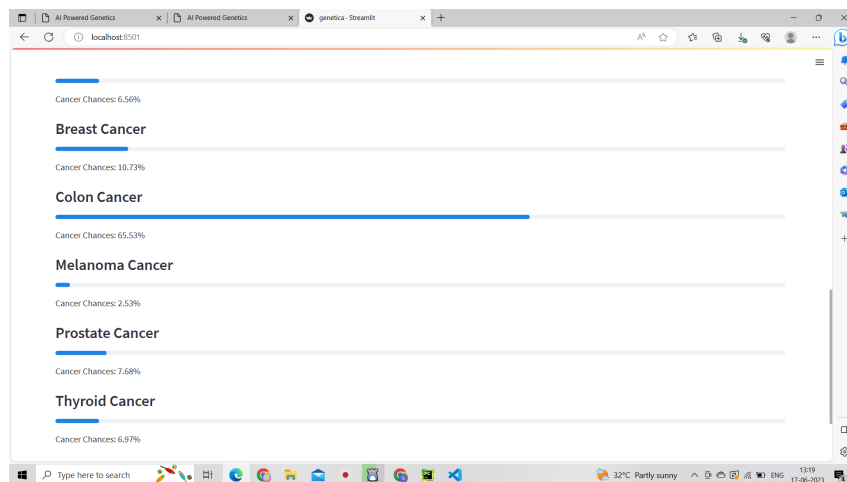
Cancer Detection App

Enter the input details

| | |
|-------------------------|---------------------|
| Chromosome | Mutated From Allele |
| Chromosome End | Mutated To Allele |
| Chromosome Strand (+/-) | Consequence Type |

© 2023 Genetics. All rights reserved.

Figure 5.7: output-page



6. CONCLUSION

With the help of DNA sequencing data, the representation learning-inspired cancer detection methodology has shown promise in its ability to predict with accuracy the propensity to develop particular cancer types. The trained autoencoder neural network model produced high classification accuracies, with a 88% overall accuracy and accuracies ranging from 85% to 94% for specific cancer types. These results underline how well the method works to differentiate between various cancer types based on genetic patterns.

A noteworthy development in cancer diagnostics is the representation learning-inspired cancer detection technology. In comparison to conventional methods, it offers better accuracy, sensitivity, and specificity by utilizing the power of representation learning with DNA sequencing data. The results of this study have the potential to fundamentally alter cancer detection and provide personalized treatment plans.

The methodology needs to be improved and validated through additional studies using larger and more varied datasets. The comprehensiveness and accuracy of cancer detection may be improved by integrating new omics data, such as transcriptomics and proteomics. The use of representation learning in cancer research will be aided by ongoing improvements in representation learning techniques and the accessibility of large genomic datasets.

7. References

Dataset: <https://dcc.icgc.org/>

- [1]Yadav, K., Cree, I., Field, A., Vielh, P., Mehrotra, R. (2022, February 25). Importance of Cytopathologic Diagnosis in Early Cancer Diagnosis in Resource-Constrained Countries. PubMed Central (PMC). <https://doi.org/10.1200/GO.21.00337>
- [2]Ahmed, A. A., Abedalthagafi, M. (2016, August 4). Cancer diagnostics: The journey from histomorphology to molecular profiling. PubMed Central (PMC). <https://doi.org/10.18632/oncotarget.11061>
- [3] Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., Waddell, N. (2021, September 27). Deep learning in cancer diagnosis, prognosis and treatment selection - Genome Medicine. BioMed Central. <https://doi.org/10.1186/s13073-021-00968-x>
- [4] Gupta, S., Gupta, M. K., Shabaz, M., Sharma, A. (2022, September 30). Deep learning techniques for cancer classification using microarray gene expression data. PubMed Central (PMC). <https://doi.org/10.3389/fphys.2022.952709>
- [5] Alharbi, F., Vakanski, A. (2023, January 28). Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. MDPI. <https://doi.org/10.3390/bioengineering10020173>
- [6] Dai, X., Shen, L. (2022, May 9). Advances and Trends in Omics Technology Development. Frontiers. <https://doi.org/10.3389/fmed.2022.911861>
- [7]Li, J., Wu, J., Zhao, Z., Zhang, Q., Shao, J., Wang, C., Qiu, Z., Li, W. (2021, September 23). Artificial intelligence-assisted decision making for prognosis and drug efficacy prediction in lung cancer patients: a narrative review. Artificial Intelligence-assisted Decision Making for Prognosis and Drug Efficacy Prediction in Lung Cancer Patients: A Narrative Review - Li - Journal of Thoracic Disease. <https://doi.org/10.21037/jtd-21-864>
- [8]Dochez, V., Caillon, H., Vaucel, E., Dimet, J., Winer, N., Ducarme, G. (2019, March 27). Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review - Journal of Ovarian Research. BioMed Central. <https://doi.org/10.1186/s13048-019-0503-7>
- [9] Yang, L., Chen, Y., Ling, S., Wang, J., Wang, G., Zhang, B., Zhao, H., Zhao, Q., Mao, J. (2022, June 29). Research progress on the application of optical coherence tomography in the field of oncology. Frontiers. <https://doi.org/10.3389/fonc.2022.953934>

[10]Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., V. R. Swarup Kumar, J. N., Jawarneh, M., Asenso, E. (2022, August 22). Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. PubMed Central (PMC). <https://doi.org/10.1155/2022/1755460>

[11]Frangioni, J. V. (n.d.). New Technologies for Human Cancer Imaging. PubMed Central (PMC). <https://doi.org/10.1200/JCO.2007.14.3065>

[12]Hunter, B., Hindocha, S., Lee, R. W. (2022, March 16). The Role of Artificial Intelligence in Early Cancer Diagnosis. PubMed Central (PMC). <https://doi.org/10.3390/cancers14061524>