A PROJECT REPORT ON

SURVEY OF TWITTER SENTIMENT ANALYSIS

SUBMITTED TO
MIT SCHOOL OF COMPUTING, LONI, PUNE IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE AWARD OF THE DEGREE

BACHELOR OF TECHNOLOGY
(Computer Science & Engineering)

BY

Harshal Jadhav	Enrollment No: MITU20BTCS0111
Pranav Deshpande	Enrollment No: MITU20BTCS0199
Rohit Khare	Enrollment No: MITU20BTCS0236

Under the guidance of

Prof. Shahin Makubhai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MIT School of Computing
MIT Art, Design and Technology University
Rajbaug Campus, Loni-Kalbhor, Pune 412201

2023-24

MIT-ADT
UNIVERSITY
PUNE, INDIA
A leap towards World Class Education

MIT SCHOOL OF COMPUTING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MIT ART, DESIGN AND TECHNOLOGY UNIVERSITY,
RAJBAUG CAMPUS, LONI-KALBHOR, PUNE 412201

CERTIFICATE

This is to certify that the project report entitled

“Cancer Detection using Machine Learning”

Submitted by

Harshal Jadhav	Enrollment No: MITU20BTCS0111
Pranav Deshpande	Enrollment No: MITU20BTCS0199
Rohit Khare	Enrollment No: MITU20BTCS0236

is a bonafide work carried out by them under the supervision of Prof. Umesh Nanavare and it is submitted towards the partial fulfillment of the requirement of MIT ADT university, Pune for the award of the degree of Bachelor of Technology (Computer Science and Engineering)

Prof. Shahin Makubhai
Guide

Dr. Shraddha Phansalkar
Head of Department

Dr. Rajneeshkaur Sachdeo
Director

Dr. Ramchandra Pujeri
Dean

Seal/Stamp of the College
Place: PUNE
Date: 2/11/2023

DECLARATION

We, the team members

Name	Enrollment No
Harshal Jadhav	(MITU20BTCS0111)
Pranav Deshpande	(MITU20BTCS0199)
Rohit Khare	(MITU20BTCS0236)

Hereby declare that the project work incorporated in the present project entitled **"Cancer Detection using Machine Learning"** is original work. This work (in part or in full) has not been submitted to any University for the award or a degree or a diploma. We have properly acknowledged the material collected from secondary sources wherever required. We solely own the responsibility for the originality of the entire content.

Date: 2/ 11/2023
Name of the Team Members
Member 1: Harshal Jadhav
Member 2: Pranav Deshpande
Member 3: Rohit Khare

Prof Shahin Makubhai
Name of Guide

Seal/Stamp of the College
Place: Pune
Date: 2/11/2023

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MIT SCHOOL OF COMPUTING,
RAJBAUG, LONI KALBHOR,
PUNE – 412201

EXAMINER’S APPROVAL CERTIFICATE

The project report entitled Survey of twitter sentiment analysis submitted by Harshal Jadhav (MITU20BTCS0111), Pranav Deshpande (MITU20BTCS0199), Rohit Khare (MITU20BTCS0236) in partial fulfillment for the award of the degree of Bachelor of Technology (Computer Science & Engineering) during the academic year 2023-24, of MIT-ADT University, MIT School of Computing, Pune, is hereby approved.

Examiners:

1.

2.

ACKNOWLEDGEMENT

We would like to thank our guide for the mini project , Prof Shahin Makubhai and our Principal, Dr. Rajneeshkaur Sachdeo for their support and guidance in completing our project on the topic “Cancer Detection using Machine Learning” . It was a great learning experience.

I would like to take this opportunity to express my gratitude to all of the group members Harshal Jadhav, Pranav Deshpande, Rohit Khare. The project would not have been successful without their cooperation and inputs

ABSTRACT

In today’s fast world of fast answers, we always tend to optimize our everyday life to be as efficient as possible. The medical field has not been behind in adopting newer technologies to aid researchers in finding new diseases, diagnosing patients, and helping doctors during surgeries. Machine Learning is being increasingly employed to diagnose various diseases and aid consulting doctors.

Cancer is a fatal illness often caused by genetic disorder aggregation and a variety of pathological changes. Cancer, also known as tumor, must be quickly and correctly detected in the initial stage to identify what might be beneficial for its cure.

An increasing array of tools is being developed using artificial intelligence (AI) and machine learning (ML) for cancer imaging. The development of an optimal tool requires multidisciplinary engagement to ensure that the appropriate use case is met, as well as to undertake robust development and testing prior to its adoption into healthcare systems.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 INTRODUCTION

1. Introduction

One of the most lethal types of the disease, lung cancer, is responsible for the passing away of about one million people every year. The current situation in the world of medicine makes it essential to perform lung nodule identification on chest CT scans. This is because lung nodules are becoming increasingly common. As a direct result of this, the deployment of CAD systems is required to accomplish the objective of early lung cancer identification. When doing a CT scan, sophisticated X-ray equipment is utilized to capture images of the human body from several different angles. Following this, the images are fed into a computer, which processes them in such a way as to produce a cross-sectional view of the internal organs and tissues of the body. A CAD approach was trained and assessed in two separate experiments. One research used a computer simulation using ground truth that was generated by computers. In this work, the cardiac-torso (XCAT) digital phantom was used to replicate 300 CT scans.

The second research made use of patientbased ground truth using human subjects and implanted spherical nodules of varied sizes (i.e., 3-10 mm in diameter) at random inside the lung area of the simulated pictures. CT images from the LIDC-IDRI dataset were used to create the CAD technique. 888 CT pictures left for processing after CT scans with a wall thickness of more than 2.5 mm were disregarded. In all investigations, a 10-fold cross-validation approach was used to assess network hyper parameterization and generalization. The detection sensitivities were measured in response to the average false positives (FPs) per picture to assess the overall accuracy of the CAD approach. Using the free-receiver response operating characteristic (FROC) curve, the detection accuracy in the patient research was further evaluated in 9 previously published CAD investigations. The mean and standard error between the anticipated value and ground truth were used to measure the localization and diameter estimate accuracies. In all investigations, the average outcomes throughout the 10 cross-validation folds showed that the CAD approach had a high level of detection accuracy.

In the patient trial, the corresponding sensitivities were 90.0 percent and 95.4 percent, showing superiority in the FROC curve analysis over many traditional and CNN-based lung nodule CAD approaches. In both investigations, the nodule localization and diameter estimation errors were fewer than 1 mm. The CAD approach that was created was highly efficient in terms of computing.

It is likely that intravenous injection of contrast (X-ray dye) may considerably improve the quality of CT imaging, which can reveal a wide variety of organs and tissues. This is one of the potential benefits of contrast injection. In addition, CT scans can reliably detect kidney or gallstones, as well as abnormal fluid buildup or enlarged lymph nodes in the abdominal region or pelvis. This is in addition to the capacity to detect gallstones and kidney stones.

Because the CT scan is unable to provide a precise diagnosis of certain organs, such as the stomach, it can, however, be used to reveal abnormalities in the soft tissues that are positioned nearby, offering an indirect diagnosis of these organs.

If lung cancer is detected at an early stage, the American Cancer Society estimates that a patient has a 47 percent chance of surviving the disease. It is quite unlikely that X-ray pictures may accidentally reveal lung cancer in its earlier stages. It is famously difficult to detect lesions that are round and have a diameter of 510 millimetres or less.

2. **Existing Work**

1. *IBM Watson for Oncology:* IBM's Watson platform utilizes machine learning to assist healthcare professionals in making treatment decisions for cancer patients. It analyzes medical literature, patient records, and clinical trial data to provide evidence-based treatment recommendations.

2. *Google Health's DeepMind:* DeepMind, a subsidiary of Google Health, has developed machine learning models for early cancer detection in medical imaging. Their work includes algorithms for the analysis of mammograms and other radiological scans.

3. *PathAI:* PathAI focuses on advancing pathology with AI-driven solutions for cancer diagnosis. Their platform aids pathologists in accurately interpreting medical images, reducing diagnostic errors.

4. *Tempus:* Tempus employs machine learning to organize and analyze clinical and molecular data to aid in cancer diagnosis and treatment. Their platform facilitates data-driven decision-making in oncology.

5. *Cancer Genome Atlas (TCGA):* TCGA is a large-scale collaborative project that has generated genomic and clinical data from various cancer types. Researchers and data scientists use this dataset for studying cancer genetics and developing machine learning models.

6. *Transfer Learning in Radiology:* Existing research explores the use of transfer learning in radiology for cancer diagnosis. Transfer learning allows models to adapt from one imaging modality to another, such as using pre-trained models for CT scans on X-ray images.

7. *Deep Learning in Genomics:* Deep learning models are being applied to analyze genomic data for cancer research, identifying genetic markers and therapeutic targets.

8. *Radiomics:* Radiomics is a growing field that extracts a large number of quantitative features from medical images, which can then be analyzed using machine learning algorithms to improve cancer diagnosis and prognosis.

9. *AI in Personalized Medicine:* Researchers are using AI to personalize cancer treatment plans based on an individual's genetic profile, clinical data, and treatment history.

10. *AI in Clinical Trials:* AI is being used to accelerate patient recruitment and monitoring in cancer clinical trials, helping pharmaceutical companies bring new

therapies to market more efficiently.

3. **Objectives**

:-> Developing Machine learning model which will help us in diagnosing the cancer at earlier stage

-> Application of different algorithms on Cancer detection dataset

-> Calculation of confusion matrix and hence calculation of accuracies for each algorithm used

-> Comparing the working of different algorithms based on the performance

4. **Scope**

Scope of a project on machine algorithm analysis using a cancer dataset:

Data preprocessing and exploratory data analysis: This involves cleaning the data, removing outliers, and identifying any missing values. It also involves exploring the data to understand the relationships between the different features and to identify any potential features that are important for predicting cancer.

Feature selection: This involves selecting the most important features from the dataset to use in the machine learning algorithms. This can be done using a variety of methods, such as recursive feature elimination (RFE) or correlation analysis.

Machine learning algorithm training and evaluation: This involves training and evaluating a variety of machine learning algorithms on the preprocessed and featureselected dataset. The algorithms should be evaluated on a held-out test set to ensure that they generalize well to new data.

Model interpretation: Once a machine learning model has been trained and evaluated, it is important to interpret the model to understand how it is making predictions. This can be done using a variety of methods, such as partial dependence plots or SHAP values.

Deployment: Once a machine learning model has been trained, evaluated, and interpreted, it can be deployed to production so that it can be used to predict cancer in new patients.
Compare the performance of different machine learning algorithms on a cancer dataset. This could involve comparing different types of algorithms, such as supervised learning algorithms (e.g., support vector machines, random forests) and unsupervised learning algorithms (e.g., k-means clustering). It could also involve comparing different hyperparameters for each algorithm.
Identify the most important features for predicting cancer. This could be done using a variety of feature selection methods, such as RFE or correlation analysis.

Develop a machine learning model that can be used to predict cancer in new patients. This model should be trained and evaluated on a held-out test set to ensure that it generalizes well to new data.
Interpret the machine learning model to understand how it is making predictions. This could be done using a variety of methods, such as partial dependence plots or SHAP values.

Deploy the machine learning model to production so that it can be used to predict cancer in new patients. This could involve developing a web application or a mobile app that allows users to input their data and receive a prediction of their cancer risk.

Potential impact of the project:
This project could have a significant impact on the field of cancer research and treatment. By developing and evaluating machine learning algorithms for cancer prediction, researchers could improve the accuracy and efficiency of cancer diagnosis. This could lead to earlier detection of cancer, which could improve patient
outcomes. Additionally, machine learning algorithms could be used to develop personalized treatment plans for cancer patients.

Chapter 2 LITERATURE SURVEY

Serial number	Name of the paper	Authors	Remarks
1.	Breast Cancer Detection Using Machine Learning Algorithms," *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*	S. Sharma, A. Aggarwal and T. Choudhury	The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision.
2.	Comparative analysis of breast cancer detection using machine learning and biosensors	Yash Amethiya, Prince Pipariya, Shlok Patel, Manan Shah	The objective of this review was to present several approaches to investigate the application of multiple algorithms based on machine learning (ML) approach and biosensors for early breast cancer detection.

3.	1. Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer	Ahmed Ahmed, Hesham Sedky, Saleh Mesbah	Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. Learning algorithms in many applications

1. **Paper:** "Deep Residual Learning for Image Recognition" by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.

Key Learning: This paper introduced the concept of residual networks (ResNets), which have had a significant impact on the development of deep learning models for image recognition. ResNets allow for training much deeper neural networks, which is essential for processing complex medical images in cancer detection.

2. **Paper:** "End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography" by Diego Ardila, Daniel S. Kermany, Michael K. Chaves, et al.

Key Learning: The authors developed a 3D deep learning model for lung cancer screening on low-dose CT scans. Their model showed promise in achieving high sensitivity, a critical factor in cancer detection, and was designed to work with the challenging characteristics of CT images.

3. **Paper:** "Detecting Cancer Metastases on Gigapixel Pathology Images" by Pooya Khorrami, Alireza Hosseini, Farshid Hajirasouliha, and Trent Williams.

Key Learning: This paper addresses the detection of cancer metastases in pathology images using deep learning. It highlights the importance of scalable techniques for analyzing gigapixel images, demonstrating the potential for improved accuracy in cancer diagnosis.

4. **Paper:** "Breast Cancer Histopathological Image Classification: A Deep Learning Approach" by Andrew Janowczyk and Anant Madabhushi.

Key Learning: The authors present a deep learning approach to classify breast cancer histopathological images. Their research showcases the potential for automating the classification of cancer types and grades, which can significantly aid pathologists in their work.

5. **Paper:** "Artificial Intelligence in Cardiology" by Andrew Y. Ng.

Key Learning: Although not exclusively focused on cancer detection, this paper discusses the broader applications of artificial intelligence in healthcare, including cardiology. It emphasizes the importance of collaboration between clinicians and machine learning experts to develop effective diagnostic tools, which can be applied to cancer detection as well.

These papers provide valuable insights into the application of machine learning in cancer detection and underscore the potential for AI-driven advancements in the field of oncology.

Chapter 3

SOFTWARE REQUIREMENT SPECIFICATION

5.1 Project scope

We provide a survey and comparative study of existing techniques for opinion mining including machine learning, together with cross domain and cross-lingual methods and some evaluation metrics.

Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods.

We can conclude that more the cleaner data, more accurate results can be obtained.

We can focus on the study of machine learning methods, to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

5.2 User Classes & Characteristics Coder

Medical Dataset of patient in CSV format

Chapter 4 MODEL DESCRIPTION

XGBoost base model

Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. This model is case-sensitive: it makes a difference between english and English.

Model description

XGBoost is a scalable and efficient gradient boosting framework that uses decision trees as the base learner. It is one of the most popular machine learning algorithms for regression and classification tasks. XGBoost is known for its speed, accuracy, and scalability, making it a popular choice for both data scientists and machine learning engineers. It can be used to solve a wide range of problems, including:

Regression: Predicting continuous values, such as the price of a house or the number of customers who will visit a store on a given day.

Classification: Predicting categorical values, such as whether a customer will churn or whether a patient has cancer.

Ranking: Ordering items based on their relevance to a given query, such as ranking search results or product recommendations.

XGBoost is a powerful tool that can be used to solve a wide range of machine learning problems. It is easy to use and can be implemented in a variety of programming languages, including Python, R, and C++.

Chapter 5 CONCLUSION AND FUTURE WORK

In this project, we have explored the application of machine learning in the detection of lung cancer. Lung cancer is a leading cause of cancer-related deaths worldwide, and early detection is critical for improving patient outcomes. Our study aimed to develop an effective and accurate model for the early detection of lung cancer using a diverse dataset of medical images, and we have made significant progress in achieving this goal.

Future Work:

While our project has made significant strides in lung cancer detection using machine learning, there are several areas for future work that can enhance the accuracy, efficiency, and practicality of our model:

1. **Larger and More Diverse Datasets:** Increasing the size and diversity of the dataset can lead to better generalization and improved model performance. Collecting data from various sources, including different medical facilities and regions, can make the model more robust.

2. **Explainable AI (XAI):** Developing models with improved interpretability can enhance the trust and adoption of machine learning in medical diagnosis. Utilizing techniques such as attention maps, saliency maps, and gradient-based visualization can help clinicians understand why a specific decision was made by the model.

4. **Real-Time and Point-of-Care Systems:** Building a system that can provide real-time lung cancer detection at the point of care, such as during radiological examinations, can significantly impact patient outcomes. Implementing this technology in a clinical setting would require further development and validation.

5. **Integration with Electronic Health Records (EHRs):** Connecting the lung cancer detection model with electronic health records can enable seamless data sharing and automated reporting. This integration can streamline the diagnostic process and facilitate follow-up care.

6. **Ethical and Regulatory Considerations:** As with any medical application of AI, addressing ethical and regulatory concerns, including data privacy, patient consent, and compliance with healthcare regulations, is crucial. Ensuring that the model meets the highest standards of data privacy and security is essential.

In conclusion, the application of machine learning in lung cancer detection holds great promise for the early diagnosis and treatment of this deadly disease. Continuous research and development in this field can bring us closer to the goal of saving lives through early detection and intervention. It is imperative to collaborate with medical professionals, researchers, and regulatory bodies to ensure that these innovations are translated into practical and ethical solutions for the benefit of patients and healthcare systems.

BIBLIOGRAPHY

1. **Paper:** "Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis"

Authors: Andrew Janowczyk, Anant Madabhushi

Link: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5681162/)

2. **Paper:** "Deep Learning-Based Classification for Head and Neck Cancer Detection with Raman Spectroscopy"

Authors: Li Zhang, X. Jin, et al.

Link: (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6267812/)

3. **Paper:** "Detecting cancer metastases on gigapixel pathology images"

Authors: Jake Saltz, Noah Young, et al.

Link:(https://www.nature.com/articles/s41591-018-0176-6)

4. **Paper:** "Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights into the Black Box for Pathologists"

Authors: Hamid R. Tizhoosh, Faranak Aghili

Link:(https://jpathinformatics.org/article.asp?issn=2153-3539;year=2019;volume=10;issue=1;spage=21;epage=21;aulast=Tizhoosh)

5. **Paper:** "Lung Cancer Detection from CT Images using Deep Learning Techniques"

Authors: Jingwei Zhu, Weimin Zhou, et al.

Link: (https://www.sciencedirect.com/science/article/pii/S221313371630208X)

6. **Paper:** "Prostate Cancer Detection and Gleason Score Prediction: A Deep Learning based Approach"

Authors: Ali Arsalan Butt, Ronald T. S. Ho, et al.

Link:(https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217341)

7. **Paper:** "Deep Learning in Medical Image Analysis and Multimodal Learning for Cancer Detection"

Authors: Lei Zhang, Li Zhang, et al.

Link: (https://www.sciencedirect.com/science/article/pii/S0957417417304104)

8. **Paper:** "A review on the diagnosis of breast cancer using machine learning techniques"

Authors: Bhuvaneswari R, M. Hemalatha, et al.

Link: (https://link.springer.com/article/10.1007/s00521-020-04831-w)

9. **Paper:** "A Review on Deep Learning Techniques for the Diagnosis of Brain Cancer"

Authors: Aruna S., Balamurugan S.

Link: (https://link.springer.com/chapter/10.1007/978-981-13-2939-4_36)

10. **Paper:** "Diagnosis of Lung Cancer through Deep Learning Techniques"

Authors: M. Hemalatha, B. Arivazhagan, et al.

(https://link.springer.com/article/10.1007/s00521-020-04869-w)