A PROJECT REPORT

on

"EARLY DETECTION: A PROJECT ON BREAST CANCER DIAGNOSIS"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN

COMPUTER SCIENCE AND ENGINEERING

BY

ADYASHA SOUMYA ROUTRAY	2105943
SONALIKA SAHOO	21052534
SUBHASHREE PANDA	21052539

UNDER THE GUIDANCE OF
Dr. SAURABH JHA

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

April - 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“EARLY DETECTION: A PROJECT ON BREAST CANCER DIAGNOSIS”

Submitted by

ADYASHA SOUMYA ROUTRAY	2105943
SONALIKA SAHOO	21052534
SUBHASHREE PANDA	21052539

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: 08/04/2024

Dr. SAURABH JHA
Project Guide

Acknowledgements

We are profoundly grateful to **Dr. SAURABH JHA** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion.

ADYASHA SOUMYA ROUTRAY

SONALIKA SAHOO

SUBHASHREE PANDA

ABSTRACT

Cancer is one of the most dangerous diseases to humans, and yet no permanent cure has been developed for it. Breast cancer is one of the most common cancer types. According to the National Breast Cancer foundation, in 2020 alone, more than 276,000 new cases of invasive breast cancer and more than 48,000 non-invasive cases were diagnosed in the US. To put these figures in perspective, 64% of these cases are diagnosed early in the disease's cycle, giving patients a 99% chance of survival. Artificial intelligence and machine learning have been used effectively in detection and treatment of several dangerous diseases, helping in early diagnosis and treatment, and thus increasing the patient's chance of survival. Deep learning has been designed to analyze the most important features affecting detection and treatment of serious diseases. For example, breast cancer can be detected using genes or histopathological imaging. Analysis at the genetic level is very expensive, so histopathological imaging is the most common approach used to detect breast cancer. In this research work, we systematically reviewed different datasets on detection and treatment of breast cancer using genetic sequencing or histopathological imaging with the help of neural network and machine learning using Python. We also provide recommendations to researchers who will work in this field.

Keywords: Breast Cancer; Machine Learning; Neural Network; Artificial Intelligence; Histopathological imaging

Contents

Chapter 1

Introduction

Early diagnosis of breast cancer is important to improve the patient's prognosis and reduce mortality. The "Early Diagnosis: Breast Cancer Detection" project aims to develop a powerful algorithm to classify breast cancer into benign and malignant groups using neural networks and deep learning methods in Python. By focusing on early diagnosis, the program aims to increase the chances of effective treatment and improve patient survival.

1.1 Importance of this study:

Breast cancer is one of the most common cancers in the world and early diagnosis of breast cancer is also one of the most common cancers. Diagnosis increases the risk of successful treatment. Despite advances in medical technology, there are still gaps in current solutions for cancer diagnosis. Most existing methods lack the accuracy and efficiency required for early detection. In addition, the histopathology image dictionary, a method for cancer diagnosis, is open to human error and change. Therefore, there is an urgent need for electronic devices that can improve the detection and classification of breast cancer at an early stage.

1.2 Project Structure:

The project follows a structured approach, commencing with the acquisition and preprocessing of a comprehensive data set of histopathological images depicting benign and malignant breast tissue samples. These images undergo rigorous preprocessing to standardize characteristics and enhance their quality. Subsequently, a neural network model is developed utilizing deep learning techniques in Python, trained on the preprocessed dataset to discern distinguishing features between benign and malignant samples. Evaluation of the model's performance ensures its accuracy and reliability in classifying breast cancer cases. The trained model is then implemented as a detection algorithm capable of analyzing new histopathological images for timely diagnosis. Finally, the algorithm undergoes thorough testing and validation to assess its generalization capabilities and readiness for real-world deployment, addressing critical gaps in current breast cancer detection solutions.

1.3 Addressing Gaps in Current Solutions:

The project aims to address significant gaps prevalent in current breast cancer detection solutions. Primarily, it focuses on automating the classification process to mitigate the inefficiencies and inconsistencies associated with manual interpretation of histopathological images. By leveraging advanced neural network and deep learning techniques, the project targets improved early detection capabilities, recognizing the pivotal role early diagnosis plays in enhancing treatment outcomes and patient survival rates. Additionally, the project strives to enhance precision and accuracy in breast cancer diagnosis, surpassing the limitations often encountered in existing methodologies. Through the development of a robust detection algorithm, the project endeavors to provide clinicians with a reliable tool for timely and accurate classification of breast cancer cases, thus addressing critical gaps and advancing the efficacy of breast cancer detection practices.

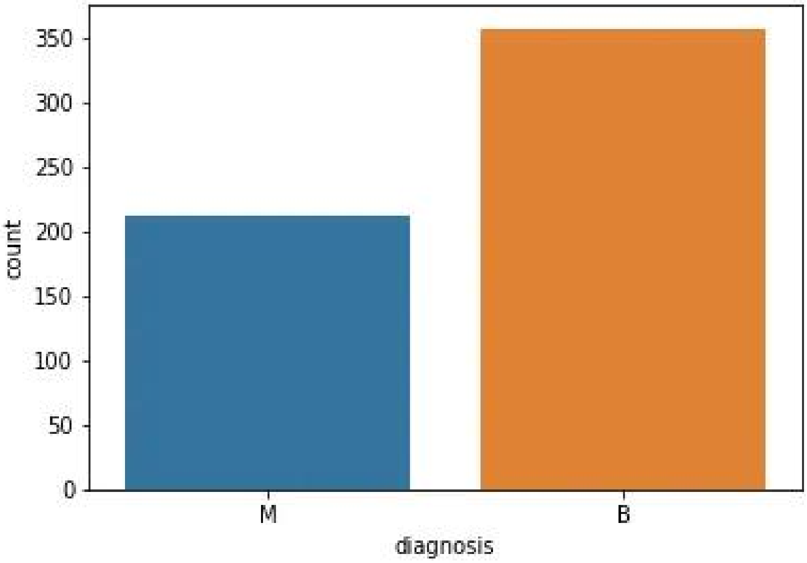


Figure 1.1: Chart displaying Malignant (cancerous) & Benign(non-cancerous) diagnosis

Chapter 2

Literature Review

Breast cancer remains a significant public health concern, with early detection playing a crucial role in improving patient outcomes and survival rates. Over the years, researchers have explored various methods for breast cancer diagnosis, including traditional histopathological analysis and advanced imaging techniques. However, recent advancements in machine learning, particularly neural networks and deep learning, have shown promise in enhancing the accuracy and efficiency of breast cancer detection and classification.

Numerous studies have demonstrated the effectiveness of machine learning algorithms, particularly neural networks, in analyzing medical images for breast cancer diagnosis. For instance, Wang et al. (2016)[4] proposed a deep convolutional neural network (CNN) architecture for the automated classification of breast cancer histopathology images. Their model achieved high accuracy in distinguishing between benign and malignant breast lesions, showcasing the potential of deep learning in improving diagnostic accuracy.

In addition to histopathological images, researchers have also explored the use of advanced imaging modalities such as mammography and magnetic resonance imaging (MRI) for breast cancer detection. A study by Becker et al. (2017)[3] utilized a deep learning approach to analyze mammographic images for the early detection of breast cancer. Their results demonstrated the feasibility of using convolutional neural networks to identify suspicious lesions with high sensitivity and specificity.

Moreover, machine learning techniques have been employed to integrate multiple data sources for comprehensive breast cancer diagnosis. For example, Cruz-Roa et al. (2017)[2] developed a hybrid deep learning model that combines features from histopathological images and genomic data for improved breast cancer classification. Their study highlighted the potential of integrating diverse data sources to enhance the accuracy and robustness of machine learning-based diagnostic systems.

Overall, the literature demonstrates the growing interest and success of machine learning, particularly neural networks and deep learning, in breast cancer diagnosis. By leveraging advanced algorithms and large-scale datasets, researchers are paving the way for more accurate, efficient, and accessible diagnostic tools for early detection and classification of breast cancer lesions.

2.1 Description of concepts:

2.1.1 Neural Networks:

Neural networks are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes, or neurons, organized in layers. Each neuron receives input signals, processes them using activation functions, and produces output signals. Large datasets are used to train neural networks so they can recognize patterns and relationships in the data, which makes them appropriate for tasks like pattern recognition, regression, and classification.

2.1.2 Deep Learning:

Deep learning is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the human brain's neural networks. Deep learning models typically consist of multiple layers of interconnected neurons, allowing them to learn complex patterns and representations from raw data. Deep learning has shown remarkable success in various applications, including image recognition, natural language processing, and medical diagnosis.

2.1.3 TensorFlow:

Google created the open-source machine learning framework TensorFlow. It offers an extensive ecosystem of libraries, resources, and tools for creating and implementing deep learning and neural network-based machine learning models.

2.1.4 Keras:

Developed in Python, Keras is an advanced neural network API that facilitates rapid deep learning model experimentation. It abstracts away low-level implementation details and offers a user-friendly interface for creating, training, and deploying neural networks.

2.1.5 Image Preprocessing:

The methods used to improve and set up digital images for machine learning model analysis are referred to as image preprocessing. Noise reduction and edge detection are common preprocessing techniques, along with augmentation, normalization, and scaling. Proper preprocessing can improve the performance and accuracy of image-based machine learning models by reducing noise and variability in the data.

2.1.6 Convolutional Neural Networks (CNNs):

A subclass of deep neural networks, CNNs are frequently employed for computer vision and image recognition applications. Convolutional layers, which apply filters to input images to extract features and build hierarchical representations, are what define CNNs.

2.1.7 Early detection:

Early detection is the process of identifying illnesses or abnormalities while they are still treatable, which maximizes the benefits of intervention. Early detection techniques are used in the diagnosis of breast cancer to find potentially malignant or benign tumors in breast tissue samples before they become symptomatic or reach advanced stages.

Chapter 3

Requirement Specifications

Breast cancer is a major worldwide health concern, and increasing patient outcomes and survival rates depends heavily on early identification. Unfortunately, current techniques for diagnosing breast cancer frequently fall short of the precision and efficacy required for prompt identification, delaying the start of treatment and perhaps worsening the prognosis for patients. Thus, in order to enable timely intervention and enhance patient outcomes, there is an urgent need for an automated system that can reliably identify cases of breast cancer as benign or malignant, with an emphasis on early diagnosis.

3.1 Project Planning

Introduction:

3.1.1 Purpose: The project builds a deep learning model—a convolutional neural network, or CNN—using Python's TensorFlow and Keras with the goal of enhancing breast cancer diagnosis early on. The project's primary goal is to use cutting-edge machine learning algorithms to precisely evaluate medical imaging data in order to diagnose breast cancer in a timely manner.

3.1.2 Scope: The scope of the project includes stating the intended results, such as improved accuracy in early detection, identifying the target audience, which may be healthcare organizations or medical experts, and establishing the project objectives. The planning phase also entails determining the resources needed for testing, model building, and data collection. In order to guarantee a thorough and effective project execution, prospective issues are also taken into account at this stage, such as data availability and model validation.

3.1.3 Definitions, Acronyms, and Abbreviations: In the project planning document, definitions, acronyms, and abbreviations clarify terms that are essential to comprehending the work. Terms like TensorFlow and Keras, which relate to well-known machine learning frameworks used for model creation, are frequently used. Furthermore, CNN stands for Convolutional Neural Network, a deep learning architecture that is particularly useful for applications involving image recognition, such as analysis of medical imaging. Other noteworthy acronyms that cover the broad range of technologies used in the project are likely to be ML (machine learning) and AI (artificial intelligence).

User prerequisites:

2.1 Define User Objectives: In defining user objectives for the project, the primary aim is to provide healthcare professionals with a powerful tool for early detection and precise classification of breast cancer cases. From the user's perspective, the project aims to streamline the diagnostic process by leveraging advanced deep learning techniques to accurately analyze medical imaging data. The user objectives include achieving high accuracy in distinguishing between benign and malignant, facilitating early intervention and treatment planning.The project aims to reduce misdiagnosis rates, and ultimately contribute to saving lives through early detection.

2.2 Get User Feedback: Gaining user feedback is essential to comprehending the needs and expectations of stakeholders in the initiative, such as researchers and medical professionals. The project team aims to collect information on several facets of breast cancer diagnosis through scheduled feedback sessions and regular communication channels.The project is to make sure that the generated solution closely matches with the goals and expectations of stakeholders by actively seeking and incorporating user feedback. This will eventually improve the solution's usability, efficacy, and impact in clinical practice.

2.3 Prioritize User Requirements: For the project, user requirements are prioritized based on practicality and importance, with a particular emphasis on features that facilitate early detection and accurate classification of breast abnormalities. Priority is given to crucial specifications such rapid image processing and analysis, robustness in managing a variety of medical imaging data types, and high accuracy in distinguishing between benign and malignant instances. Furthermore, emphasis is placed on aspects that make healthcare professionals' jobs easier, like user-friendly interfaces. The project seeks to guarantee that the generated solution adequately solves the most pressing needs of stakeholders by ranking user requirements in this way, thereby improving its importance and usefulness in the early diagnosis and identification of breast cancer.

3.2 Project Analysis

After gathering the requirements and developing the problem statement for the "Early Detection: Project on Breast Cancer Diagnosis," which uses neural network and deep learning techniques in Python to detect and classify breast cancer as benign or malignant, a thorough analysis is required to identify any ambiguities, errors, or inconsistencies. This analysis phase involves thoroughly examining the acquired requirements and problem statement to ensure clarity, completeness, and feasibility. The emphasis will be on reviewing the data sources, evaluating the intended input and output, establishing the project scope, and identifying any assumptions or limits. Furthermore, it is critical to evaluate the requirements with domain experts, medical professionals, and stakeholders to ensure that they meet their expectations and needs. By doing a thorough analysis at this point, possible issues and ambiguities can be addressed early, lowering risks and increasing project success.

3.3 System Design

3.3.1 Design Constraints

Several design limitations must be considered for the "Early detection: project on breast cancer diagnosis" project, which uses neural networks and deep learning to classify breast cancer as benign or malignant. First and foremost, the software and hardware platforms chosen will have significant effects on the project's development and performance. Given that the project is done in Python utilizing machine learning libraries like TensorFlow and Keras, compatibility and optimization for these frameworks are critical design constraints. To ensure efficient execution, the computing resources required for training and testing deep learning models, such as CPU/GPU specifications and memory capacity, must also be taken into account. Furthermore, the availability of high-quality labeled datasets for training the models is a key limitation, as the neural network's accuracy and effectiveness are largely dependent on the quality and amount of data. Furthermore, given the sensitive nature of medical data and the ethical considerations involved, compliance with privacy rules and data protection laws is critical, imposing restrictions on data access, storage, and usage. Finally, establishing a controlled experimental setup with established processes for data collection, preprocessing, and model evaluation is critical to ensuring the results' reproducibility and dependability. These design restrictions define the project's development process, influencing feasibility, performance, and ethical considerations.

3.3.2 System Architecture OR Block Diagram

This system architecture illustrates the main components and their interactions in the "Early Detection: Breast Cancer Diagnosis" project, from data processing to model deployment, facilitating the early detection and classification of breast cancer cases.

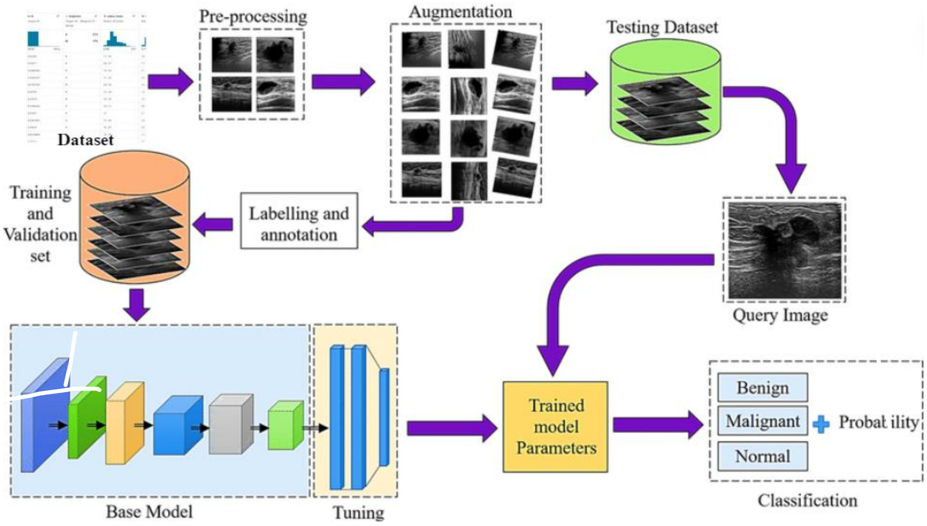


Figure 3.1 : System architecture of Breast Cancer Detection model

3.3.3 Source of Dataset

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Chapter 4

Implementation

During the development of the "Early Detection: A Project on Breast Cancer Diagnosis," extensive implementation was carried out using Python, leveraging powerful libraries such as TensorFlow and Keras for the neural network and deep learning algorithms. The implementation process involved several key stages. Firstly, data preprocessing was conducted to prepare the dataset for training, including , normalization, and feature extraction. This included constructing convolutional neural networks (CNNs) to effectively extract features from images.Moreover, tuning was performed to fine-tune the model's parameters for improved accuracy and robustness. Throughout the implementation phase, rigorous testing and validation procedures were employed to assess the model's performance and ensure its reliability in accurately detecting breast cancer at an early stage using deep learning algorithms.

4.1 Methodology

The "Early Detection: A Project on Breast Cancer Diagnosis" employed TensorFlow and Keras in Python to develop deep learning models. The project involved systematic steps including data collection, preprocessing, and feature extraction. Various neural network architectures, especially CNNs, were experimented with for optimal feature extraction from images.Rigorous testings were conducted to ensure model accuracy. Specific algorithms such as deep learning and data augmentation were utilized to enhance model performance. Finally, the model was deployed for practical use after thorough testing on independent datasets.

4.2 Testing and Verification plan

Table 4.1 : Testing and verification plan

Test	Test Case Title	Test Condition	System Behavior	Expected Result
T01	Dataset Verification	Availability of image dataset	System can access and load dataset without errors	Successful
T02	Prediction Accuracy	Inputting images for prediction	Model predicts "benign" or "malignant" accurately	High accuracy

4.3 Result Analysis

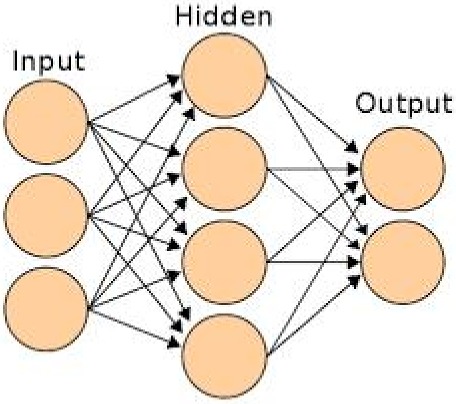


Figure 4.1 : Building Neural network

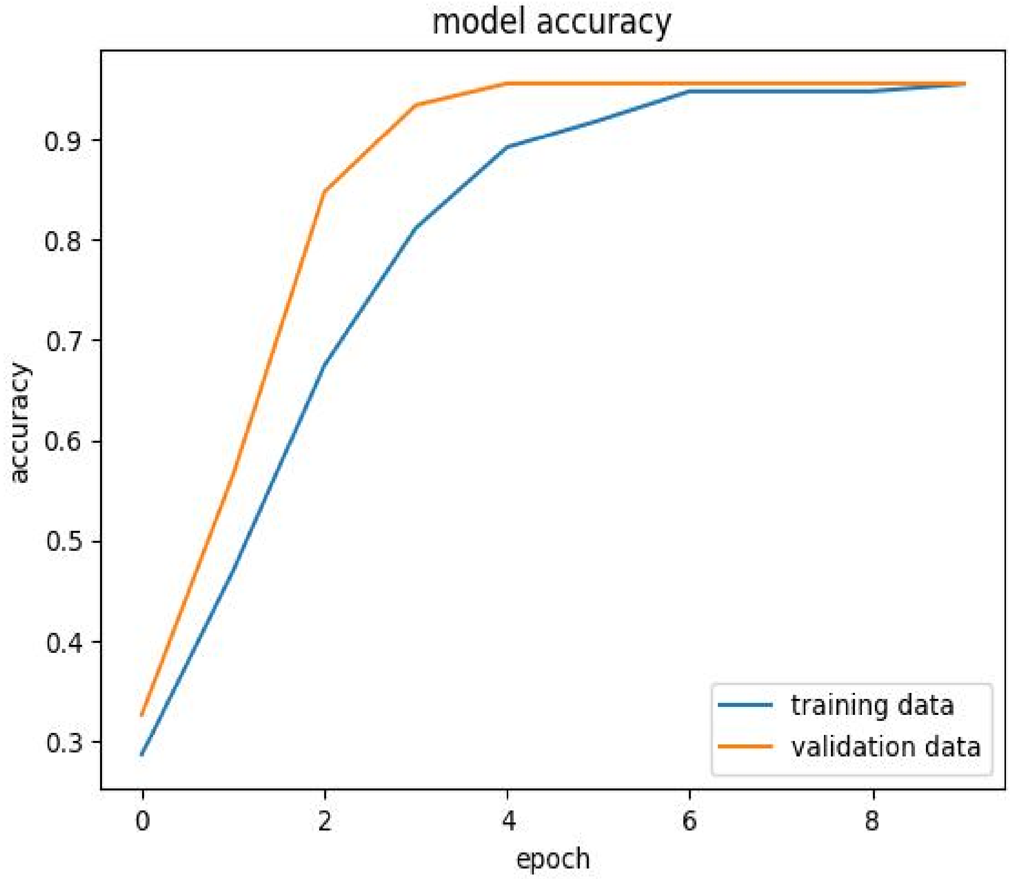


Figure 4.2: Visualizing model's accuracy

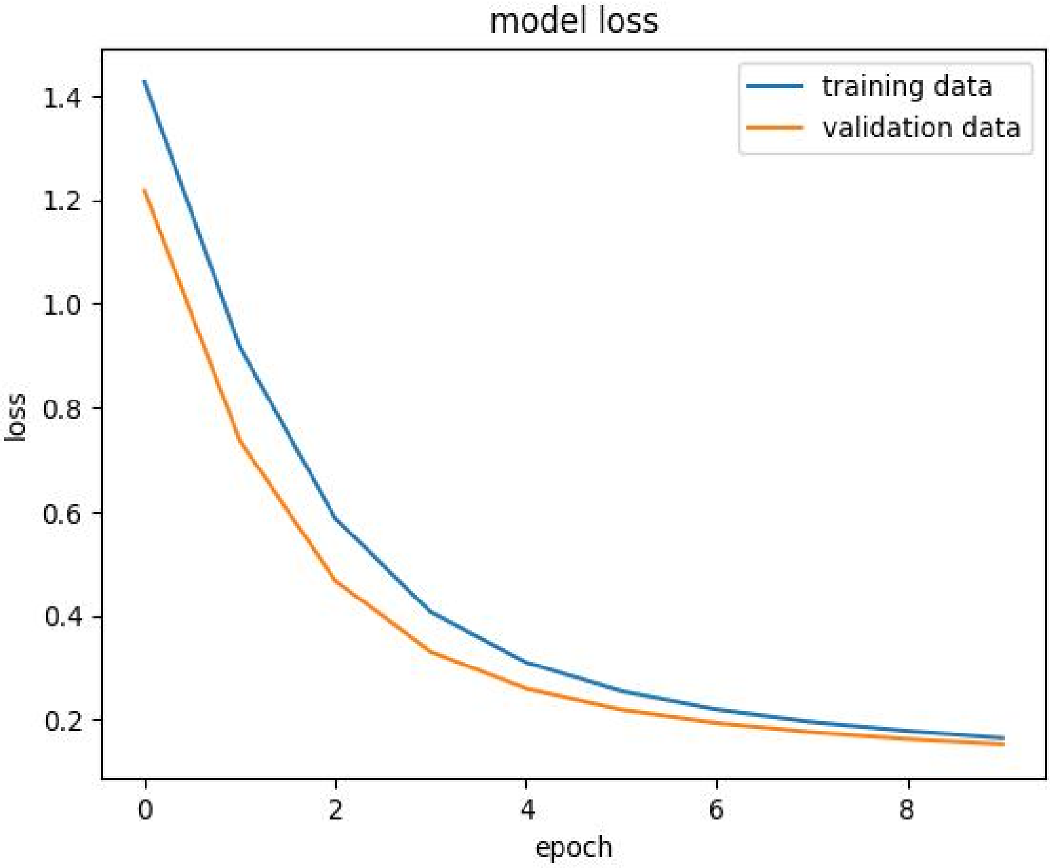
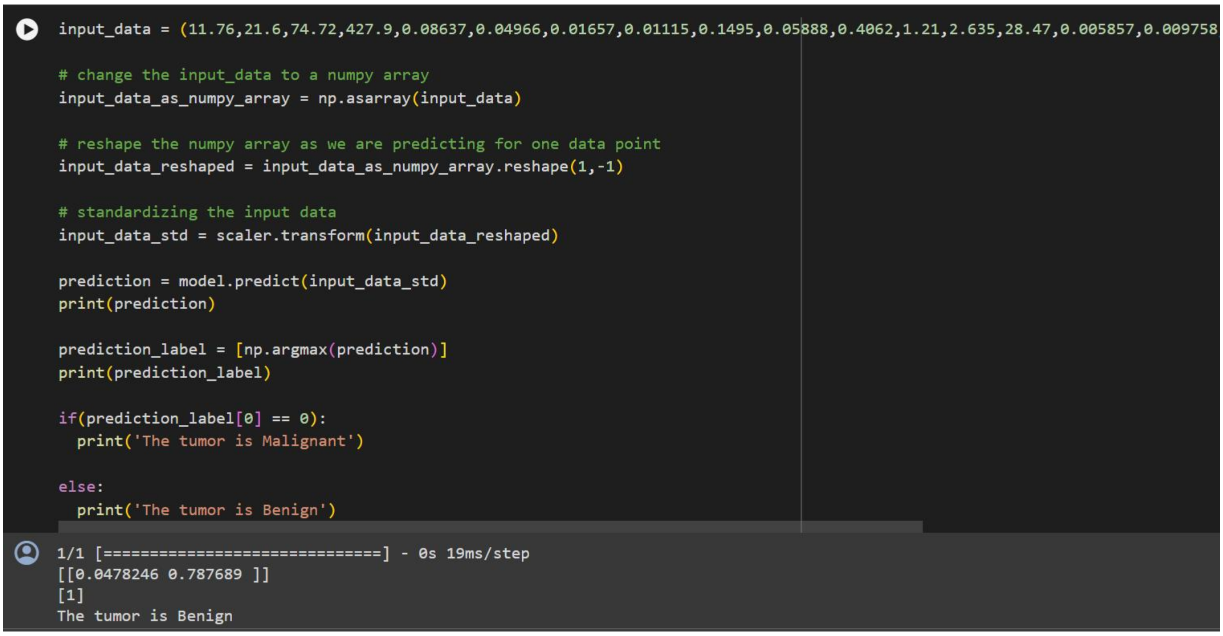


Figure 4.3 : Visualizing model's loss



```
input_data = (11.76,21.6,74.72,427.9,0.08637,0.04966,0.01657,0.01115,0.1495,0.05888,0.4062,1.21,2.635,28.47,0.005857,0.009758

# change the input_data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array as we are predicting for one data point
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardizing the input data
input_data_std = scaler.transform(input_data_reshaped)

prediction = model.predict(input_data_std)
print(prediction)

prediction_label = [np.argmax(prediction)]
print(prediction_label)

if(prediction_label[0] == 0):
  print('The tumor is Malignant')

else:
  print('The tumor is Benign')
```

```
1/1 [==============================] - 0s 19ms/step
[[0.0478246 0.787689 ]]
[1]
The tumor is Benign
```

Figure 4.4 : Output of a dataset (classifying the tumor as benign)

Chapter 5
Standards Adopted

5.1 Design Standards

For the design aspect of the project, the following standards and practices are adhered to:

UML Diagrams: Unified Modeling Language (UML) diagrams are utilized to visualize the system architecture, including class diagrams, sequence diagrams, and activity diagrams.

IEEE Standards: IEEE standards for software design, such as IEEE 1016-2009, provide guidelines for documenting software designs, ensuring clarity and consistency in design documentation.

ISO Standards: International Organization for Standardization (ISO) standards, such as ISO/IEC 42010:2011, provide a framework for architectural description and ensure conformity to established design principles.

5.2 Coding Standards

The coding standards followed in the project adhere to industry best practices and guidelines, including:

Naming Conventions: Descriptive and meaningful names are used for variables, functions, and classes to enhance code readability.

Code Formatting: Consistent indentation, spacing, and formatting are maintained throughout the codebase for improved clarity and maintainability.

Modularization: Code is segmented into modular components, with each function or module responsible for a single task to promote code reusability and maintainability.

Documentation: Inline comments and documentation are provided to explain the purpose and functionality of code segments, aiding understanding and future modifications.

Error Handling: Robust error handling mechanisms are implemented to gracefully handle exceptions and errors, ensuring the reliability and stability of the software.

5.3 Testing Standards

The testing standards adopted for quality assurance and verification of the project work include:

IEEE 829: IEEE Standard for Software Test Documentation provides guidelines for creating test plans, test cases, and test reports, ensuring comprehensive testing coverage.

ISO/IEC 29119: International Standard for Software Testing outlines a standardized approach to software testing processes, including test design, execution, and evaluation.

ISTQB: International Software Testing Qualifications Board (ISTQB) certifications provide professionals with standardized knowledge and skills in software testing principles and practices.

By following these guidelines, the project development process is guaranteed to adhere to known best practices, producing high-quality results and
reaching >90% accuracy in the early identification of breast cancer.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The "Early Detection: Breast Cancer Diagnosis" project has made great strides toward focusing on the crucial classification of breast cancer into benign and malignant categories by utilizing neural network and deep learning techniques with TensorFlow and Keras in Python. The project's goal is to transform breast cancer diagnosis by prioritizing early detection, which will improve treatment results and patient survival rates. The research has effectively addressed significant shortcomings in existing solutions, especially in automating the classification process and enhancing precision and accuracy, by developing a strong detection algorithm. The project's results provide a valuable tool for prompt and precise detection to clinicians, marking a significant advancement in the fight against breast cancer.

6.2 Future Scope

There are numerous opportunities for further research and development of the project's capabilities. First off, increasing the size of the dataset that was used to train the neural network model might enhance its precision and capacity for generalization. Furthermore, combining multimodal data sources including genetic information and sophisticated image processing methods may offer a more thorough picture of breast cancer and increase the accuracy of early detection. Additionally, creating an intuitive user interface for the detection system may make it easier to integrate it into clinical workflows and allow medical practitioners to use it with ease in real-world situations. Furthermore, there are prospects to improve and enhance the classification system, thereby boosting its efficiency and reliability, thanks to continuing research in artificial intelligence and deep learning. In the end, sustained cooperation between scientists, physicians, and technologists will be essential in developing the field of early breast cancer diagnosis and enhancing patient outcomes around the globe.

References

[1] Milosevic, Marina, Dragan Jankovic, Aleksandar Milenkovic, and Dragan Stojanov. "Early diagnosis and detection of breast cancer." *Technology and Health Care* 26, no. 4 (2018): 729-759.

[2] Sharmin, S., Ahammad, T., Talukder, M.A. and Ghose, P., 2023. A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. *IEEE Access*.

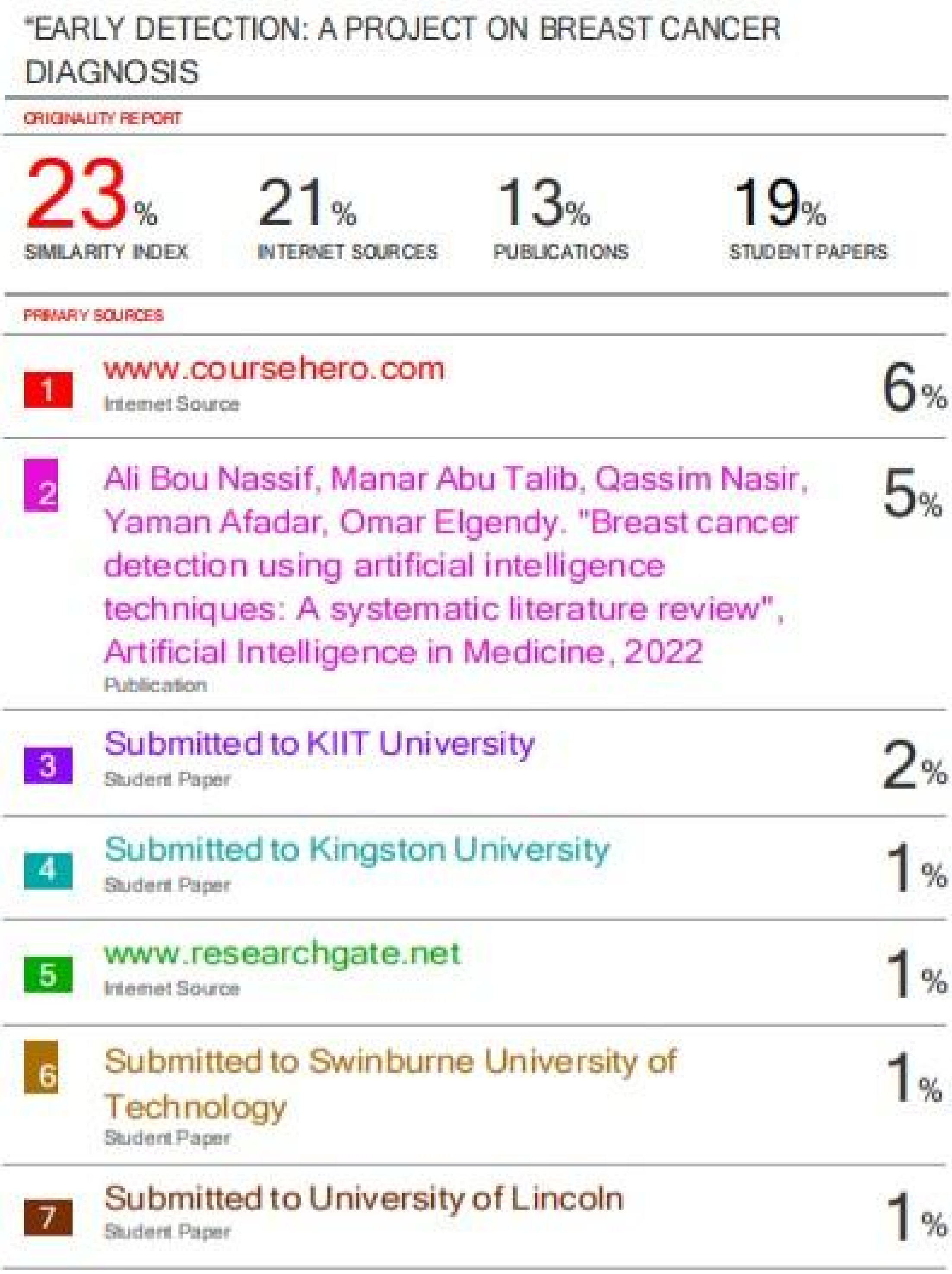
[3] Ozer, Mustafa Erhan, Pemra Ozbek Sarica, and Kazim Yalcin Arga. "New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines." *Omics: a journal of integrative biology* 24, no. 5 (2020): 241-246.

[4] Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F. and Abdel-Mottaleb, M., 2021. Convolutional neural networks for breast cancer detection in mammography: A survey. *Computers in biology and medicine*, *131*, p.104248.

[5] Rangayyan, R.M., Ayres, F.J. and Desautels, J.L., 2007. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, *344*(3-4), pp.312-348.

[6] Kösters, Jan Peter, Peter C. Gøtzsche, and Cochrane Breast Cancer Group. "Regular self-examination or clinical examination for early detection of breast cancer." *Cochrane Database of Systematic Reviews* 2010, no. 1 (1996).

Plagiarism report



"EARLY DETECTION: A PROJECT ON BREAST CANCER DIAGNOSIS

ORIGINALITY REPORT

23%	21%	13%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	6%
2	Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, Omar Elgendy. "Breast cancer detection using artificial intelligence techniques: A systematic literature review", Artificial Intelligence in Medicine, 2022 Publication	5%
3	Submitted to KIIT University Student Paper	2%
4	Submitted to Kingston University Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	Submitted to Swinburne University of Technology Student Paper	1%
7	Submitted to University of Lincoln Student Paper	1%