**Breast Cancer Detection**

Minor project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

# **Computer Science and Engineering**

By

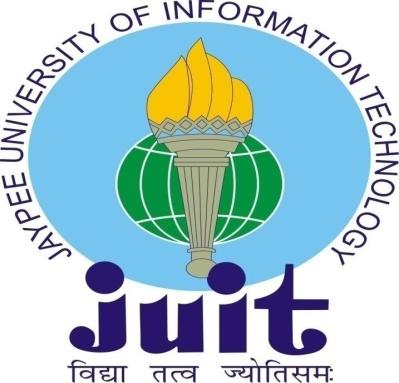
## PARUL SHARMA(191206 )

KUNIKA SHARMA(191227)

RIA MAHAJAN(191236)

**UNDER THE SUPERVISION OF**

RAVINDARA BHATT



Department of Computer Science & Engineering and Information Technology

# **Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA**

**TABLE OF CONTENTS**

| **Title** | **Page No.** |
| --- | --- |
| **Declaration** | I |
| **Certificate** | Ⅱ |
| **Acknowledgement** | Ⅲ |
| **Abstract** | Ⅳ |
| **Chapter-1 (Introduction)**  1.1 Introduction  1.2 Objective  1.3 Motivation  1.4 Language Used  1.5 Technical Requirements  1.5.1 Software Requirements  1.5.2 Hardware Requirements  1.6 Outcomes | 1  2  2  3  3-4  3-4  4  4 |
| **Chapter-2 (Feasibility Study, Requirements Analysis and Design**  2.1 Feasibility Study  2.1.1 Problem Definition  2.1.2 Problem Analysis  2.1.2.1 Related Work  2.1.3 Solution  2.2 UML(Unified Modelling Language) | 5-7  5  5  5-7  7  8 |
| **Chapter-3 (Implementation)**  3.1 Date Set Used in the Minor Project  3.2 Date Set Features  3.2.1 Types of Data Set  3.2.2 Number of Attributes, fields, description of the data  3.2.2.1 Wisconsin Breast Cancer Dataset Description  3.2.2.2 Breast Histopathology Dataset Description  3.3 Design of Problem Statement  3.4 Pseudo code of the Project Problem  3.4.1 Machine Learning  3.4.1.1KNN  3.4.1.2 SVM  3.4.2 Deep Learning  3.5 Flow graph of the Minor Project Problem  3.6 Screenshots of the various stages of Project  3.6.1 Machine Learning  3.6.1.1 Data Preprocessing  3.6.1.2 SVM(Support Vector Machine)  3.6.1.3 KNN(K-Nearest Neighbour)  3.6.2 Deep Learning  3.6.2.1 Custom based CNN Model  3.6.2.2 Resnet-50 Pretrained model  3.6.2.3 VGG-16 Pre-trained Model | 8  8  8-9  9  9-10  10-11  11-12  12  12  13-14  15-16  15  16  16  17-29  18-22  22-26  26-29 |
| **Chapter-4 Results**  4.1 Discussion on the Results Achieved  4.2 Result Validation  4.3 Application of the Minor Project  4.4 Limitation of the Minor Project  4.5 Future Work | 30-33  33-34  34-35  35  36 |
| **References** | 37-38 |

**DECLARATION**

I hereby declare that this project has been done by us under the supervision of Dr. Ravindara Bhatt**,** Associate Professor**,** Jaypee University of Information Technology. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**

**(Dr. Ravindara Bhatt )**

Associate Professor

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

**Submitted by:**

**Parul Sharma**(191206 )

**Kunika Sharma**(191227)

**Ria Mahajan**(191236)

Computer Science & Engineering Department

Jaypee University of Information Technology

**CERTIFICATE**

This is to certify that the work which is being presented in the project report titled “Breast Cancer Detection” in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by “Parul Sharma(191206 ), Kunika Sharma(191227), Ria Mahajan(191236)” during the period from January 2022 to May 2022 under the supervision of Dr. Ravindara Bhatt, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Parul Sharma(191206 )

Kunika Sharma(191227)

Ria Mahajan(191236)

The above statement made is correct to the best of our knowledge.

(Dr. Ravindara Bhatt)

Associate Professor

Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat

**ACKNOWLEDGEMENT**

Firstly, I express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible for us to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor **Dr. Ravindara Bhatt, Associate Professor**, Department of CSE Jaypee University of Information Technology, Waknaghat. The deep Knowledge & keen interest of my supervisor in the fields of “**Machine Learning**” and “**Deep Learning**” encouraged us to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Ravindara Bhatt,** Department of CSE, for his kind help to finish the project.

I would also generously welcome each one of those individuals who have helped straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated our undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Parul Sharma(191206 )

Kunika Sharma(191227)

Ria Mahajan(191236)

**ABSTRACT**

Breast cancer is the world's second leading cause of cancer deaths. It begins when cells in the breast begin to proliferate uncontrollably. These cells typically form tumours, which can be seen on an X-ray or felt as lumps in the breast. Early diagnosis and detection, on the other hand, can increase the likelihood of successful treatment and survival. The key challenge to detection is determining whether tumours are malignant (cancerous) or benign (non-cancerous). A tumour is considered malignant if the cells can spread to other parts of the body or grow into them. In contrast to cancerous tumours, benign tumours do not invade nearby tissue or spread to other parts of the body. However, benign tumours can be dangerous if they press on vital structures like blood vessels or nerves. Despite the fact that successful detection of malignant tumours from histopathological pictures is mostly dependent on radiologists' long-term knowledge, specialists occasionally disagree with their decisions. A computer-aided diagnosis is a second option for image diagnosis that can help experts make more reliable decisions. In clinical applications for identifying malignant tumours from histopathological images, automatic and precision classification for breast cancer histological images are critical. Advanced convolution neural network technology has been widely employed in biomedical image processing and has had considerable success in natural image classification. In this work, we propose an accurate and inclusive computational breast cancer diagnosis by comparing the accuracies of different machine learning and deep learning models on structured data and histopathological microscopy images respectively. The machine learning models like KNN and SVM achieve an accuracy of 96.27% and 93.7% respectively on the Breast Cancer Wisconsin dataset. Furthermore, three deep learning models are proposed and analysed. The simulation results showed that our first custom-based CNN model achieves comparatively lesser accuracy (nearly 70%). The other two models employ transfer learning techniques of the powerful ResNet-50 and VGG-16 Convolutional Neural Networks, pre-trained on ImageNet to train and classify the BreakHis dataset(Histopathological Image Dataset) into benign or malignant. The resultant accuracies were obtained to be 92.17% and 97.96% respectively.

**Chapter 01: INTRODUCTION**

**1.1 Introduction**

Breast cancer is the most prevalent cancer diagnosed in women, accounting for 30% of all new cancer diagnoses in women (excluding skin cancers). Breast tissue samples allow doctors to examine the tissue's microscopic structure and components histologically. Hematoxylin and eosin (H&E) is the primary stain of tissue specimens for routine histopathological diagnoses. It is used to identify between normal tissue, non-malignant (benign) and malignant (carcinomas) lesions, as well as to perform a prognosis evaluation. Breast carcinomas come in a variety of shapes and sizes, each with its own tissue morphology. Breast carcinomas grow from the mammary epithelium and produce ductal carcinoma in situ, which is a premalignant epithelial proliferation inside the ducts. The ability of cancer cells to break past the basal membrane characterises invasive carcinoma.

The pathologist used to undertake morphological evaluation and tumour grading visually, but this technique is time-consuming and subjective, resulting in inter-observer variances even among experienced pathologists. Because the use of morphological criteria in visual categorization is subjective, computer-aided diagnosis (CAD) systems are used to enhance diagnostic accuracy, minimise human error, boost interobserver agreement, and improve repeatability. For digital pathology picture analysis, a variety of approaches have been developed, ranging from rule-based to machine learning applications. Deep learning-based systems have recently been demonstrated to outperform traditional machine learning methods in a variety of image analysis tasks, automating the entire process.

Convolutional neural networks (CNN) have been effectively applied in the medical imaging area for diabetic retinopathy screening, bone disease prognosis and age assessment, and other challenges. Previous deep learning-based applications in histological microscopic image processing have shown promise in assisting in the diagnosis of breast cancer. We provide a method for histology microscope image analysis for breast cancer type categorization in this research. Our method uses deep CNNs for feature extraction and gradient boosted trees for classification, and it outperforms other similar approaches to our knowledge.

In radiology and other medical science fields, machine learning-based algorithms play an important role in diagnosing disease in a much simpler way than ever before, thus providing a viable alternative to surgical biopsy for breast tumours. We attempted to identify and classify breast tumours in this research, comparing the outcomes of binary and multi-class classification of breast tumours with and without Transfer Learning, utilising pre-trained Keras models such as VGG16 and Convolutional Neural Network (CNN) architecture.

**1.2 Objective**

The proposed study's major objective is to:

* Study of various machine learning and deep learning algorithms
* Classify input data (structured data and histopathological breast tumour image data) into benign (non-cancerous) and malignant (cancerous) categories.
* Application of Various Machine Learning & Deep Learning Models for the classification.
* Use of metrics such as accuracy, Confusion Matrix, and others to assess the performance of the applied models and identification of the best model.

**1.3 Motivation**

Maintaining one's health is crucial in today's world. The increased prevalence of health issues has put a considerable strain on doctors throughout the world, making it more difficult to treat each patient successfully. To reduce the strain on doctors, there have been several significant new innovations that have revolutionised the healthcare industry. Because of advancements in computer technology, more accurate and faster results may be obtained, and patients can be treated accordingly. Early identification of cancer remains a challenge in healthcare. According to the World Health Organisation, cancer cells can be detected and destroyed if they are discovered in their early stages. As a result, early detection of cancer cells is critical for reducing cancer-related problems.

Patients with breast cancer are at the very top of the cancer patient list. Because many women are at risk for breast cancer, finding and eliminating cancer cells from the patient's body is crucial. According to the World Health Organisation, breast cancer claimed the lives of over a million women in 2011, and the situation is only getting worse. Breast cancer is the second most common cancer in both women and men throughout the world. It was responsible for around 12% of all new cancer cases in 2012, and 25% of all cancers in women. Breast cancer risk rises when cells in the breast begin to grow out of control. These cells usually develop into a tumour, which can be seen on x-rays or felt as a lump on the skin.

This research aims to use and analyse a variety of Machine Learning and Deep Learning algorithms in order to develop a solution that can assist a patient in determining whether she is at risk for breast cancer at an early stage, allowing the breast cancer cells to be eliminated with appropriate medication. Neurosurgeons and healthcare experts can utilise the system. The method, which combines image processing and pattern analysis, is projected to enhance breast cancer screening performance metrics. The main goal of medical imaging projects is to extract as much useful and accurate information as possible from these pictures with the least amount of inaccuracy. The right combination and parameterization of the phases allow for the construction of a tool that can aid in tumour early detection or surveillance.

**1.4 Language Used**

Python 3 - The main reasons for the use of this programming language for the implementation of the project are stated below:

**Simple and dependable:**

Python provides code that is both concise and readable. Machine learning and AI are based on sophisticated algorithms and flexible workflows, and Python's simplicity allows developers to design dependable solutions. Instead of focusing on the technical subtleties of the language, developers can devote all of their attention to solving an ML problem.

**Extensive selection of libraries and frameworks:**

To enable developers to come up with the greatest coding solutions, it's critical to have a well-structured and well-tested environment.

Python makes a large number of libraries and frameworks available.

These frameworks and libraries are used by programmers to reduce development time. Here are a few examples:

Machine learning frameworks include Keras, TensorFlow, and Scikit-learn.

* NumPy is a Python package for scientific computing and data analysis.
* SciPy is a Python package for advanced computation.
* Pandas is a data analysis tool that can be used for a variety of purposes.
* Seaborn is a data visualisation platform.

**Independency of platform**

Platform independence refers to a programming language or framework that allows developers to create things on one system and then utilise them on another with few (or no) modifications. Python's success stems from the fact that it is a platform-independent language. It is available on a variety of operating systems, including Linux, Windows, and macOS. utilizedPython code may be used to produce standalone executable programs for the majority of mainstream operating systems, allowing Python software to be distributed and utilised without the need for a Python interpreter.

**1.5 Technical Requirements**

**1.5.1 Software Requirements:**

For the reasons described above, Python 3 was utilised to construct the project. The dataset was obtained using Kaggle. In the event of programming syntax mistakes, GitHub and StackOverflow were used as resources.

Google Colaboratory is a high-GPU Jupyter Notebook interface that is open-source. Google Colab is a free Jupyter notebook environment that runs fully in the cloud and requires no setup. Colab allows users to develop and execute code, store and share analysis, and access sophisticated computational resources all from the browser, all for free. Jupyter Notebook is a useful tool for iterating and writing Python data analysis scripts. Instead of creating and rewriting a full programme, lines of code may be written and run one at a time.

**1.5.2 Hardware Requirements:**

* **Processor:** Intel® Core™ i5-10300H
* **Installed memory (RAM):**8.00GB
* **System Type:** 64-bit Operating System

**1.6 Outcomes**

Our study intends to use several machine learning (KNN, SVM) and deep learning (ResNet-50, VGG-16, Custom CNN) classification models to histopathological pictures and statistical data in order to determine the type of tumour the patient has: benign-noncancerous or malignant-cancerous.

The research would assist us in determining the optimal model for this type of classification. Further research might be conducted based on this comparison, and eventually, a fully functional Breast cancer detection system could be built for use in the medical industry.

**Chapter 02: Feasibility Study, Requirements Analysis and Design**

**2.1 Feasibility Study**

**2.1.1 Problem Definition**

Application of various machine learning and deep learning models to classify input data into benign (non-cancerous) and malignant (cancerous) categories given certain input (structured data input/ Histopathological Image of Breast Tissue).

Further, utilise measures like accuracy, Confusion Matrix, and others to evaluate the performance of the applied models and, as a response, visualisations to identify the optimal model.

**2.1.2 Problem Analysis**

An examination of the problem statement was carried out using a literature review and the previously done related works. An extensive literature review aided in a better understanding of past work, approaches and algorithms employed, and performance indicators employed. The main goal of the literature review was to see what was available in the scholarly literature on the subject, and a literature survey of numerous journal articles would help us figure out what models were constructed and what research findings came from them, as well as compare and contrast them. This comparison model aims to provide a better understanding of all the types of machine and deep learning based classification models to help us classify the authenticity of the presence of tumour , as a result, eventually, a fully functional Breast cancer detection system could be built for use in the medical industry.

**2.1.2.1 Related Work:**

**Saad Awadh Alanazii et al.2021 [2]-** For the automatic identification of breast cancer, the study developed a CNN approach that evaluates the IDC tissue areas in WSIs. In this study, three distinct CNN architectures were described, along with a proper comparison. The proposed approach, which employed CNN Model 3, achieved an accuracy of 87 percent. Model 3's five-layer CNN was best suited for this task, despite the fact that it was deeper than Models 1 and 2. A large collection of roughly 275,000 50 50-pixel RGB picture patches guided all structures. On comparing the suggested model to the machine learning (ML) algorithm, it was found that the proposed model outperformed the algorithm by 8%. The proposed model was proven to produce accurate findings, which could eliminate human error in the diagnosis process and lower the cost of a cancer diagnosis. The use of a secondary database like Kaggle is the study's biggest weakness, and future studies should be based on primary data for more accurate breast cancer detection outcomes.

**Mehedi Masud et al.2020 [1]-** This work used transfer learning to observe the classification performance of breast cancer from ultrasound pictures using eight pre-trained CNN models with fine-tuning. The photos were integrated from two separate datasets, and the Adam, RMSprop, and SGDM optimizers were used to evaluate the fine-tuned pre-trained models. The ResNet50 with Adam optimizer had the highest accuracy of 92.4 percent, and VGG16 had the highest AUC 0.97 score. It also suggested a shallow custom model because the pre-trained models had not produced the expected results, and all of the pre-trained models had many convolutional layers and required a long training period. As feature extractors, the proposed custom model used only one convolutional layer. The custom model was 100 percent accurate and had an AUC value of 1.0. In terms of training time, the custom model outperformed all other models and necessitated a small number of trainable parameters. The model was to be validated with other datasets, including new ultrasound pictures, in the future.

**Ahmad LG et al. (2013) [3] -** There were 1189 records in the dataset, 22 predictor variables, and one outcome variable. To create the predictive models, the researchers used machine learning approaches such as Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN). The major purpose of this work was to examine the sensitivity, specificity, and accuracy of these three well-known algorithms on the data. The accuracy of the DT, ANN, and SVM was 0.936, 0.947, and 0.957, respectively, according to the analysis. With the lowest error rate and maximum accuracy, the SVM classification model predicted breast cancer recurrence. The DT model's anticipated accuracy was the lowest of all. The results were obtained by utilising 10-fold cross-validation to assess each model's unbiased prediction accuracy.

**Min et al. ( 2017) [4] -** The goal of this study was to evaluate the efficacy of SVM and SVM ensembles in predicting breast cancer outcomes on small and big size datasets. Training SVM and SVM ensembles were compared in terms of classification accuracy, ROC, F-measure, and computational durations. The experimental results demonstrated that for a small-scale dataset, linear kernel based SVM ensembles based on the bagging method and RBF kernel based SVM ensembles using the boosting method were preferable alternatives, and feature selection must be done during the data pre-processing step. RBF kernel-based SVM ensembles based on boosting outperformed the other classifiers on a large-scale dataset.

**Assegie et al. (2021) [19] -**The k-Nearest Neighbours algorithm (k-NN) was used to classify breast cancer illness in this investigation. Furthermore, k-NN was implemented for various k values, and the resulting classification accuracies were compared. Breast cancer disease was successfully classified using k-NN, according to the study's findings. To evaluate the success of k-NN, the classification accuracies and error values were obtained. The acquired classification accuracy of k-NN was roughly 97 percent, according to the test findings. Furthermore, the findings of the study suggest that k-NN is a good classifier for classifying breast cancer disease.

**Sunil Kumar et al. (2020) [5] -** On the standard benchmark, Wisconsin Diagnostic Breast Cancer (WDBC) database retrieved from the UCI Machine Learning Repository, the performance of the proposed SVM technique had been validated. The WDBC dataset contains 569 records (one record for each patient), 357 of which are classified as benign breast cancer patients and the rest as malignant breast cancer patients. Prior to feature selection, data pre-processing in the form of data normalisation was undertaken in this investigation. 10-fold cross validation was used to produce accurate and impartial classification findings. The dataset was divided into 10 equal sections in a 10-fold CV. The SVM classifier was trained using nine parts (train set) in each fold, and the classification accuracy for the tenth part was calculated (test set) . This procedure was performed ten times, with each portion being classed as a test set. The SVM classifier utilised in this study was a two-class classifier, which classified the selected subset of characteristics as benign or malignant breast cancer tumours. A confusion matrix containing information about the classification system's actual and expected classifications was then constructed. Using 10-fold cross validation, the proposed method reduced the feature dimensionality from 30 to just six, resulting in a classification accuracy of 98.24 percent.

**Subham Sadhukhan et al.(2019) [7] -** The research described a computerised system for detecting cancer in its early stages in a short amount of time. The researchers used machine learning to train a model based on the predicted features of cell nuclei.. The accuracy of each classifier was tested in a comparison study of two separate methods, KNN and SVM. Following that, image processing was used to analyse a digital image of a fine needle aspirate (FNA) of breast tissue in order to determine the nuclei of the cells. The trained model's feature values were then used to determine whether the tumour produced was benign or malignant.

**Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun et al.(2015) [10]** - The paper presented a residual learning framework for making it easier to train networks that are significantly deeper than those previously used. Instead of learning unreferenced functions, the layers were explicitly created as learning residual functions with reference to the layer inputs. The empirical evidence showed that residual networks were easier to optimise and could benefit from significantly increased depth. On the ImageNet dataset, residual nets with up to 152 layers—8 are deeper than VGG nets but have lower complexity. On the ImageNet test set, an ensemble of these residual nets achieves a 3.57 percent error. This result took first place in the classification task at the ILSVRC 2015.

**2.1.3 Solution**

We propose a study of several machine learning and deep learning classification models for detecting whether breast cancer is malignant or benign. This study will aid in the development of an automated breast cancer detection system for use in the medical business, as well as establishing the optimal model for such classification.

**2.2 UNIFIED MODELLING LANGUAGE (UML) DIAGRAM :**

|  |
| --- |
| **Figure 1 : Unified Modelling Language Diagram**  *\*Flow chart drawn using draw.io [https://app.diagrams.net/]* |

**Chapter 03: IMPLEMENTATION**

**3.1 Date Set Used in the Minor Project**

The following two datasets have been used in the project**:**

* **Wisconsin Breast Cancer Dataset**

**[**[**https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data**](https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data)**]**

* **Breast Histopathology Dataset**

**[**[**https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images**](https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images)**]**

**3.2 Date Set Features**

**3.2.1 Types of Data Set**

Two types of datasets were used for the proposed comparison model, one of them included the structured .csv data and the other included the histopathological images of the breast tissue. The dataset was then divided into Training and testing sections for the necessary validation. Further detailed description of the dataset for the binary classification has been provided above as well.

**3.2.2 Number of Attributes, fields, description of the data set**

The number of classes in this dataset are 2, for the presence of cancerous and non-cancerous tissue in the breast. Malignant or carcinoma and benign or non cancerous. The major two types of breast cancers :

Ductal or lobular carcinoma is a type of cancer that affects the intestines.

The majority of breast cancers are carcinomas, which are tumours that begin in the epithelial cells that line the organs and tissues in the body. Adenocarcinoma, which begins in cells in the ducts (milk ducts) or the lobules, is the most common type of carcinoma that forms in the breast (glands in the breast that make milk).

### In situ vs. invasive breast cancers : The type of breast cancer can also refer to whether the cancer has spread or not. In situ breast cancer (ductal carcinoma in situ or DCIS) is a pre-cancer that starts in a milk duct and has not grown into the rest of the breast tissue. The term invasive (or infiltrating) breast canceris used to describe any type of breast cancer that has spread (invaded) into the surrounding breast tissue.

**3.2.2.1 Wisconsin Breast Cancer Dataset Description**

A digitised image of a fine needle aspirate (FNA) of a breast mass is used to compute features. They characterise the properties of the cell nuclei that appear in the photograph.

The 3-dimensional space is described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Attribute Information:

1) ID number

2) Diagnosis (M = malignant, B = benign)

3) Ten real-valued features are computed for each cell nucleus:

* a) radius (mean of distances from centre to points on the perimeter)
* b) texture (standard deviation of grey-scale values)
* c) perimeter
* d) area
* e) smoothness (local variation in radius lengths)
* f) compactness (perimeter^2 / area - 1.0)
* g) concavity (severity of concave portions of the contour)
* h) concave points (number of concave portions of the contour)
* i) symmetry
* j) fractal dimension ("coastline approximation" - 1)

For each image, the mean, standard error, and "worst" or worst (mean of the three largest values) features were computed, yielding 30 features.

For example, field 3 represents Mean Radius, field 13 represents Radius SE, and field 23 represents Worst Radius.

With four significant digits, all feature values are captured.

Attribute values that are missing: none

357 benign tumours, 212 malignant tumours

**3.2.2.2 Breast Histopathology Dataset Description**

* 162 whole mount slide pictures of Breast Cancer (BCa) specimens were scanned at 40x in the original dataset.
* 277,524 50 x 50 patches were retrieved from that (198,738 IDC negative and 78,786 IDC positive).
* The images were in png format.
* example uxXyYclassC.png uxXyYclassC.png
* where u is the patient ID (10253idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C is the class, with 0 being non-IDC and 1 representing IDC.

The image array of the Histopathological dataset has 2 categories : Benign(0) and Malignant(1).   
The shape of the given image was found out to be of the dimension 50X50 with 3 channels indicating the RGB coloured image.

Since the Image dimensions were large and would hence result in less training rate of the model, I standardised the image using the pixel range. The pixel range ranges from 0 - 255.

Therefore , we normalised our data in the range of 0 - 1 by dividing the obtained pixel dimension of the image with maximum pixel range i.e 255.  
The obtained Min and Max Normalised pixel range obtained was 0.188 and 0.957.  
Next I converted the RGB coloured image to the grayscale since it is easier to process the grayscale images.

The images were standardised to obtain the bell shaped Gaussian Curve with mean = 0 and Standard Deviation = 1.

|  |
| --- |
| **Figure 2: Breast Histopathological Images** |

**3.3 Design of Problem Statement**

The severity of the situation was taken into consideration when creating the problem statement. Breast tumours are deemed fatal because they can push on healthy breast tissue or spread to other parts of the body. Some breast tumours are cancerous or may become cancerous in the future. If they restrict the flow of fluid surrounding the breast tissue, resulting in a rise in pressure inside the breast region, they can create difficulties. As a result, early detection and categorization are critical, and models like these, which are constructed using various algorithms, assist in precisely determining them.

|  |
| --- |
|  |
| **Figure 3 : Design of Problem Statement**  *\*Flow chart drawn using draw.io [https://app.diagrams.net/]* |

**3.4 Pseudo code of the Project Problem**

**3.4.1 Machine Learning:**

**3.4.1.1 KNN:**

1. The scikit-learn package is used to import the k-nearest neighbour algorithm.
2. Construct a feature and a target variable.
3. Split the data into training and testing sets.
4. Create a k-NN model with the neighbours value.
5. The data is trained or fitted into the model.
6. Predict the result.

#### Using the scoring method, determine the model's accuracy

**3.4.1.2 SVM:**

1. The scikit-learn package is used to import the SVC (Support vector Classifier)algorithm.
2. Construct a feature and a target variable.
3. Split the data into training and testing sets.

#### Create a Linear SVC object.

#### Using the training data,train the linear SVC classifier.

#### Predict the result.

#### Using the scoring method,check the accuracy of the model.

**3.4.2 Deep Learning:**

**3.4.2.1 Pseudo code 1 Pre Processing:**

1. Create Function for training the image dataset.
2. Divide the dataset into desired target categories.
3. Create an array of images and read images.
4. Create a new array and place the images after resizing.
5. Now new training data consists of an array of resized images and target classes.

**3.4.2.2 Pseudo code 2 Model Building :**

1. Initialise model as sequential()
2. Add input layer
3. Add convolution layer, set filters(64 or 128 or 256), kernel size(3x3 or 5x5),etc
4. Add max pooling layer, set pool size(2x2) and furthermore layers.
5. Add dropout layer, set the dropout percent
6. Repeat steps 3-5 as required
7. Flatten the output of these layers
8. Add dense and dropout layers
9. Get classification as output

**3.5 Flow graph of the Minor Project Problem**

|  |
| --- |
|  |
|  |
| **Figure 4: General Working Of CNN in Image Classification**  **Image Source:** [**http://surl.li/bxlgd**](http://surl.li/bxlgd) |

A convolution is depicted in the image above. To obtain the convolved feature, we apply a filter/kernel(3 matrix) to the input picture. The next layer receives this convolved feature. A convolution is depicted in the image above. To obtain the convolved feature, we apply a filter/kernel(3 matrix) to the input picture. The next layer receives this convolved feature.

|  |  |
| --- | --- |
| **Figure 5(a): VGG-16 Architecture** | **Figure 5(b) : ResNet-50 Architecture** |

**The above architectures depict that :**

The VGG architecture is divided into blocks, each of which is made up of 2D Convolution and Max Pooling layers.

ResNet or Residual Network uses Skip Connections which allow the vanishing gradient problem to be overcomed.

**3.6 Screenshots of the various stages of Project**

**3.6.1 Machine Learning:**

|  |
| --- |
| **Figure 6: Importing all the necessary libraries and dataset** |

|  |
| --- |
| **Figure 7: Dropping Null Values** |

|  |
| --- |
| **Figure 8: Train Test split** |

**3.6.1.1 Support Vector Machine (SVM)**

|  |
| --- |
| **Figure 9: SVM model** |

**3.6.1.2 K-Nearest Neighbour (KNN)**

|  |
| --- |
| **Figure 10 : KNN model** |

**3.6.2 Deep Learning:**

|  |
| --- |
| **Figure 11: Importing all the necessary libraries** |

|  |
| --- |
| **Figure 12 : Importing the dataset from Kaggle using API** |

|  |
| --- |
| **Figure 13: Unzipping dataset** |

**3.6.2.1 Custom CNN Model**

|  |
| --- |
| **Figure 14: Displaying a picture from the dataset** |

|  |
| --- |
| **Figure 15 : Resizing the previously displayed image** |

|  |
| --- |
| **Figure 16 : Training the dataset and converting the image from RGB to Grayscale** |

|  |
| --- |
| **Figure 17 : Creating the CNN model framework** |

|  |
| --- |
| **Figure 18: Setting the epoch and batch size** |

|  |
| --- |
|  |
| **Figure 19: Training the model** |

|  |
| --- |
| **Figure 20: Validation on Performance metric ACCURACY** |

|  |
| --- |
| **Figure 21: Saving the model** |

**3.6.2.2 ResNet-50 Pre-Trained Model**

|  |
| --- |
| **Figure 22: Data Segmentation** |

|  |
| --- |
| **Figure 23: Train-Test Split** |

|  |
| --- |
| **Figure 24: Data Augmentation** |

|  |
| --- |
| **Figure 25: CNN(Resnet50) model** |

|  |
| --- |
|  |
|  |
| **Figure 26: Training the model** |

|  |
| --- |
|  |
| **Figure 27: Saving and Predicting result** |

**3.6.2.3 VGG-16 Pre-Trained Model**

|  |
| --- |
|  |
|  |
|  |
| **Figure 28: Data Segmentation** |

|  |
| --- |
| **Figure 29: Train-Test Split** |

|  |
| --- |
| **Figure 30: CNN(VGG-16) model** |

|  |
| --- |
|  |
| **Figure 31: Training the model** |

|  |
| --- |
| **Figure 32: Saving Model** |

**Chapter 04: RESULTS**

**4.1 Discussion on the Results Achieved**

|  |  |
| --- | --- |
| **Figure 33 (a)-33(b) Accuracy Vs Loss curve for Custom CNN model** | |

|  |  |
| --- | --- |
| **Figure 34(a)-34(b) Accuracy Vs Loss curve for Resnet50 CNN model** | |

|  | |
| --- | --- |
| **Figure 35: Training and validation Accuracy and Loss curve for Resnet50 CNN model** | |

|  |
| --- |
| **Figure 36: Accuracy and Loss curve for VGG-16 CNN model** |

|  |
| --- |
| **Figure 37: Training and Testing Accuracy & Loss for VGG-16 CNN model** |

The plots for accuracy metrics comparing the performance of training accuracy and validation accuracy obtained during the training process are given in above figures. Both accuracy curves are steadily increasing according to the plot, with a faster ceiling level obtained for training accuracy. In addition, it shows plots for the loss function comparing the behaviour of training loss and testing loss obtained during the training process. Both losses are systematically decreasing.

| **Model** | **Dataset** | **Accuracy Obtained** |
| --- | --- | --- |
| KNN | Wisconsin Dataset | 96.27% |
| SVM | Wisconsin Dataset | 93.7% |
| Custom CNN Model | Histopathological Image Dataset | 69.71% |
| ResNest-15 Pre-Trained Model | Histopathological Image Dataset | 92.17% |
| VGG-16 Pre-Trained Model | Histopathological Image Dataset | 97.96% |

|  |
| --- |
| **Figure 38: Accuracy Plot for the different Models** |

In machine learning, KNN performs better than SVM on the Wisconsin Dataset.

While in deep learning models, the accuracy metric shows that VGG16 outperformed Custom CNN model and Resnet 50 model. The training and validation accuracy were maximum for the VGG16 .The possible reason for its accuracy could be the higher number of the parameters that makes it effective to overfit to the dataset faster. Eventually, we also encountered early stopping which was countered by the callback function..More parameters makes it a better feature extractor. All of these results clearly demonstrate VGG16's classification capabilities and its robust performance on our dataset.

**4.2 Result validation**

Based on the above observations and metrics performed on various machine learning and deep learning models on both structured and image datasets, we finally concluded that our datasets (Wisconsin dataset & Histopathological Image dataset) performed well on KNN and VGG16 pre-trained model techniques respectively.

To prove our validation for the same, two methods could be adopted :

**MANUAL VALIDATION :**

In this, we could provide our predicted results to the domain expert mainly from the healthcare domain to validate the actual and predicted output results.

**MACHINE ORIENTED PREDICTION :**

In this, we can provide the predicted results to the machine, which would further predict the actual and predicted results based on the performance metrics , for instance, confusion matrix or accuracy to get the authentication on the achieved results.  
  
Based on these observations, we can then provide a better solution to the healthcare department to get a reliable solution in detecting breast cancer.

**4.3 Application of the Minor Project**

The application of computer-based learning models have emerged as a key area of cancer research. Several researchers have focused in recent years on developing systems, both hybrid and fully automatic systems, that could aid in the diagnosis, prognosis, and prediction of breast cancer outcomes by utilising Statistics and Artificial Intelligence. The development of these systems necessitates a variety of techniques, the most common of which are machine learning (ML) algorithms. Several scientific studies have published algorithms and nomograms for predicting the pathologic stage of patients with clinically localised cancer or Gleason score upgrading. Specifically, ML enables the integration or combination of various layers of data, such as those from medical images, laboratory results, clinical outcomes, biomarkers, and biological features, for improved patient prognosis and stratification toward personalised medicine. Deep learning approaches for extracting characteristics and improving the efficiency of medical image analysis have recently been developed. Deep learning is a type of machine learning that employs multilayer convolutional neural networks (CNN). Unlike other feature extraction techniques, CNN can extract image features directly from the dataset. Using convolution, this type of feature extraction extracts features from different parts of an image. Scientists have achieved promising results with CNNs for the diagnosis of breast cancer in recent years.

**4.4 Limitation of the Minor Project**

Our proposed model comparison approach is tested on a dataset with a small number of images, the performance of our proposed network could be further evaluated using a larger dataset. Furthermore, real-world data often differs from publicly available datasets.The main limitation of this project is that it relied on a secondary database such as Kaggle, and future studies should rely on primary data to improve the accuracy of the results related to breast cancer identification. If we could further experiment with real-world data, we could see how the model performs on it. According to the confusion matrix,some images were misclassified,even after employing a variety of image processing techniques, the model still struggles to distinguish cancer cells from dense breast tissue. However, the model performs admirably in the majority of the test images, correctly classifying the cases. Despite its minor issues, the model is robust.

There are a few small drawbacks to the project:

* It has a high computational complexity due to its several layers, and it also has a much reduced computational complexity due to operations such as maxpool.
* If the machine on which the model is being performed does not have a decent GPU, it will take a long time.
* One of the drawbacks of histopathological diagnostic procedures is that they are time-consuming and prone to sample errors.

|  |
| --- |
| **Figure 39: Confusion Matrix Showing Misclassifications In VGG-16 Model** |

**4.5 Future Work**

In future, we will be working on improving accuracies of the models and comparing them on the basis of other performance metrics. We will evaluate models on the basis of different optimizers and learning rates. Furthermore,we will be analysing other models like VGG-19 and DenseNet to bring out the best model for the detection of breast cancer. Once we come up with the most optimal model, we would try to develop an appropriate User Interface for the deployment of the model into a full-fledged Breast Cancer Detection System to be hosted on a server and be put into actual use .

**References**

[1] Masud, M., Eldin Rashed, A. E., & Hossain, M. S. (2020). Convolutional neural network-based models for diagnosis of breast cancer. Neural Computing and Applications. doi:10.1007/s00521-020-05394-5.

[2] Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting breast cancer detection using convolutional neural network. Journal of Healthcare Engineering, 2021.

[3] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), 3.

[4] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. PloS one, 12(1), e0161501.

[5] Agarap, A. F. M. (2018, February). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd international conference on machine learning and soft computing (pp. 5-9).

[6] Dey, S. CNN application on structured data-Automated Feature Extraction. URL: https://towardsdatascience. com/cnn application-on-structured-data-automated-feature extraction-8f2cd28d9a7e.(accessed: 20.05. 2019).

[7] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[8] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9(4), 611-629.

[9] Gandhi, R. (2018). Support vector machine—introduction to machine learning algorithms. Towards Data Science, 7.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun(2015)-Deep Residual Learning for Image Recognition.

[11] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321, 321-331.

[12] Dabiri, S., & Heaslip, K. (2018). Inferring transportation modes from GPS trajectories using a convolutional neural network. Transportation research part C: emerging technologies, 86, 360-371.

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[14] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[15] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), 2402-2410.

[16] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... & CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210.

[17] Christ, P. F., Elshaer, M. E. A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., ... & Menze, B. H. (2016, October). Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In International conference on medical image computing and computer-assisted intervention (pp. 415-423). Springer, Cham.

[18] Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology, 195(1), 215-243.

[19] Assegie, T. A. (2021). An optimized K-Nearest Neighbor based breast cancer detection. Journal of Robotics and Control (JRC), 2(3), 115-118.