BREAST CANCER DETECTION USING MACHINE LEARNING

BY

Momenunnessa Meem
ID: 191-15-2634

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Mahfujur Rahman
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project/internship titled **"Breast Cancer Detection Using Machine Learning"**, submitted by Momenunnessa Meem, ID No:191-15-2634 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *04-02-2023*.

BOARD OF EXAMINERS

5/2/23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. S. M. Aminul Haque
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

4/2/2023

External Examiner

Dr. Shamim H Ripon
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Dr. S.M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S.M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Mahfujur Rahman
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Momenunnessa Meem

Momenunnessa Meem
ID: 191-15-2634
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to Dr. S.M. Aminul Haque, Associate Professor Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Data science” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan, and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

Breast tissues swell and grow out of control, generating a lump or fine layer as just cancer. This is how breast cancer arises. The second most widespread cancer in girls, after melanoma, is breast cancer. Women above Fifty are most likely to experience this. [1] However, it can affect anyone at any age. Reducing health risks and boosting prevention strategies may help prevent cancer. Cancer prevention human trials are used to explore strategies for preventing cancer, and novel approaches to avoiding breast cancer are indeed being investigated in clinical studies. In this article, i used machine learning algorithms to determine whether cancer is benign or malignant. If the condition is benign, the doctor can begin therapy right once. For this project, the data is collected by Kaggle. There are four machine learning algorithms used for analysis. They are decision trees, logistic regression, bagging, and random forest. Here, i find out the best algorithm with accuracy and precision, f1-score, recall, and confusion matrix.PCA and correlation matrices are utilized for data visualization. The random forest does have the best performance (98%) for the datasets that was provided and the bagging algorithm gives the best AUC value (99%).

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction:

Breast cancer is a disease where the cells in the breast multiply uncontrollably. There are various different types of breast cancer. The type of breast cancer depends on which breast cells turn malignant. Breast cancer can spread beyond the chest through the lymphatic and circulatory systems. When breast disease develops in certain other anatomical areas, it is considered to have disseminated. Men too can contract breast cancer, albeit it's rare. In the United States men, melanoma affects roughly twenty-six thousand men a year, making up less than one percent of all cases.[3] Quasi-white women have a slightly higher chance of developing this disease than women of any other color or ethnicity. Quasi White women and quasi-Black women both had roughly the same odds of acquiring the disease

1.2 Motivation:

1. I Apply machine learning algorithms to analyze cancer patient stage, such as benign or malignant.
2. Most importantly, after the breast cancer type has been established (recognized), considerably more specific queries may be done, identifying the supplied cancer kind as well as any other breast cancer-related questions.

1.3 Rationale of the study:

1. To provide for a more alleviated property perspective on the datasets utilized.
2. To find a more accurate predictive machine-learning system with a lower FN count.

1.4 Research Question:

1. Does it predict an actual output by given sample data with the system?

Yes it predict actual output what I want for my analysis.

2. Can it classify if it initial stage breast cancer or not by using a machine learning algorithm?

Yes

3. Does every algorithm work perfectly (yes/no)?

Yes.

1.5 Research Outcome:

1. I identify the most accurate and effective algorithm.

2.Comparison with previous research work.

3. Applying PCA for better visualization.

1.6 Project Management and Finance

For analysis breast cancer I used data set from an online platform then I used weka machine learning tool to analysis my datasets then I find my necessary result. This is completed by my personal money.

1.7 Report Layout

In the chapter 1 Introduction of my report written.

In chapter 2 I added Background study

In chapter 3 I keep Research Methodology

In chapter 4 I keep Experimental Result Analysis

In chapter 5 Social Impact is discuss

After that conclusion is discused in chapter 6.

1.5 Types of breast cancer:

1. cancer known as intrusive (trying to infiltrate) duct adenocarcinoma starts in the breast's milk-producing glands, breaks through the channel membrane and then spreads to surrounding breasts. With over 80% of cases being this type of cancer, it is the most common.

2. lobular cancer that is aggressive: This cancer started in the fascicles of the breast, where breastfeeding is generated, and it has subsequently spread to surrounding breasts. It leads to ten to fifteen percent of breast cancer cases.

3. Acinar cancer with Vivo, commonly known as diagnostic phase cancer, is generally thought to either be pituitary tumors or malignant because cancer has not spread beyond the dairy glands. Amazingly, this condition is treatable. Treatment must begin very away to prevent the illness from spreading quickly and attacking other systems.

4. Triple-negative breast cancer (TNBC): One of the most difficult breast cancers to treat is triple-negative breast cancer, which makes up 15 percent of the entire of instances overall. It is referred to as triple-negative breast cancer because it lacks Three of the markers connected to other types of cancer. This makes the treatment process difficult.

5. Intra Vivo clavicular melanoma: Papillary cancer, a precancerous condition, is the term for mammary granulomas having cancerous results. Even if it isn't true cancer, this symptom could indicate a future breast cancer possibility.

6. Acute inflammatory cancer: That form of cancer seems rare, severe, and infectious-looking. Breast skin pitting, dimpling, and swelling are common side effects of metastatic breast cancer. Blocked cancerous cells in the lymph system beneath their skin create it.

7. Breast Paget's cancer: This cancer affects both the tissue of your nipple and the areola.

1.6 Symptoms :

Tissue in the chest that appears larger and sloppier than just the remainder of something like the breast.

1.alterations to the breast's skin, including slight discoloration.

2.A breast's shape, size, and appearance changing

3.a recently flipped nip

4.There could be redness and orange-like spots across each breast.

5.Scales, flaking, scabbing, and peeling of the chest and areola's basal pigmentation.

6.Nipple and breast discomfort

7.Blunt discharge.

Self-examinations for breast cancer are crucial, but if you are over 40, you should also get yearly mammograms, at least until you turn 70. Women over 70 years old can choose to get mammograms every two years.[2]

1.7 Causes of breast cancer:

Research found that a number of variables could increase the chances of getting breast cancer. That included

Age: The risk of breast cancer rises if you are older than 55.

Genetics and Family history: DNA screening can be used to identify specific faulty genes that are handed down from parents to kids and are responsible for 5 percentage points to 10 percentage points of breast cancers.

Sex: Breast cancer is far more common in women than in males.

Exposure to radiation: People are more likely to get breast cancer whether they've previously undergone radiotherapy, especially to the neck, chest, or head.

Smoking and alcohol: Alcohol use can increase the likelihood of several forms of breast cancer, and tobacco use has been related to numerous cancers, especially breast cancer.

Replacement hormone: Replacement hormone treatment Utilizers of hormone replacement treatment (HRT) are more likely to get breast cancer.

1.8 Stage of breast cancer :

Stage 0: The illness has no aggressive features. This indicates that it has not emerged from your breast tissues.

Stage 1:The adjacent breast tissue has been infected with cancerous cells.

Stage 2: Either the tumor is less than two cm in size but has weakened immune systems under the arms, or it is more than five cm in size and it hasn't. Tumors varied in diameter between Two to five cm at this stage and may or may not affect the regional lymph glands.

Stage 3: At this point, the cancer has spread beyond its previous website. Although it could have spread to nearby lymph nodes or muscles, it has not yet affected other areas of the body. Breast cancer at stage 3 is typically referred to as regionally progressed.

Stage 4:The liver, bones, brain, or lungs are among the organs where cancer has metastasized outside of your breast. Metastatic breast cancer is another name for end-stage breast cancer.

1.9 Prevent the occurance of breast cancer:

Check with your doctor about mammograms. With the doctor, you should decide when to begin radiography as well as many other mammography examinations & procedures, like physical mammograms.

To increase awareness of cancer, undertake a self-reflection and self to learn further about breasts. Women may find it helpful to periodically inspect their breasts throughout a self-reflection and self in as to get more familiar with them. If you discover any new abnormalities, malignancies, and other odd signs in your breasts, ask your healthcare provider straight once.

Attempt to squeeze in at least 30 minutes of activity many days of each week. If you have not exercised in a while, ask your doctor for permission to start again. Begin slowly.

When you already have a good lifestyle, strive to keep it. If necessary, seek guidance from your physician regarding the best ways to lose weight. Per day, you must consume less food while slowly increasing your exercise.

Extra-virgin olive oil and mixed nuts can be added to a Mediterranean diet by women to lower their risk of breast cancer. Fruits and whole grains, vegetables, nuts, and legumes are the mainstays of the Mediterranean diet. People who consume a Mediterranean-style diet prefer fish over red meat and healthy fats like olive oil to butter.

1.10 How is breast cancer diagnosed :

Doctors frequently use extra tests to locate or diagnose breast cancer. An expert or breast doctor may be recommended for women.

Breast ultrasound imaging is a technique that uses sound waves to create sonogram images of specific breast areas.

If you suffer from a breast problem, such as a malignancy, or if you see portion of your breast looks strange on a diagnostic mammogram, your doctor might advise definitive radiography. That specific chest X-ray is much more detailed.

Biopsy. In this test, material and liquid first from mammary is removed and put through additional testing after being inspected under a lens. The many sorts of biopsy are numerous.

A technique for scanning the body using a magnet and a computer is magnetic resonance imaging (MRI) for the breast. The MRI scan will provide a detailed image of areas inside the breast.

Thanks to tremendous expenditure in aimed at raising awareness initiatives, breast cancer diagnosis and screening have advanced. Survival rates for breast cancer have increased and the death rate associated with the disease is quickly declining thanks to earlier discovery, a revolutionary tailored approach to therapy, and better understanding of the disease. This project tries to predict breast cancer early so that doctors can start their treatment properly and early and also understand patients' previous history.

CHAPTER 2

Background

2.1 Related Work

Asri et al. applied data mining techniques in their research paper. Using the primary information for Wisconsin Breast Cancer, they used several ML algorithms: Support Vector Machine (SVM), Decision Tree, k-NN, and Naive Bayes .Their objective is to evaluate these techniques' accuracy, efficacy, or efficiency, sensitivity, specificity, and precision. WEKA data tools are employed in the experiment. SVM produces the best precision and error rate results, and SVM provides the maximum accuracy of 97.13 percent.[4]

Rawal detects and categorizes normal and cancerous patients, with the goal of optimizing our classification systems to attain high accuracy. Four breast cancer detection algorithms are evaluated in the study, Logistic Regression, SVM, KNN, and Random Forest, using distinct datasets. In JUPITER, A predictive model called the 10-fold cross-validation test is being used to assess and analyze data in terms of effectiveness and efficiency. Breast Cancer Dataset from the UCI Wisconsin Machine Learning Repository drew attention since it included a high number of cases with multivariate variables. SVM has the largest percentage of properly classified examples (97.13 percent) and a smaller percentage of wrongly correctly classified than the other classifications. [5]

Singh develops the breast cancer prediction model used different of ML classification methods and they are SVM, k Nearest Neighbor (kNN), Gaussian Naive Bayes (NB) and Logistic Regression (LR). The Wisconsin breast cancer dataset from the UCI Machine Learning Repository was used to collect the data. Regarding the given data, k Nearest Neighbor has the highest accuracy. The small number of samples utilized for training and testing is a drawback of this investigation. [6]

Using machine learning approaches, Ganggayah et al. created models for discovering and displaying pertinent predictive markers of breast cancer survival rates. To identify the key prognostic factors of breast cancer survival rate, prediction models were developed using random forests, decision trees, extreme boost neural networks, support vector machines and logistic regression.All algorithms produced comparable results in terms of predictive performance and calibrating measure, with decision tree giving the lowest accuracy (purity - 79.8%) and the random forest producing the highest (accuracy = 82.7%). Finally, decision trees were created and validated using survival analysis.[7]

Six supervised machine learning techniques are provided by Gupta et al., including support vector machines with radial base function cores, decision trees, random forests, logistic regression, and k-NN. In addition to machine learning, an ideal deep learning approach has been used for classification. The data utilized in the current investigations to carry out the tests comes from the WBCD, which was previously been categorized as malignant and benign. Adam Gradient Learning achieves maximum accuracy since it combines the advantages of AdaGrad and RMSProp.[8]

Rane et al. compared 6 ML algorithms: NB, Artificial Neural Networks (ANN), Random Forest (RT), Support Vector Machine (SVM), Nearest Neighbour (KNN), and Decision Tree . The dataset utilized was obtained from the WDBC dataset, that {: gap {:kind:userinput}} retrieved from a digitized MRI picture. These initiatives assist real-life patients and clinicians gather as much information as possible. Using machine learning approaches, we would be capable of categorizing and predict whether cancer is benign or cancerous.[9]

Prat et al looked at the newest current statistics on breast cancer fundamental classification, with such a focus on the Trust and trustworthiness subgroup. We anticipate merging these fundamental types with 4 basic clinical therapy groups, such as HR/HER2, HR/HER2, HR/HER2, and quintuple, to improve patient outcomes. Additionally, they highlight how combining commonly diagnosed indicators with the data offered by these genetic domains might aid in better understanding the disease's molecular intricacy. They examine the CSC theory and the productivity - boosting ancestry of each fundamental category before wrapping up.[10]

Recent research by Weigelt et al. concentrating on distinct semi of carcinoma cells from various locales have revealed neuropathies alterations and particular fusing proteins. For example, adenoma cancers continuously exhibit this same t(6;9) MYBeNFIB migration, Breast secretory tumors are always carrying a t(12;15) mutation, which results in the formation of the ETV6eNTRK3 fusion protein, and papillary cancers continuously exhibit the inhibitory effect of the Gene encoding via a range of molecular processes. Reviewing the relationships between breast cancer's molecular taxonomy and its histology special kinds, they also address potential causes of the disease's diversity.[11]

In a study that Ruiz et al. studied, specialists were able to diagnose cancer at mammography with greater accuracy when assisted by an artificially intelligent system. ScreenPoint Medical provided financial support for the research. They used digital mammograms that were retrospectively acquired and anonymised from diagnostic exams for their investigation. According to the Health Insurance Accountability and Portability Legislation, this retrospective study was legal. The Artificial system's AUC alone was comparable to the typical radiologists' AUC (0.89 vs 0.87).[12]

In order to identify breast cancer, Sadoughi et al. looked at a number of AI techniques that use image processing. To illustrate various strategies and their outcomes over the last few years, the findings were supplied in tables. 18,651 papers were taken for this analysis from 2007 to 2017. After eliminating those that employed similar methods and presented comparable findings, 40 publications were then looked at. The findings of a study that employed a suitable segmentation technique to extract the target region in the image showed the maximum accuracy in the SVM approach. The best accuracy was achieved by combining morphological features, Pratio features,matrix of gray-level co-occurrence, and features.[13]

The research by Sechopoulos et al. that assessed the existing abilities of AI will be discussed, along with suggestions regarding how these abilities might be used in the healthcare setting and the problems that need to be resolved before this vision is becoming a reality.Provided this same limitations of the present hold achievement evaluations of such methodologies, which can only be assessed during massive vetting tests, it is still unclear how well these new algorithms' performance will be measured to that of breastscreen radiology in the real screening world.[14]

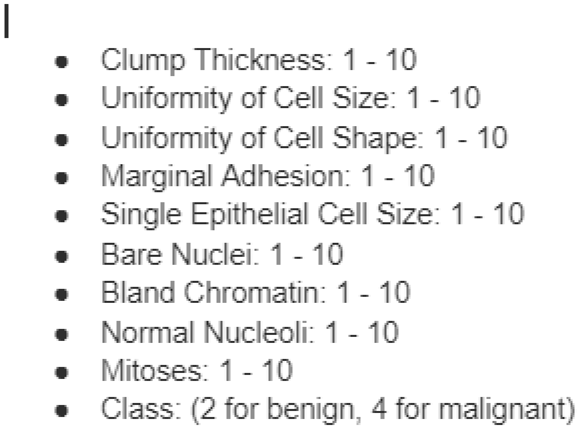
Huang et al. evaluated several texturing techniques and provided the most popular genetic algorithm: The three primary kinds of texture extraction techniques are architectural, analytical, and spectrum.[15]

CHAPTER 3

Research Methodology

3.1 Explanation of the data:

The information was obtained through kaggle.. There are 683 data and 10 attributes such as



- Clump Thickness: 1 - 10
- Uniformity of Cell Size: 1 - 10
- Uniformity of Cell Shape: 1 - 10
- Marginal Adhesion: 1 - 10
- Single Epithelial Cell Size: 1 - 10
- Bare Nuclei: 1 - 10
- Bland Chromatin: 1 - 10
- Normal Nucleoli: 1 - 10
- Mitoses: 1 - 10
- Class: (2 for benign, 4 for malignant)

Figure 1: Attributes name

3.2 Algorithm description:

3.2.1. Logistic regression:

This model is frequently in use in stats to simulate the probability of a particular class or occurrence happening, like the chance that a squad will succeed, that a patient will be in excellent health, etc. This may be expanded to represent a number of other situations, such as determining whether an image is of a creature such as a dog,cat, lion, or the other. It was anticipated that there would be one major detected object in the picture, with each object's significance ranging from zero to one. The logistic model's log-odds for the value designated "1" are a linear synthesizing of one or maybe more connections between the predictions; the two parameters can each be a binary classifier issue or any actual value. The logistic function, thus the title, translates file to likelihood; the average diameter of the value labelled "1" might fluctuate between 0 and 1, therefore the labeling. The standard unit of measurement for the logarithmic scale is the logit, from the logistic unit; consequently, the relatively distinct. The distinguishing characteristic of the logistic regression model is that increasing one of the individual variables computer resources scales the likelihood of the specific result at a constant speed, with each predictor variable having its own parameter; for binary predictors, this extrapolates the hazard ratio. Equivalent models, such as the probit model, can be used in place of the likelihood of stagnation. A binary logistic regression model with two levels represents the dependent variable.When there are more than two results in an output, multinomial logistic regression is utilized to model it. In the event that the numerous categories are ordered, ordinal logistic regression is utilized.

3.2.2 Random Forest:

A call trees and perhaps a fabric classification has hyper parameters that are akin to a random forest. The model's unpredictability is increased by developing these plants in a Random Forest. While ripping a node, it looks for the most basic characteristic out of a random selection of possibilities rather than the most important one. Operation of the Random Forest algorithm

The following stages can help us understand how the Random Forest algorithmic program functions:

The first step is to choose random samples from a particular dataset.

One pair of steps After that, each sample's wire tree may be created by this algorithmic software. Following that, each call tree will experience the prediction impact.

Options are carried out for each predicted outcome in step three.

In the last step, select the predicted result that received the most votes as the outcome.

When applying the Random Forest method for regress situations, the mean square error is used to identify how your data forks out from every node.

$$MSE=1ni=1N(xi-yi)2$$

3.2.3 Bagging:

The Bootstrapping integration is a ML technique that combines a mechanism intended to increase the legality and reliability of algorithms employed during processing and retrieval, as well as what is thought to just be content. By adding, variability is decreased and overdosing is prevented. Bulk predictions are frequently made using bootstrap predictive models, which may be used to include a variety of predictive models.The subdivision or retrospective rule was implemented to every random subset, and a new forecasting measuring device forecasts from viewers of each foundation in the setting of hindsight.This resource might be a simple general rule for all duplicates of the bootstrapping measurement device of a main training session carried out. Once the issue of segregation has been handled, the lowest student forecast grading system includes a bully vote in mass or by evaluating any open division possibilities.X is a guessable record, fbag is the bagged forecast, and f1(X), f2(X),..., fb(X) are forecasts from users of each basis. It's going to be connected to Calculation.

$$fbag= f1(X)+f2(X)+.....+fb(X)$$

Because of the aggregation strategy, bagging efficiently lowers the variation of the a personal learning algorithm (— in other words, average reduces variability); yet, bagging does not constantly enhance a private base learner. Bagging is extremely beneficial with volatile, multi-variable trainees that prediction accuracy varies dramatically in response to minute adjustments in coaching input. This includes the call tree and K nearest algorithms. Sacking, on the other hand, produces less rise in anticipated outcomes for systems with high unit stability or bias because there is less fluctuation.

3.2.4 Decision tree:

In a decision tree, the leaf nodes represent results or class labels, while the non-leaf or decoration nodes represent decisions. One or more designations are examined by each internal node, which results in one or more connections or branches. This connection has a decision value assigned by the government. The aforementioned connections are both both exclusive and inclusive. This indicates that you will be shielded from every scenario if you merely click on one of the links. The best tools for weighing several possibilities are decision trees. They offer a very useful framework for outlining alternatives and evaluating potential outcomes for each optionEach node of a binary decision tree indicates the comparison to be made or the alternative to be chosen. There are two edges entering into and leaving the nodes. The outcome "yes" or "true" is represented by one edge, and the outcome "no" or "false" is represented by the other edge. The lettersA,B,C, and d appear to be printed on four coins, three of which are equal in strength and one of which is lighter. Find the heavier version of this coin. The prediction model for this issue is shown in Figure *. The bodyweights of A + B and C + D are compared and assessed at the root node. The answer is "yes," and the left subsidiary is certainly true if A+ B surpasses C + D. Apart from that, though, A + D is the more challenging option since it calls for the usage of the proper branching. The node on the left branch compares how a and b train for weights. The greater load coin is chosen as an if the response to this question is "yes." If you respond "no," it selects b as the currency with the much larger value. The very same method is applied to c and d if the main node's outcome is "no."

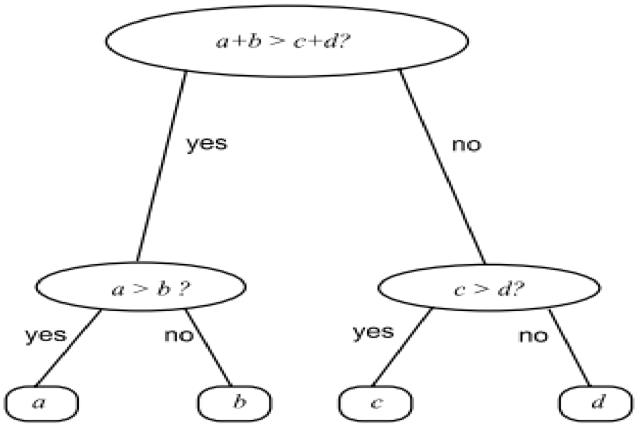


Figure 2: Decision tree figure

A straightforward decision tree in the context of the organization is shown in the example in Figure *. The following issues were mentioned:

The four potential outcomes are represented by four-leaf nodes. The weight of a coin is represented by each leaf node.

A definite finding or a leaf node requires two pairwise comparisons. Each weighing procedure corresponds to a certain stage. From the roots to every leaf, there are several decision nodes.

There is a rule for every tree, from the root to the leaf. For the followed by a decision, for instance, the criterion is "if A + B>C + D and A>B, then an is light."

3.3 Data Implementation:

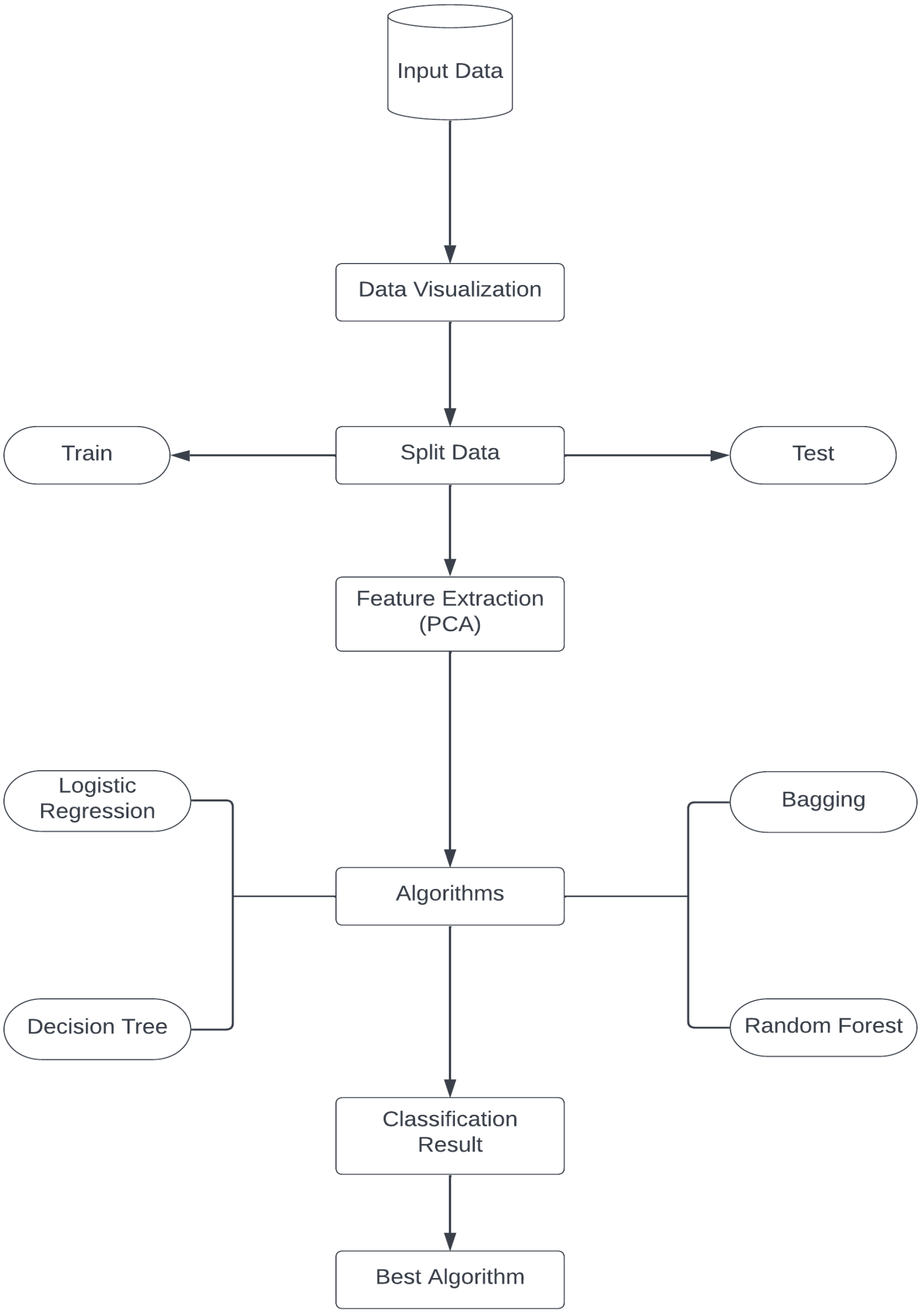


Figure 3: Data implementation

3.3.1 Input data : A system file holding data that will be utilized as input by a system or piece of software is known as input data. the source file Computational engineering is the branch of engineering technology that uses machines to study processes and structures that are calculable.

3.3.2 Split data : Machine learning frequently employ the approach of dividing data into separate groupings. To train the model, data information is often separated into train and test sets. choose the model's hyperparameters, and test the prediction error or accuracy of the model. The training datasets won't include enough data for the model to learn an effective translation from source to destination when the datasets is split into train and test sets. Additionally, the test technique would lack enough information to properly evaluate the model's performance.

3.3.3 Algorithm: There are four algorithm are used and they are:

1. Logistic Regression.(LR)
2. Decision Tree
3. Bagging
4. Random Forest

3.3.4 Classification Result: After applying this four algorithm we find out the best result.

3.3.5 Best algorithm: Applying four algorithm which algorithm give the best performances is called best algorithm.

3.4 Data visualization:

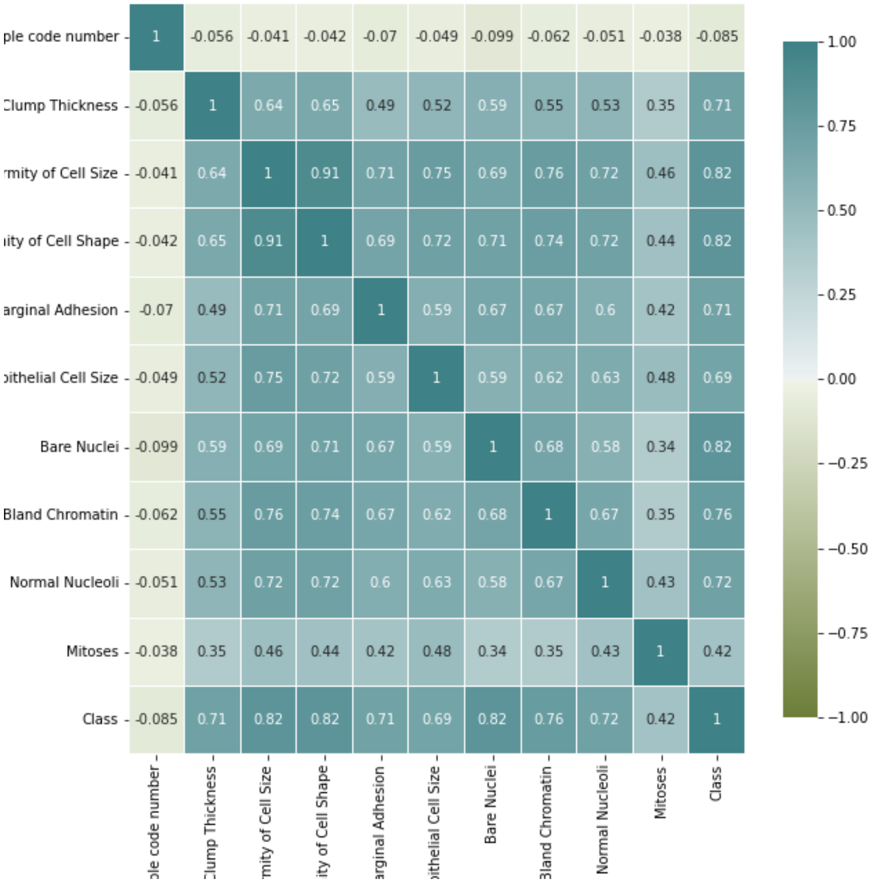


Figure 3: Corelation matrix

Corelation matrix:

Calculate and add the component products for each position of the two columns with respect to the first to get the cross-correlation of the two matrices. This may be used to determine the offset necessary to get two matrices of related values to overlap, however there are certain restrictions. A dataset is readily summarized using a correlation matrix. The relationships between all the variables in a dataset may be easily summarized using a correlation matrix. Regression diagnostics use a correlation matrix.

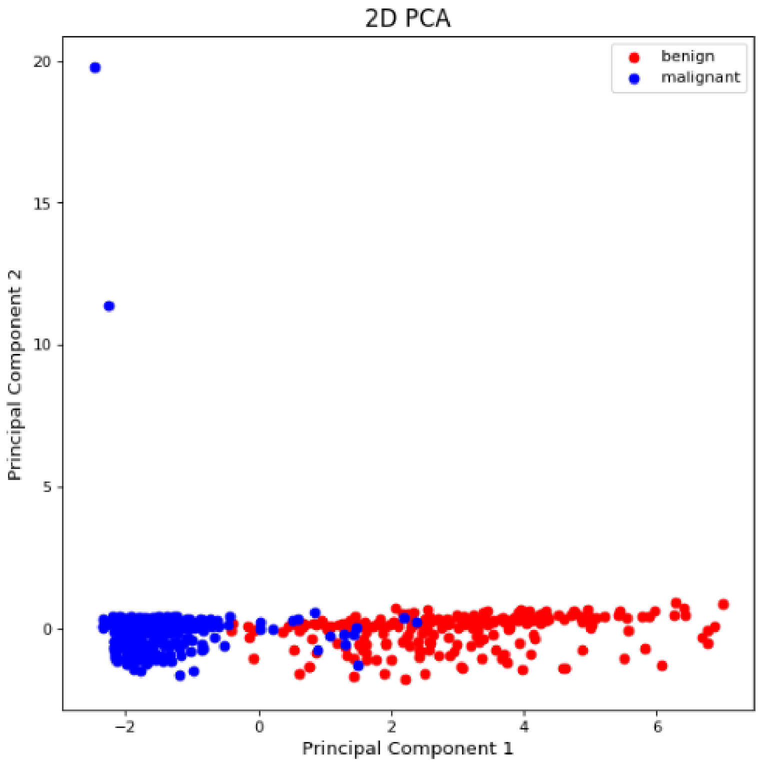


Figure 4: Principle component analysis

PCA:

A lot has been written about PCA on the internet, and there are several excellent articles on it, but many of them spend too much time getting too technical when most of us just would like to understand how it works simply.

Unsupervised ML algorithms like Principal Component Analysis are frequently employed in a variety of applications, such as exploratory

1. analysis of information
2. compression of information,
3. decrease of dimensions
4. data cleaning, among other things.

CHAPTER 4

Experimental Result Analysis and Discussion

4.1 Experiment result analysis:

Table 1: Evaluated Result For all Algorithm

Classification	class	Precision	Recall	F1-score
Logistic regression	2 4	0.65 0.00	1.00 0.00	0.79 0.00
Random Forest	2 4	1.00 0.95	0.97 1.00	0.98 0.97
Bagging	2 4	0.98 0.95	0.97 0.96	0.97 0.95
Decision tree	2 4	0.97 0.91	0.95 0.94	0.96 0.93

In table 1 I evaluated the result for algorithm . Here I apply 4 types of machine learning algorithm and they are They are decision trees, logistic regression, bagging, and random forest. In class 2 means benign or 4 means malignant. Here, I find out the precision, f1-score, and recall.

Table 2: Result of confusion matrices

Classification	Accuracy	AUC	Label	Predictive Negative(%)	Predictive Positive(%)
Logistic regression	64%	54%	Actual Negative Actual Positive	133 72	0 0
Random Forest	98%	98%	Actual Negative Actual Positive	129 0	4 72
Bagging	96%	99%	Actual Negative Actual Positive	129 3	4 69
Decision tree	94%	94%	Actual Negative Actual Positive	126 4	7 68

4.2 Discussion

In table 2 I evaluated the result of confusion matrices and find out the accuracy and AUC value.The Accuracy of decision trees 94%, logistic regression 64%, bagging 96%, and random forest 98%. Comparing the result here random forest give the best accuracy . And the AUC rate of decision trees 94%, logistic regression 54%, bagging 99%, and random forest 98%. so if i compare the result of these algorithm bagging give the best AUC value 99%. In this table bagging algorithm give the best performance . In this table, I also find out the result of confusion matrix. I labeling it as Actual Negative, and Actual Positive. Here logistic regression algorithm’s true positive is highest (133%) and false negative (0%).

CHAPTER 6

Impact on Society, Environment and Sustainability

6.1 Impact on society:

In addition to probable financial strain and maybe physical dislocation if a move is necessary for therapy, family members frequently deal with worry and sadness. Strengthening social support is a promising method to decrease psychological stress if you or someone you care about has been diagnosed with breast cancer. Breast cancer awareness is crucial because early identification, frequently through screening, can catch the illness when it is most curable. We now have a vast understanding of the basic mechanisms underlying the start, growth, and spread of cancer in the body thanks to research. These discoveries have produced more precise, effective therapies as well as preventative measures.

6.2 Sustainability:

Cancer is the sixth most common cause of death in Bangladesh, according to the Bangladesh Bureau of Statistics. At the National Institute of Cancer Research Hospital (NICRH) and the Oncology Department of Bangabandhu Sheikh Muijb Medical University (BSMMU), a hospital-based cancer registry has been established (10). The rate of cancer-related deaths in Bangladesh was anticipated by the International Agency for Research on Cancer (IARC) to be 7.5% in 2005 and will rise to 13% in 2030. According to IARC (2008), the top 10 cancer-related causes of death in men are lung, mouth and oro-pharyngeal, esophageal, pharynx, stomach, larynx, colorectal, lymphoma, liver, and bladder cancers, while the top 10 cancer-related causes of death in women are mouth, cervical, breast, oro-pharyngeal, lung, esophageal, gallbladder, stomach. There is a critical lack of qualified oncologists and physicists in Bangladesh. For better cancer care, oncopathology and cytopathology abilities must be improved. Additionally, technical personnel for imaging modalities and tissue diagnosis must be established. To produce complete and balanced services in a fair amount of time, the training requirements must be analyzed and the best training must be given[16].

CHAPTER 6

Conclusion

6.1 Conclusion: Cancer could be prevented by lowering health risks and enhancing preventative measures.A variety of factors can influence our risk of acquiring breast cancer throughout the course of our lives.By taking care of our health, we may help lower the chances of getting breast cancer, but other variables, like becoming older or having a family history, are beyond our control. We used ML techniques within that investigation. to distinguish between benign and malignant tumors. If the illness is mild, the doctor can start treatment immediately once. In this study, we identify the optimal method in terms of accuracy, precision, recall, and confusion matrix. Principal component analysis (PCA) and correlation matrices are utilized for data visualization. The bagging technique has the best AUC value (99%) and For the datasets presented, Random Forest has the highest accuracy 98%.

6.2 Future work: We used four algorithms to predict breast cancer here, and we want to use more in the future to identify it. We can employ a variety of techniques to achieve better and ideal results. Deep learning or artificial neural networks should be used in our future study for the greatest results.

References:

[1] ClevelandClinic. “Breast Cancer Overview: Causes, Symptoms, Signs, Stages and Types.” Cleveland Clinic, my.clevelandclinic.org/health/diseases/3986-breast-cancer. Accessed 23 Jan. 2023.

[2]Admin. “Seven Warning Signs of Breast Cancer &Mdash; Bay Imaging Consultants.” Bay Imaging Consultants, 29 Aug. 2019, www.bicrad.com/blog/seven-warning-signs-of-breast-cancer.

[3]ClevelandClinic. “Breast Cancer Overview: Causes, Symptoms, Signs, Stages and Types.” Cleveland Clinic, my.clevelandclinic.org/health/diseases/3986-breast-cancer. Accessed 23 Jan. 2023.

[4] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[5] Rawal, R. (2020). Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR), 13(24), 7.

[6] Singh, G. (2020). Breast cancer prediction using machine learning. Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., 8(4), 278-284.

[7]Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making, 19(1), 1-17.

[8] Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. Procedia Computer Science, 171, 593-601.

[9] Rane, N., Sunny, J., Kanade, R., & Devi, S. (2020). Breast cancer classification and prediction using machine learning. International Journal of Engineering Research & Technology, 9(02), 576-580.

[10]Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. Molecular oncology, 5(1), 5-23.

[11]Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: how special are they?. Molecular oncology, 4(3), 192-208.

[12]Rodríguez-Ruiz, A., Krupinski, E., Mordang, J. J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I., & Mann, R. M. (2019). Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology, 290(2), 305-314.

[13]Sadoughi, F., Kazemy, Z., Hamedan, F., Owji, L., Rahmanikatigari, M., & Azadboni, T. T. (2018). Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer: Targets and Therapy, 10, 219.

[14]Sechopoulos, I., Teuwen, J., & Mann, R. (2021, July). Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. In Seminars in Cancer Biology (Vol. 72, pp. 214-225). Academic Press.

[15]Huang, J., Cai, Y., & Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern recognition letters, 28(13), 1825-1844.

APPENDIX

Abbreviation:

MSE = Mean square error

PCA= Principle component analysis

MRI=Magnetic resonance imaging for the breast

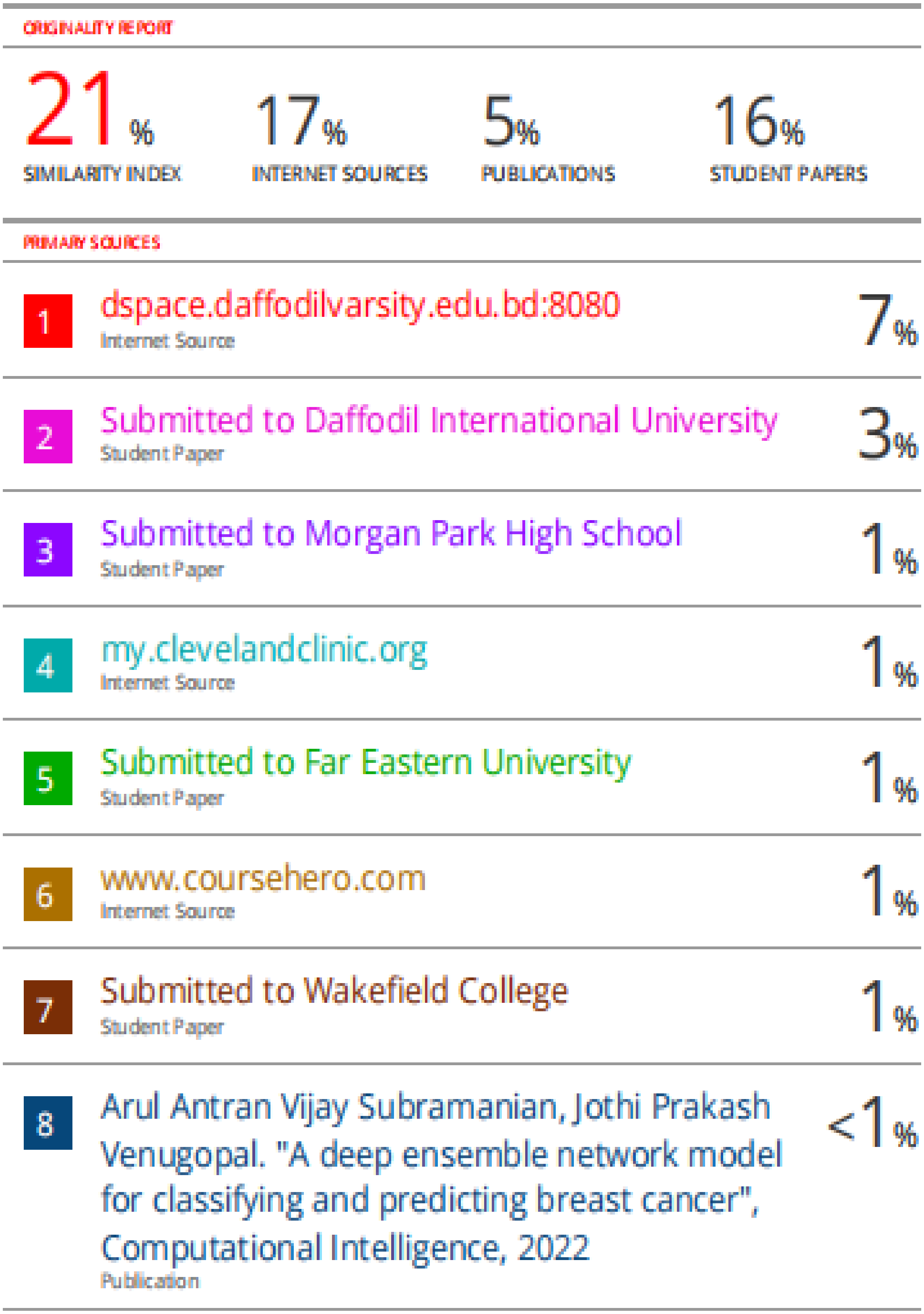
HRT=Hormone replacement treatment

TNBC=Breast cancer with three negatives

Appendix : Research reflection

We know a little bit about machine learning at the start of the project. We are unable to comprehend how algorithms function or how to anticipate the future. Our supervisor treated us with kindness and generosity. He provided us with helpful advice and was a great assistance. We learn a lot throughout this study period of work. We get new skills in technique, algorithm, and other methodologies.

Finally, through conducting the research, we have grown braver and motivated to conduct other research in the future.



ORIGINALITY REPORT

21 %	17%	5%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to Morgan Park High School Student Paper	1%
4	my.clevelandclinic.org Internet Source	1%
5	Submitted to Far Eastern University Student Paper	1%
6	www.coursehero.com Internet Source	1%
7	Submitted to Wakefield College Student Paper	1%
8	Arul Antran Vijay Subramanian, Jothi Prakash Venugopal. "A deep ensemble network model for classifying and predicting breast cancer", Computational Intelligence, 2022 Publication	<1%