UNIVERSITY OF MALAYA

WQD 7001 Principles of Data Science

Session 2023/2024 Semester 1 - Group Assignment 1

Improving Breast Cancer Diagnosis: Data-Driven Insights and Predictive Analysis

Group 7 - 5A's

Name & Matric No.	Role
WONG KAI THUNG (17101556)	Leader
GUNTUPU SOWJANYA (22085697)	Maker
XIAO YUE (22110867)	Oracle
WANG PENGXIANG (22105619)	Detective
TAN FONG YI (22108436)	Presenter

Lecturer

Dr. Rohana Mahmud

Table of Contents

Project Background

Breast cancer is the most prevalent cancer in the world with more than 4 million US women with a history of invasive breast cancer being alive on January 1, 2022.[1] According to the American Cancer Society, it is the second leading cause of cancer death in women. (First is lung cancer due to its low mortality rates.) The chance of a woman dying from breast cancer in the United States is about 1 in 39 (about 2.5%). On the other hand, there's only about 1% of all breast cancer cases occur in men.[2]

The development of breast cancer arises from the abnormal growth of cells in breast tissue, often identified as a tumour. A tumour can be benign(not cancerous) or malignant (cancerous). [3] To determine if a tumour is malignant or benign, a breast biopsy is required to remove tissue or fluid from the suspicious breast area. Fine-needle aspiration(FNA) is one of the common types of biopsies that remove fluid from the breast lump. The fluid samples are then examined under a microscope by a pathologist, looking for abnormal or cancerous cells. [4] This process is known as tumour grading.

Pathologists utilise tumour grading to categorise malignant breast cancer tumours according to the severity of mutations and the probability of spreading. Examination of breast cancer cells is performed under a microscope to assess factors such as the histologic grade (degree of resemblance to healthy cells), nuclear grade (shape and size of tumour cells' nuclei), and the rate of cell division and multiplication. This systematic evaluation aids in understanding the characteristics and potential aggressiveness of the cancer. [5]

However, since the Microscopic examination of FNA results are highly operator dependent [6], the tumour grading result is prone to bias and inaccuracy. Its reproducibility has been the subject of debate for decades and the inter- and intra-observer variation has been extensively reported. [7]

Over the past few years, machine learning has emerged as a transformative force in the healthcare industry, changing the way how medical data is conventionally analysed and interpreted. By utilising advanced algorithms and computational techniques, machine learning

has been instrumental in predicting various healthcare outcomes. One prominent application is in predictive analytics, where machine learning models are employed to forecast diseases such as diabetes[8], cancer[9], cardiovascular disease[10], mental health disorder[11] and etc.

Therefore, by harnessing the power of machine learning, we hope to improve the accuracy of breast cancer diagnosis. These algorithms can assist healthcare professionals in making more accurate and timely assessments of breast cancer risk, distinguishing between the benign and malignant tumours. This, in turn, will enable patients to receive treatment earlier due to the more timely and accurate diagnosis result.

Problem Statement

Breast cancer stands as the most prevalent cancer globally, claiming a significant toll on women's health. Despite advancements in diagnostic methodologies, the accuracy of distinguishing between benign and malignant tumours, crucial for determining appropriate treatment plans, remains a challenge. The result interpretation of Fine-Needle Aspiration (FNA), a biopsy method to extract fluid samples from breast lump, heavily relies on manual tumour grading through microscopic examination. The operator-dependent nature of this process not only introduces potential biases and inaccuracies but also contributes to significant inter- and intra-observer variations. Recognizing the limitations of conventional diagnostic approaches, this project aims to address the pressing issue of inconsistent tumour grading in breast cancer diagnoses by leveraging machine learning. The integration of advanced algorithms seeks to provide a more objective, reproducible, and accurate assessment of breast cancer risk, thereby enabling timely interventions and improving the overall effectiveness of breast cancer diagnosis and treatment.

Project Objective

1) **To Identify Features Associated with Breast Cancer Diagnosis**

   Conduct an in-depth analysis to find features that may be relevant and associated with breast cancer diagnosis. Prioritise the feature with a correlation or the ability to recognize benign and malignant tumours.

2) **To Explore Pairwise Correlations of Numerical Features in the Dataset**

   To analyse the numerical features in the dataset and their pairwise correlations. Exploration of numerical features using statistical correlation analysis (e.g., Pearson correlation coefficient) and visualisations to identify relationships between these features, which could find the dependencies and interactions.

3) **To Evaluate and Compare Models for Breast Cancer Assessment**

   Create and analyse machine learning models for predicting the risk of breast cancer using the identified features. Using techniques such as random forest, logical regression, and few others to develop a predictive model using the identified features To assess the model's performance based on metrics such as precision, accuracy, recall, F1-score, and so on.

Project Scope / Domain

Health industry:

The health industry continues to be a critical domain that necessitates continuous advancements in order to improve patient care, treatment efficacy, and diagnostic accuracy. This research, by focusing on breast cancer diagnosis, aims to directly contribute to the improvement of healthcare practices, benefiting both patients and healthcare professionals.

Cancer Diagnostics - Breast Cancer Emphasis:

Breast cancer is a common and terminal illness that claims millions of lives each year. The project's focus on breast cancer prediction aligns with addressing the challenges distinct to this cancer type, especially the accuracy of distinguishing between benign and malignant tumours. The research hopes to have a significant impact on cancer diagnosis and subsequent treatments by dealing with this critical aspect.

The adoption of data science/machine learning in the industry: Integrating data science and machine learning into healthcare has the potential to completely transform traditional diagnostic methods. These advanced computational methods allow the analysis of large datasets, allowing meaningful patterns and insights to be extracted. The project aims to introduce more objective, reproducible, and reliable evaluations of breast cancer diagnosis by leveraging these techniques, in turn mitigating the biases and inaccuracies common in manual grading methods.

Literature Study

Globally, breast cancer remains a major concern, especially among women. Due to manual assessments, traditional methods of diagnosis, such as Fine-Needle Aspiration (FNA) biopsies, often have accuracy issues. Machine learning (ML) is now recognized as a useful tool that can enhance breast cancer diagnosis and prognosis. In an effort to increase the accuracy of breast cancer diagnosis and prediction, this review looks into various machine-learning techniques and their effectiveness.

S. No.	Reference Article	ML Technique	Summary	Limitations
1	Assessing the performance of fully supervised and weakly supervised learning in breast cancer histopathology[12]	A comparison of two major machine learning paradigms: Fully Supervised Learning (FSL) and Weakly Supervised Learning based Multiple Instance Learning (WSLMIL) is used.	A comparison between the two approaches, Fully Supervised Learning (FSL), where the computer learns a lot from labelled data, and Weakly Supervised Learning (WSLMIL), where the computer learns a lot from less detailed information, examines the performance of various computer network architectures on a dataset pertaining to the spread of breast cancer nodes Newer networks are found to be more effective in FSL at identifying cancer in both small areas and across entire slides. However, in WSLMIL, the information fusion process has a greater impact on the outcomes than the network architecture.	With its focus on a single dataset, the study may not cover all forms of breast cancer. Their findings may not apply in any case because they only looked at a few computer techniques and didn't explore all available options. It may not be easy to apply these findings in actual hospitals because doctors may not fully understand the computer's prediction process. Furthermore, the primary focus of the discussion was the computer's capabilities rather than how doctors could use it in actual hospitals.

2	Feature learning based on connectivity estimation for unbiased mammography mass classification[13]	To identify breast cancer characteristics, they used DenseNet, a specialised computer network. By identifying comparable examples from the training data for every new case, they improved their predictions. By preventing the computer from picking up incorrect information, this technique aims to improve the computer's capacity for breast cancer detection and prediction.	The study enhances the detection of breast cancer by utilising a unique computer network known as DenseNet. They trained this network to identify the textures and characteristics of breast tissue. For every new case they tested, they identified a comparable few instances from the training data to ensure the computer didn't pick up biased information. Comparing this method to conventional methods, the computer was able to produce more accurate predictions. Additionally, it displayed graphics that clarified the computer's decision-making process, which may help in understanding when the machine makes errors.	The study's findings, which focused on particular data, might not apply to all datasets or forms of breast cancer. Although their approach enhances predictions and provides useful visuals, it may be difficult for non-experts to understand those visuals. Even though it functions better, there are a number of practical issues that make its use in actual hospitals challenging.

3	Random forest for breast cancer prediction [14]	To predict breast cancer, they employed the Random Forest technique, in which several computer decision trees collaborate. When specific portions of the dataset were used for training. This method demonstrated a high level of precision in predicting breast cancer.	This study predicts breast cancer using the Random Forest technique. The Wisconsin Breast Cancer Database was the dataset they used. The technique demonstrated extremely high accuracy, sometimes even hitting 100% when only a portion of the dataset was used for training. They think this approach could aid medical professionals in making more informed decisions regarding breast cancer.	It's important to realize that the study didn't look closely at potential computer errors. Furthermore, their method may not work as well on various datasets because it was only tested on one dataset. Despite the positive outcomes, there are practical issues that could make it application of this computing method in actual hospitals difficult.

Looking into machine learning techniques for the diagnosis of breast cancer shows promise for increasing accuracy and predicting outcomes better. Several machine learning (ML) models, including Fully Supervised Learning (FSL), Weakly Supervised Learning (WSLMIL), DenseNet, and Random Forest, have demonstrated success in identifying breast cancer traits and predicting whether it is malignant. However, there are still issues, particularly when using these techniques in actual hospitals. These methods need to be improved for wider usage in clinics, even though they function well in certain scenarios. This review emphasises that for cutting-edge techniques to be successfully applied in real medical settings, clarity and practicality are just as important as sophistication.

Methodology

In the pursuit of improving breast cancer diagnosis through data-driven insights and predictive analysis, our project adopts a comprehensive methodology guided by the OSEMN framework: Obtain, Scrub, Explore, Model, and Interpret. This framework serves as the backbone of our approach, providing a structured pathway to extract meaningful knowledge from the data. In this project, we will discuss the first three steps of OSEMN—Obtain, Scrub, and Explore.

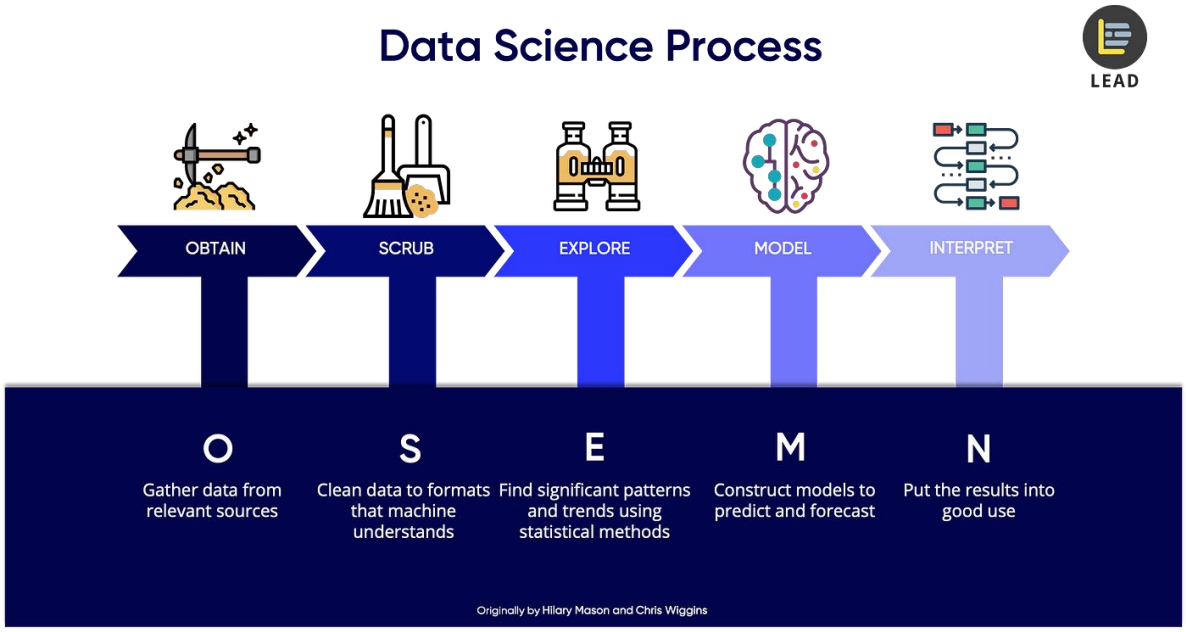

Figure 1: The OSEMN framework [15]

a. Obtain – Data Collection:

The first phase of our data science project involves obtaining the necessary data from relevant sources. Data reliability is a critical factor, and we prioritise sources known for their accuracy, consistency, trustworthiness, and relevance to the project objectives. After multiple rounds of research and careful consideration, we have chosen the dataset from The UCI Machine Learning Repository, titled "Breast Cancer Wisconsin (Diagnostic),"[16] . This dataset aligns seamlessly with our project's objectives and is recognized for its completeness and reliability.

b. Scrub – Data Cleaning:

Following data acquisition, the focus shifts to the scrubbing phase, aimed at ensuring the quality and integrity of the dataset.

```
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), int64(1), object(1)
```

First, we use the Python script `df.info()`, which provides a concise summary of the DataFrame to check for non-null values. The summary includes information such as the count of non-null values, data types for each column, and the total number of rows and columns in the DataFrame. This is helpful for gaining initial insights into the structure of the data, identifying any missing values, and understanding the data types of each column. The result shows that all 32 columns contain 569 non-null values, which means the dataset does not contain any missing values.

Figure 2: Output of Code Snippet `df.info()`

This is further confirmed by using the Python code `df.isna().sum()` which identify and count the number of missing values (NaN or null values) in each column of the DataFrame `df`.This information is useful for assessing the data quality and determining if there are any gaps in the dataset that need to be addressed, such as through imputation or removal of incomplete records.

```
id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
dtype: int64
```

Figure 3: Output of Code Snippet `df.isna().sum()`

	columns	nulls	nulls_%	unique
1	diagnosis	0	0.000000	2
26	smoothness_worst	0	0.000000	411
10	symmetry_mean	0	0.000000	432
2	radius_mean	0	0.000000	456
22	radius_worst	0	0.000000	457
⋮				
9	concave points_mean	0	0.000000	542
25	area_worst	0	0.000000	544
21	fractal_dimension_se	0	0.000000	545
16	smoothness_se	0	0.000000	547
0	id	0	0.000000	569

Figure 4: Summarization of The Null and Unique Values for Each Row in The Dataset

Then, we used the `unique` function to identify if there is any duplicated identity in the dataset. The 'id' column shows 569 unique values which is the same as the number of rows. Therefore, we can say that there are 569 unique identities in the dataset with no duplicated identity.

The figure on the left summarises the null values, the percentage of null values and the number of unique values relative to the total number of rows. The colour gradient of the 'unique' column, ranging from light to dark blue, helps highlight variations in the number of unique values across different columns.

c. Explore – Data Analysis:

The next step is exploratory data analysis (EDA), a crucial phase to gain insights into the structure and characteristics of the data.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rad
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 32 columns

Figure 5: Output of Code Snippet `df.head()`

`df.head()` shows the top rows of the DataFrame, allowing us to inspect the data and understand its structure. Our dataset contains variable such as ID number, Diagnosis (M = malignant, B = benign), ten real-valued features computed for each cell nucleus which are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The standard error and worst case of each feature make up the rest of the columns.

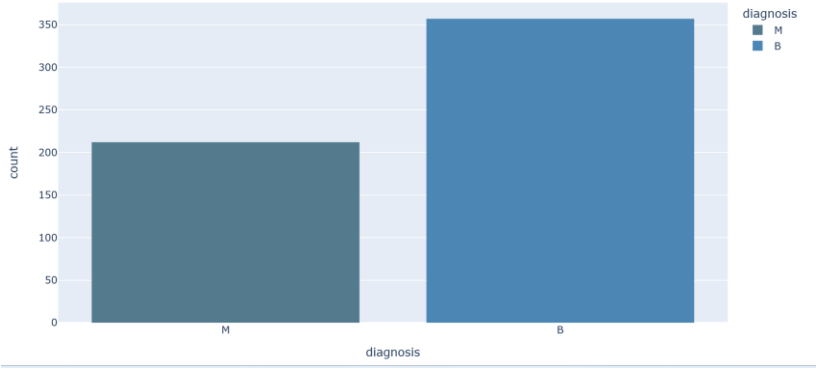

Figure 6: Histogram of Number of Tumours Diagnosed as Malignant (M) vs Benign (B)

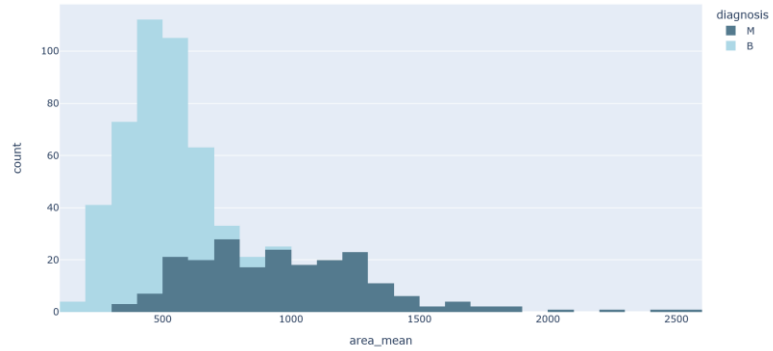

Figure 7: Histogram Visualising The Distribution of area_mean of Cells in The Malignant (M) and Benign (B) Categories

We also use histogram (Figure 6) to visualise the number of malignant (M) and benign (B) tumour diagnosis. There are more benign tumours diagnosed as compared to malignant tumours. Histograms like one in Figure 7 were also used to visualise the distribution of mean tumour areas, radii, perimeters and other numerical variables of the cells.

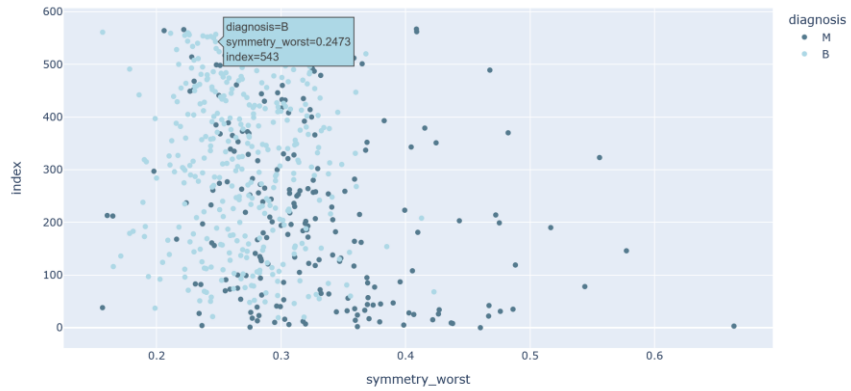

Figure 8: Scatter Plot Visualising The Distribution of symmetry_worst of Cells in The Malignant (M) and Benign (B) Categories

On the other hand, scatter plots were used to visualise the distribution of worst-case symmetry, concavity, fractal dimension and other numerical variables of tumours for different diagnostic categories.

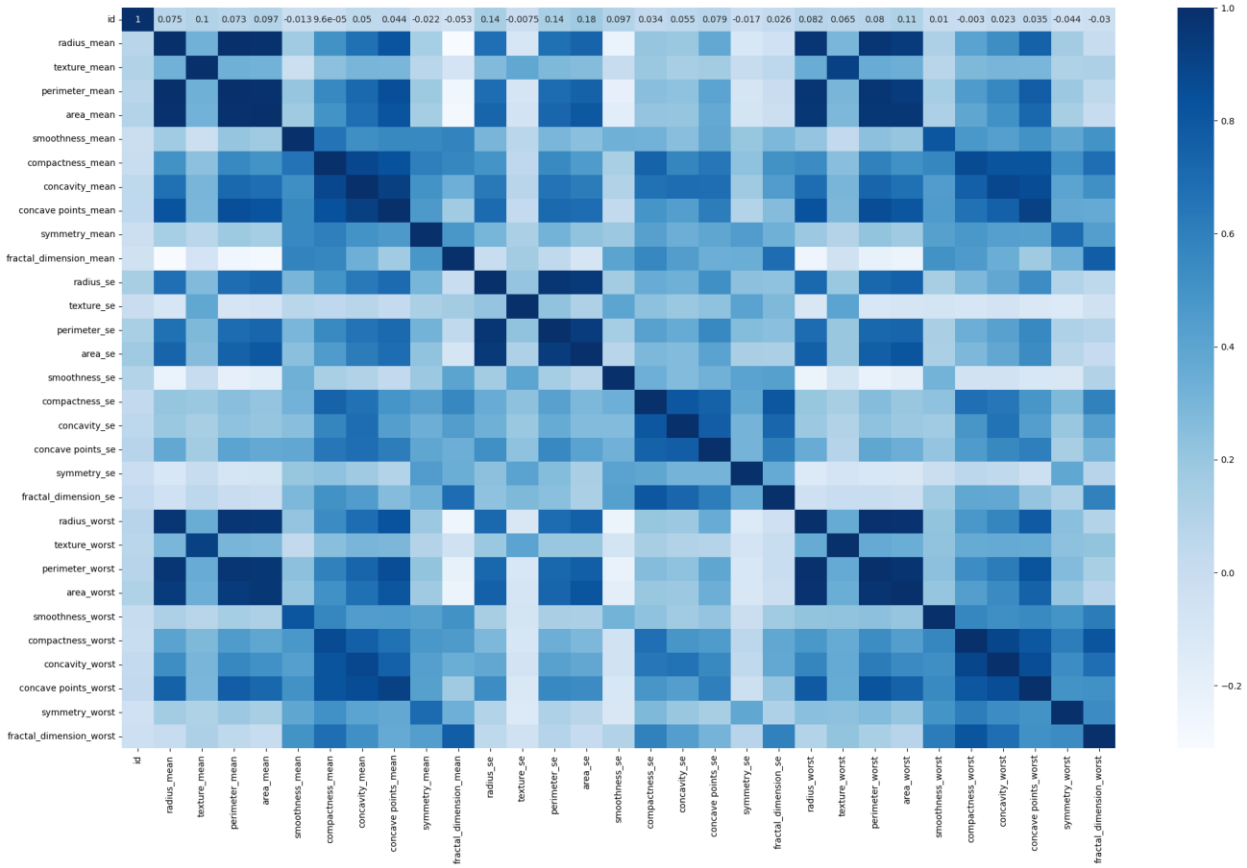

Figure 9: Heatmap Showing the Correlation between One Variable and Another

To identify the correlation between the variables, a heatmap was used where darker colours represent a stronger positive correlation.

From figure 9, we can see that the heatmap is annotated with the correlation values, and the diagonal from the top-left to bottom-right is dark blue with a correlation of 1, which is to be expected as it's the correlation of each variable with itself.

Aside from that, we observed that ‘radius_mean’ and ‘perimeter_mean’ are highly correlated with ‘area_mean’. We also notice that ‘perimeter_worst’ and ‘radius_worst’ are highly correlated.

Ethical Considerations

As we embark on the journey of integrating machine learning into breast cancer diagnosis, ethical considerations remain our priority. While the potential benefits are immense, we must navigate carefully to ensure that the project upholds principles of fairness, transparency, and privacy throughout its implementation. Below are a few ethical considerations that we have taken into account:

i) Anonymity:

Privacy of the participants is of paramount importance.The collected data in this project does not contain any personal-identifiable information such as name, addresses, phone numbers, email address, photos, videos, etc. This will safeguard the participant's privacy as we would not know who the participants are and we can't link any individual participant to their data.

ii) Fairness and Bias Mitigation:

In the development and deployment of machine learning models, it is crucial to identify potential biases that may emerge from the data used for analysis and modelling. This is why we have chosen data with anonymous participants in the first place to ensure no form of bias will occur. We will prioritise fairness in our project, striving to ensure that the benefits of accurate diagnosis are accessible to all individuals, irrespective of their demographic backgrounds.

iii) Transparency and Explainability:

Transparency is fundamental to earning public trust for both ethical and clinical reasons. Machine learning algorithms, while powerful, can be complex and challenging to interpret. We are committed to providing clear explanations of how the machine learning models function, enabling healthcare professionals and patients to comprehend the basis for diagnostic outcomes.

iv) Potential for Harm:

Acknowledging the potential for unintended consequences is a fundamental ethical precept. We are acutely aware of the responsibility inherent in deploying machine learning for healthcare applications. We always strive to ensure no type of possible harm such as psychological harm, social harm, physical harm or legal harm were made to anyone involved in this project.

v) Results Communication:

In the communication of results, we are committed to deliver findings that are honest, reliable, and credible. We will ensure that the results we communicate accurately reflect the outcomes of our research. We emphasise the importance of originality and integrity by strictly avoiding plagiarism in any form. Additionally, we adhere to stringent ethical standards to prevent research misconduct, which includes the prohibition of manipulating data analyses, falsifying data, or misrepresenting results in research reports.

vi) Continual Evaluation and Improvement:

Ethical considerations extend beyond the project's initiation. Continuous monitoring and evaluation of the machine learning models must be incorporated to identify and rectify any biases, errors, or unintended consequences that may arise during implementation. Feedback from all parties will be actively sought and incorporated to ensure ongoing compliance with evolving ethical standards.

Impact of the Project to the Society

We believe the implementation of machine learning in breast cancer diagnosis will bring a positive impact to individuals, healthcare systems and societal level. On an individual level, by enhancing the accuracy and timeliness of breast cancer detection, patients can initiate treatment plans sooner if diagnosed with a malignant tumour, improving their chances of recovery. This also minimises unnecessary treatments due to diagnosis error, thereby mitigating potential side effects and financial burden of a patient.

As for the healthcare system, timely and accurate diagnosis can significantly reduce the burden of the healthcare systems, lowering the overall cost of healthcare associated with advanced-stage cancer treatment. This, in turn, may lead to more effective resource allocation, optimising the efficiency of healthcare delivery. The public will also be more confident with the overall healthcare system given the enhanced quality provided.

On a societal level, the project contributes to the advancement of medical research and technological innovation. The introduction of machine learning in breast cancer diagnosis not only improves our understanding of breast cancer but also paves the way for the development of more sophisticated and effective diagnostic tools to be applied to other aspects of healthcare in the future.

In summary, the positive impact of this project is multifaceted, encompassing improved health outcomes for individuals, a more efficient use of resources for the healthcare system, and potential progress in medical research.

Reference

[1] Miller, K. D., Nogueira, L., Devasia, T., Mariotto, A. B., Yabroff, K. R., Jemal, A., Kramer, J., & Siegel, R. L. (2022). Cancer treatment and survivorship statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(5). https://doi.org/10.3322/caac.21731

[2] American Cancer Society. (2022). *Breast Cancer Facts & Figures 2022-2024*. Atlanta, Georgia.

[3] Patel, A. (2020). Benign vs Malignant Tumors. In *JAMA Oncology* (Vol. 6, Issue 9). https://doi.org/10.1001/jamaoncol.2020.2592

[4] “Breast Biopsy: Procedure Types, What to Expect & Results Guide.” *National Breast Cancer Foundation*, June 2023, www.nationalbreastcancer.org/breast-cancer-biopsy/.

[5] Bloom, H. J., & Richardson, W. W. (1957). Histological grading and prognosis in breast cancer a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, *11*(3). https://doi.org/10.1038/bjc.1957.43

[6] Giard, R. W. M., & Hermans, J. (1992). The value of aspiration cytologic examination of the breast a statistical review of the medical literature. *Cancer*, *69*(8). https://doi.org/10.1002/1097-0142(19920415)69:8<2104::AID-CNCR2820690816>3.0.CO;2-O

[7] van Dooijeweert, C., van Diest, P. J., & Ellis, I. O. (2022). Grading of invasive breast carcinoma: the way forward. In *Virchows Archiv* (Vol. 480, Issue 1). https://doi.org/10.1007/s00428-021-03141-2

[8] Zafar, F., Raza, S., Khalid, M. U., & Tahir, M. A. (2019). Predictive analytics in healthcare for diabetes prediction. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3326172.3326213

[9] Choudhury, A., & Eksioglu, B. (2019). Using predictive analytics for cancer identification. *IISE Annual Conference and Expo 2019*.

[10] Muniasamy, A., Muniasamy, V., & Bhatnagar, R. (2021). Predictive analytics for cardiovascular disease diagnosis using machine learning techniques. *Advances in Intelligent Systems and Computing*, *1141*. https://doi.org/10.1007/978-981-15-3383-9_45

[11] Hahn, T., Nierenberg, A. A., & Whitfield-Gabrieli, S. (2017). Predictive analytics in mental health: Applications, guidelines, challenges and perspectives. In *Molecular Psychiatry* (Vol. 22, Issue 1). https://doi.org/10.1038/mp.2016.201

[12] Kang, H., Xu, Q., Chen, D., Ren, S., Xie, H., Wang, L., Gao, Y., Gong, M., & Chen, X. (2024). Assessing the performance of fully supervised and weakly supervised learning in breast

cancer histopathology. *Expert Systems with Applications*, *237*. https://doi.org/10.1016/j.eswa.2023.121575
[13] Guobin Li, Reyer Zwiggelaar.(2024).Feature learning based on connectivity estimation for unbiased mammography mass classification.Computer Vision and Image Understanding, Volume 238.https://doi.org/10.1016/j.cviu.2023.103884
[14] Octaviani, T. L., & Rustam, Z. (2019). Random forest for breast cancer prediction. *AIP Conference Proceedings*, *2168*. https://doi.org/10.1063/1.5132477
[15] LEAD. (2019). *Data Science Process*. 5 Steps of a Data Science Project Lifecycle. Retrieved November 10, 2023, from https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492.
[16] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.