

# **BREAST CANCER SURVAIVAL PREDICTION USING MACHINE LEARNING AND DEEP LEARNING**



**MARTEENA ROY – 10616560**

Applied Research Project submitted in partial fulfilment of the requirements for the degree of  
Master of Science in Data Analytics at Dublin Business School

Supervisor: Dr.Syed Mustufa

August 2023

## **Declaration**

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of Master of Science in Data Analytics is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Marteena Roy

Student Number: 10616560

Date:29/08/2023

## **Acknowledgements**

First and foremost, I would like to offer my sincere gratitude to my Dissertation Supervisor Dr.Syed Mustufa of MSc in Data Analytics course at the Dublin Business School. Dr.Syed Mustufa was always open to help me whenever I was in trouble or had questions about my research or writing. He consistently ensured that my Masters research paper was unique and my own work. He also corrected me at times when he felt any problems or issues in my research artifact or writing. He was always open for any type of discussions regarding the research. Every meeting with him was a great learning experience since he asked me to try and do investigation on a lot of options in the areas which were very new for me. Also, I would like to thank and I'm in-debt to other teaching faculty at Dublin Business School in order to help me learn and implement the necessary knowledge, technologies and tools in order to successfully accomplish my dissertation. Lastly, my family deserves endless gratitude.

## **Abstract**

Breast cancer continues to be a highly widespread and formidable kind of malignancy on a global scale. The precise estimation of survival is of utmost importance in order to customize treatment plans and enhance patient outcomes. The objective of this study was to improve the accuracy of breast cancer survival prediction by combining two well-known datasets: METABRIC, which provides extensive molecular profiling data, and CBIS-DDSM, a valuable collection of mammographic images. The study utilized three sophisticated machine learning models, namely Inception Net, Adam Net, and DenseNet, to make the most of how molecular and imaging information can work together effectively. The DenseNet model stood out, achieving an 81% accuracy in predicting breast cancer patient survival, surpassing Inception Net (66%) and Adam Net (71%). We can create a helpful tool using DenseNet121 to study how well breast cancer patients might survive by looking at their mammograms.

**Keywords: CBIS-DDSM, METABRIC, Mammography, Gene Expression, Inception Net, Adam Net and DenseNet.**

# Table of Contents

1	Introduction .....	11
1.1	Motivation .....	12
1.2	Research Question .....	14
1.3	Research Objectives .....	14
1.4	Research Outline .....	14
2	Literature Review .....	17
2.1	General Literature: .....	17
2.1.1	Diagnostic Techniques: .....	17
2.1.2	Treatments: .....	18
2.2	Related Literature: .....	19
2.3	Conclusion .....	30
3	Methodology .....	31
3.1	Data Collection .....	31
3.1.1	CBIS-DDSM .....	31
3.1.2	METABRIC .....	32
3.2	Exploratory Data Analysis .....	34
3.3	Pre-processing .....	35
3.3.1	Finding The Common Attribute .....	35
3.3.2	Determining tumor size in the CBIS-DDSM dataset .....	36

3.3.3	Resampling of the classes .....	37
3.3.4	Splitting the Data .....	38
3.4	Modeling .....	<b>Error! Bookmark not defined.</b>
3.4.1	Inception Net .....	39
3.4.2	Adam net .....	44
3.4.3	Dense Net .....	46
3.5	Evaluation.....	51
3.5.1	Confusion Matrix.....	51
3.5.2	Accuracy.....	52
3.5.3	Precision.....	52
3.5.4	Recall (Sensitivity).....	52
3.5.5	F1-Score .....	52
4	Results .....	53
4.1	Results for the Inception Net.....	53
4.1.1	Confusion Matrix.....	53
4.1.2	Classification Report.....	54
4.2	Results for Adam Net.....	55
4.2.1	Confusion Matrix.....	55
4.2.2	Classification Report.....	56
4.3	Results for Dense Net.....	57

4.3.1	Confusion Matrix.....	57
4.3.2	Classification Report.....	58
4.4	Conclusion.....	59
5	Conclusion and Future Work.....	60
5.1	Future Work.....	61
6	References.....	62

## List of Figures

Figure 1.1: Research Outline .....	15
Figure 3.1: Methodology flow .....	31
Figure 3.2: Snapshot of the 'dicom_info.csv' file in CBIS-DDSM dataset (Source: Notebook)...	34
Figure 3.3: Presence of null values in the dicom_info.csv file (Source: Notebook) .....	35
Figure 3.4: Binning the tumor size in METABRIC.....	37
Figure 3.5: Distribution of Classes .....	38
Figure 3.6: Architecture of the Inception Net V3 model (Source: <a href="http://www.medium.com">www.medium.com</a> ).....	39
Figure 3.7: Implementation of Inception Net V3 (Source: Notebook) .....	42
Figure 3.8: Training performance of Inception Net.....	44
Figure 3.9: Adam Net model summary .....	45
Figure 3.10: Adam Net Training Performance .....	46
Figure 3.11: DenseNet121 architecture (Source: <a href="http://www.towardsdatascience.com">www.towardsdatascience.com</a> ).....	48
Figure 3.12: Implementation of DenseNet121 (Source: Notebook) .....	50
Figure 3.13: Training Performance of the DenseNet121 model (Source: Notebook) .....	51
Figure 4.1: Confusion matrix for Inception Net (Source: Notebook).....	53
Figure 4.2: Classification report for Inception Net (Source: Notebook) .....	54
Figure 4.3: Subset of test images and their corresponding prediction by Inception Net (Source: Notebook).....	55
Figure 4.4: Confusion matrix for Adam Net (Source: Notebook) .....	56
Figure 4.5: Classification report for Adam Net (Source: Notebook).....	56
Figure 4.6: Testing on test data (Source: Notebook) .....	57
Figure 4.7: Confusion matrix for DenseNet121 (Source: Notebook).....	58

Figure 4.8: Classification report for DenseNet121 .....58

Figure 4.9: Testing DenseNet121 on Test Data (Source: Notebook) .....59

## List of Tables

Table 3.1: Dataset files and their contents .....32

Table 3.2: Attributes in METABRIC dataset .....33

# 1 Introduction

Breast cancer, a malignant tumor that originates in the cells of the breast, is a global health concern that has garnered significant attention from the medical community, researchers, and the public alike (Coughlin and Ekwueme, 2009). It stands as one of the most diagnosed cancers worldwide, with both developed and developing countries facing its impact. The World Health Organization (WHO) has identified breast cancer as the leading cause of cancer-related deaths among women, emphasizing its significance in global health agendas (Coughlin and Ekwueme, 2009).

The pathogenesis of breast cancer is multifaceted, influenced by a combination of genetic, environmental, and lifestyle factors. Genetic mutations, family history, age, exposure to estrogen, and certain reproductive patterns are among the well-established risk factors. Over the years, advancements in medical imaging, like mammography and magnetic resonance imaging (MRI), have improved the early detection of breast tumors, leading to timely interventions and better patient outcomes<sup>1</sup>.

However, the journey from diagnosis to treatment and prognosis is intricate. The heterogeneity of breast cancer, characterized by its various subtypes and stages, makes each patient's journey unique<sup>1</sup>. Factors such as the size of the tumor, its grade, lymph node involvement, and the presence of specific receptors like HER2, estrogen, and progesterone receptors play a crucial role in determining the course of treatment and the subsequent prognosis<sup>1</sup>.

In recent years, the concept of personalized medicine has gained traction. The idea is to tailor medical treatment to the individual characteristics of each patient, moving away from a 'one-size-fits-all' approach. This shift is particularly relevant in the context of breast cancer, where the

---

<sup>1</sup> <https://www.ncbi.nlm.nih.gov/books/NBK20362/>

variability in tumor characteristics and patient profiles necessitates a personalized approach to care<sup>1</sup>.

Enter the realm of machine learning and big data. With the exponential growth in data generation and storage capabilities, there is now an unprecedented opportunity to analyze vast datasets, extracting meaningful patterns and insights (Siu, 2016). Machine learning, a subset of artificial intelligence, is particularly poised to make significant contributions in this domain (Siu, 2016). By training algorithms on large datasets, machine learning can identify patterns and make predictions that are often beyond the capability of traditional statistical methods (Siu, 2016).

In the context of breast cancer, machine learning offers the potential to predict patient outcomes, optimize treatment pathways, and even identify at-risk populations based on a combination of clinical, genetic, and demographic data (Siu, 2016). This fusion of technology and medicine holds the promise of revolutionizing breast cancer care, making it more precise, personalized, and effective (Siu, 2016).

## **1.1 Motivation**

The motivation for doing this study is complex and encompasses multiple factors. The field of medical study has historically placed significant emphasis on breast cancer, given its widespread occurrence and consequential effects (Jemal *et. al.*, 2011). The intricate nature of the disease, characterized by a multitude of subtypes and phases, underscores the ongoing need for enhanced diagnostic and prognostic instruments (Sørliie *et. al.*, 2001). Although classic statistical methods provide inherent value, they frequently encounter challenges in effectively managing the extensive and complex datasets that are currently accessible. Machine learning, because to its capacity to process vast amounts of data and identify complex patterns, emerges as a viable approach for

improving the precision of predictions (Esteva *et. al.*, 2017). Enhancing the predictability of breast cancer survival holds the potential to enable doctors to make better-informed decisions on treatment programs, afford patients a more comprehensive comprehension of their prognosis, and optimize the allocation of healthcare resources (Hwang *et. al.*, 2013).

One notable advancement in recent times is to the creation of multi-omics data. The integration of gene expression data and mammography pictures offers a distinct possibility to get a thorough comprehension of breast cancer (Gao *et. al.*, 2022). The analysis of gene expression data offers valuable information regarding the underlying molecular and cellular mechanisms that contribute to the progression of a disease (Perou *et. al.*, 2000). On the other hand, mammography images provide a visual depiction of the tumor's characteristics, including its dimensions, morphology, and density (Lehman *et. al.*, 2015). The integration of these two data sets holds the potential to provide a comprehensive perspective on the condition, facilitating more precise prognostications and tailored therapeutic approaches (Yu *et. al.*, 2016).

One illustrative instance involves the utilization of gene expression data to elucidate the molecular pathways and genes that are correlated with the manifestation of aggressive tumor behavior (Desmedt *et. al.*, 2007). In parallel, mammography serves as a valuable diagnostic tool that can offer preliminary indications of tumor growth and dissemination at an early stage. By utilizing machine learning algorithms capable of processing and integrating both types of data, there exists the possibility of constructing predictive models that exhibit higher accuracy and clinical relevance compared to models based solely on one data source (Li *et. al.*, 2018).

Nevertheless, notwithstanding the considerable promise, the obstacles are equally formidable. The computational integration of gene expression data with mammography pictures is a complicated task, necessitating the use of advanced algorithms and models due to the large volume and intricate

nature of the data. The objective of this work is to address the existing disparity by utilizing machine learning techniques to exploit the valuable information provided by the integration of gene expression and mammography data. The ultimate goal is to improve the accuracy of breast cancer survival predictions.

## **1.2 Research Question**

The primary research question guiding this study is: *"How can machine learning techniques be effectively utilized to predict the survivability of breast cancer patients based on clinical and demographic data?"*

## **1.3 Research Objectives**

The objectives of this research are as follows:

- To review the current literature on breast cancer survivability prediction and the application of machine learning in healthcare.
- To identify relevant clinical variables that influence breast cancer survivability.
- To design and implement machine learning models such as Inception Net, Adam Net, and Dense Net for breast cancer survivability prediction.
- To evaluate the performance of the developed models against traditional statistical methods.
- To provide insights and recommendations for future research and clinical applications.

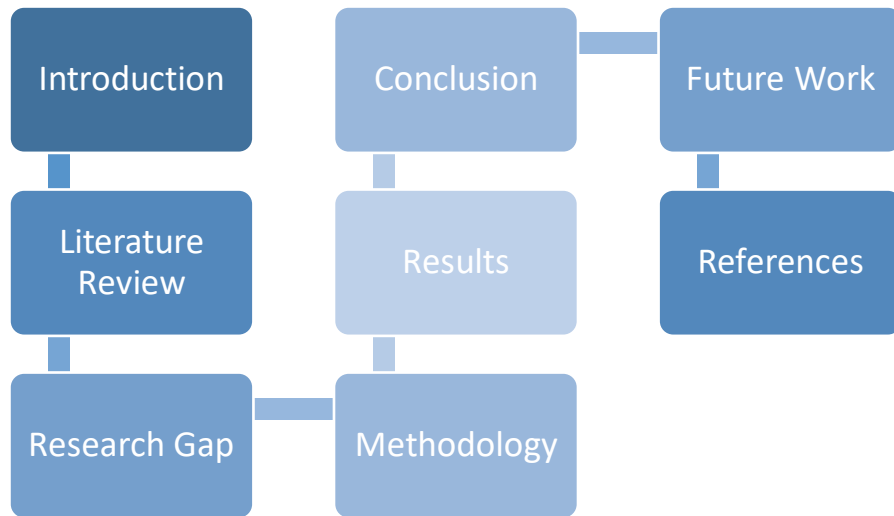
## **1.4 Research Outline**

This chapter introduced the reader to the background and the aim of the presented study. The rest of the report is structured as follows:

## Chapter 2: Literature Review

A comprehensive review of existing literature on breast cancer survivability prediction, highlighting traditional methods and recent advancements with the integration of machine learning.

Figure 1.1 below shows the research outline



**Figure 1.1: Research Outline**

## Chapter 3: Methodology (Implementation)

Detailed description of the data collection process, preprocessing techniques, selection of machine learning algorithms, and model evaluation metrics.

## Chapter 4: Results and Discussion

Presentation of the results obtained from the implemented machine learning models, followed by a discussion comparing the findings with existing methods and highlighting the significance of the results.

## **Chapter 5: Conclusion and Future Scope**

Summarizing the key findings of the study, drawing conclusions based on the research objectives, and suggesting areas for future research and potential clinical applications.

## **2 Literature Review**

Breast cancer remains one of the most prevalent and extensively researched malignancies worldwide. Its early detection and effective treatment are paramount to improving patient outcomes. Over the years, various diagnostic techniques and treatments have been developed to enhance our understanding and management of this disease.

### **2.1 General Literature:**

#### **2.1.1 Diagnostic Techniques:**

Effective communication between the physician and patient is crucial when a breast cancer diagnosis is made. This ensures that the emotionally overwhelmed patient is educated about her disease and available treatments, enabling her to participate in decisions about her care. A study by Roberts et al. (1994) postulated that patients whose surgeons employed psychotherapeutic techniques during the cancer diagnostic interview exhibited better psychological adjustment to their cancer.

This has emerged as a potential alternative or supplementary tool to conventional X-ray mammography for breast cancer detection. Wang (2018) emphasized that the quality of microwave breast imaging is influenced by factors such as the microwave sensor, sensor array, and the number of sensors in the array. The advantage of microwave sensors is that they offer better cancer detection and improved image resolution.

The human epidermal growth factor 2 (HER2) is a critical marker in breast cancer. Traditional diagnostic techniques to assess HER2 status include immunohistochemistry and fluorescence in situ hybridization. However, these methods require invasive biopsies. An alternative approach is molecular imaging using probes against HER2, allowing for a noninvasive, whole-body

assessment of HER2 status in real time (Massicano *et. al.*, 2018). This technique can potentially select patients who may benefit from HER2-directed therapy.

Triple-negative breast cancer (TNBC) is known for its aggressive proliferation and poorer clinical diagnosis compared to other breast cancer types. While chemotherapy, surgical removal, and radiation therapy are common treatments for TNBC, they come with side effects and costs. An emerging area of interest is the use of nanobodies, which have advantages in therapeutic agents and diagnostic kits. Nanobodies, due to their size and stability, can be used against antigens expressed in TNBC, offering a potential diagnostic and therapeutic solution (Bakherad *et. al.*, 2022).

Mammography has been the primary tool for breast cancer screening for over three decades. With technological advancements, newer diagnostic procedures like digital mammography, computer-aided testing, and tomosynthesis have emerged. Tomosynthesis, also known as 3-D mammography, captures multiple images of the breast from different angles, providing a three-dimensional view. This technique has shown improved breast cancer detection rates and reduced false positives (Hassan, 2022).

### **2.1.2 Treatments:**

Fallowfield and Jenkins (2014) highlighted that while modern breast cancer treatments offer many women greater prospects of cure or lengthier, good quality survival, there are concerns that these improvements have not led to similar benefits in the psychosocial, functional, and sexual well-being of women. They emphasized the need for systematic monitoring of quality of life-threatening side effects to permit early implementation of effective interventions.

Exosomes, biological nanovesicles present in bodily fluids, have shown potential in cancer diagnosis and therapy. These vesicles can deliver bioactive moieties such as proteins, lipids, and drugs, making them a promising tool for breast cancer management. Their role in identifying genes with pronounced expression differences and as drug delivery carriers offers a novel approach in breast cancer treatment (Kumar *et. al.*, 2022).

For young women diagnosed with breast cancer, fertility preservation is a significant concern. Controlled ovarian stimulation (COS) for oocyte/embryo cryopreservation is standard before starting chemotherapy. However, concerns arise, especially in hormone-sensitive tumors. Recent studies suggest that COS or assisted reproductive technology (ART) before or after breast cancer treatments does not seem to negatively impact survival outcomes (Arecco *et. al.*, 2022).

## **2.2 Related Literature:**

Breast cancer remains a leading cause of mortality among women worldwide. The urgency for an accurate and timely diagnostic system is underscored by the alarming statistics, with breast cancer accounting for an estimated 15.5% of cancer deaths in women in 2020. This paper by D. Santos (2022) leverages machine learning to address this pressing concern. By employing algorithms that iteratively learn from data, the study achieves an impressive accuracy and precision level of approximately 79% in predicting breast cancer survival based on gene expression profiles.

The paper's claim of a 79% accuracy in predicting breast cancer survival is commendable. However, without details on the Xgboost model employed, the dataset characteristics, or the validation techniques, it becomes challenging to assess the robustness and replicability of the study. The emphasis on gene expression profiles as a predictor is noteworthy. Yet, the paper could delve deeper into the challenges encountered during the study, the limitations of the model, and

potential areas of improvement. A comparative analysis with other data types or an integration approach with multi-omics data might offer a more holistic perspective on breast cancer survival prediction.

This paper by Jiang *et. al.* (2022) delves into the potential of somatic genomic variants, including copy number variation and SNP locus, in predicting the five-year survival rate of breast cancer patients. The study's choice of the CatBoost model is commendable, offering a machine-learning approach that can handle categorical data without the need for extensive preprocessing. However, the rationale behind selecting the CatBoost model over other potential machine learning models could be elaborated upon. This would offer readers insights into the decision-making process and the perceived advantages of CatBoost in this context.

While the study claims an AUC of 0.70 on an independent external dataset, a comparative analysis with other existing models or studies would provide a clearer benchmark for the study's contributions. The identification of potential biomarkers like TP53, DNAH11, and MAP3K1 is a significant contribution to the field. However, a more in-depth exploration of their roles, significance in breast cancer prognosis, and potential therapeutic implications would amplify the paper's impact.

Triple Negative Breast Cancer (TNBC) is a subtype of breast cancer that lacks the three most common types of receptors known to fuel breast cancer growth. This paper provides a comprehensive analysis of the impact of chemotherapy on early elderly patients diagnosed with TNBC. By leveraging the Surveillance, Epidemiology, and End Results (SEER) Database, a significant and authoritative source of cancer statistics, the study by Huang *et. al.* (2022), offers insights into the treatment outcomes and survival rates of this specific patient group.

The paper's focus on early elderly patients diagnosed with TNBC offers a niche perspective, addressing a specific subset of breast cancer patients that might have unique treatment responses and outcomes. The utilization of the SEER Database lends credibility to the study due to its extensive and detailed cancer statistics. However, a discussion on potential biases, limitations, or discrepancies associated with this database would provide a more balanced view.

The choice of the LightGBM model for survival prediction and its comparative analysis with other algorithms is well-presented. Yet, a deeper exploration of the model's hyperparameters, optimization techniques, and potential biases would offer readers a more comprehensive understanding of the study's methodology.

Given the unique challenges faced by elderly patients, such as potential comorbidities and different physiological responses to treatment, the paper could delve deeper into the implications of its findings for personalized treatment plans and healthcare policies.

Breast cancer and osteoporosis are two conditions that, at first glance, might seem unrelated. However, the paper by Ji *et al.* (2022) bridges the gap by investigating the relationship between these two conditions. The study aims to predict the risk of osteoporosis, fracture occurrence, and overall prognosis in breast cancer patients using machine learning models. The paper's unique approach of linking osteoporosis risk with breast cancer prognosis is both innovative and crucial for a holistic understanding of patient health. The study's comprehensive use of six different machine learning methods provides a robust evaluation of the best predictive approach for this unique problem. The superior performance of the XGB model, especially when compared to existing models like FRAX and OSTA, is a testament to the potential of machine learning in healthcare. However, the study's dataset, while significant, could benefit from further diversification. Incorporating data from different regions or ethnicities might offer a more

generalized model that can be applied across diverse patient populations. The paper could probe deeper into the biological and clinical implications of its findings. Understanding the relationship between osteoporosis and breast cancer, especially the potential shared pathways or risk factors, could pave the way for integrated treatment plans and preventive measures.

Early diagnosis of breast cancer can significantly improve survival rates and reduce treatment complexities. The study by Dehdar *et. al.* (2023) focuses on the factors leading to delayed breast cancer diagnosis among women in Iran. By employing four distinct machine learning methods, the research identifies specific demographic and social factors that increase the risk of diagnosis delay.

The paper's emphasis on the delay in breast cancer diagnosis offers a fresh perspective on the challenges faced in early detection and treatment. The identification of high-risk groups, such as urban-residing women who marry or have their first child after the age of 30 and those without children, is crucial for targeted interventions and awareness campaigns. While the study's focus on Iranian women offers valuable regional insights, its findings might not be directly applicable to other populations with different socio-cultural dynamics. The use of multiple machine learning methods, including XGBoost, random forest, neural networks, and logistic regression, provides a comprehensive analysis. However, a deeper exploration of the performance metrics of each model and their comparative advantages would enhance the paper's impact. The study could further delve into the implications of these findings on healthcare policies, suggesting interventions to reduce diagnosis delays and improve breast cancer awareness.

While genetic factors play a significant role in breast cancer prognosis, clinical attributes can also offer valuable predictive insights as suggested by Nair and Sahai (2022). This study focuses on predicting a specific label, the overall survival months, for breast cancer patients based solely on

clinical attributes. By employing Multivariate Regression and Random Forest models, the research assesses the relative importance of each clinical variable in prognosis prediction.

The paper's unique approach of focusing on a specific label, overall survival months, offers a granular perspective on breast cancer prognosis. The exclusion of genetic attributes, while limiting in some respects, allows for a focused exploration of the predictive power of clinical attributes. However, a discussion on the potential advantages of integrating genetic data could provide a more holistic view.

The Random Forest model's superior performance in accounting for 44% of the variance in the testing dataset is commendable. Yet, the paper could benefit from exploring other advanced machine learning models to potentially improve predictive accuracy. A comprehensive evaluation of the model's performance in real-world clinical settings, including its potential impact on treatment decisions and patient outcomes, would enhance the study's practical relevance.

Breast cancer recurrence or the emergence of a secondary primary tumor is a significant concern for survivors. This study focuses on predicting the likelihood of patients diagnosed with first-episode breast cancer developing a second primary tumor (Wang and Fan, 2022). By applying eight distinct machine learning algorithms to data from the SEER database, the research identifies the Random Forest model as the most predictive. The study's emphasis on predicting dual primary tumors offers a fresh perspective in breast cancer research. The comprehensive use of multiple machine learning models provides a robust evaluation, with the Random Forest model's accuracy increase from 63.25% to 97.19% being particularly noteworthy. Relying solely on the SEER database, while providing a vast amount of data, might introduce certain biases. Incorporating data from different sources or regions could offer a more balanced and generalizable model. The paper

could delve deeper into the clinical implications of its findings, exploring potential interventions or monitoring strategies for high-risk patients.

Breast cancer survivors often grapple with a range of health challenges post-treatment, with insomnia being a prevalent issue. The study by Uneo *et al.* (2021), based on a nationwide survey in Japan, underscores the high incidence of insomnia among breast cancer survivors, pegged at 37.5%. By leveraging machine learning algorithms, the research aims to predict comorbid insomnia, offering a potential avenue for early intervention and targeted support.

The paper's focus on insomnia among breast cancer survivors offers a nuanced understanding of the post-treatment challenges faced by patients. The reported prevalence of insomnia is alarmingly high, underscoring the need for holistic post-cancer care that addresses both physical and mental health challenges. The study's use of both the L2 penalized logistic regression model and the XGBoost model provides a comprehensive analysis, though the performance metrics of each model are closely matched. While the study offers valuable insights into the Japanese population, it would be beneficial to explore the prevalence and predictors of insomnia among breast cancer survivors in other cultural or geographical contexts. The emphasis on routine screening for sleep problems is a crucial takeaway, highlighting the importance of comprehensive post-treatment care.

Predicting breast cancer prognosis is a complex task, with various models and methods vying for supremacy in terms of accuracy and clinical relevance. This study by Xiao *et al.* (2021) compares traditional Cox models with contemporary machine learning methods to determine the most effective approach for breast cancer prognostic prediction. The research reveals the RSF model's superior performance, emphasizing the potential of machine learning in enhancing predictive accuracy.

The paper's comparative approach offers valuable insights into the evolving landscape of breast cancer prognosis prediction. The study's comprehensive comparison between traditional Cox models and modern machine learning methods provides a clear benchmark for future research in the field. The RSF model's standout performance offers a promising avenue for further exploration and potential clinical application. While the study provides a robust evaluation, it could delve deeper into the reasons behind the RSF model's superior performance, offering insights into its potential advantages over other methods.

The paper could also explore the real-world implications of its findings, discussing how improved predictive models might influence treatment decisions and patient outcomes.

Breast cancer diagnosis often relies on the analysis of medical images, and this study seeks to harness the power of machine learning to enhance the diagnostic process. The research by Apporva, Yogish and Chayadevi (2021) employs Convolution Neural Networks (CNNs) for image datasets and a range of machine learning algorithms, including KNN, Decision Tree, SVM, and Naïve Bayes, for numerical datasets derived from digitized images of breast masses. The overarching goal is to refine the accuracy of breast cancer predictions. The study's multi-pronged approach, integrating both image and numerical data, offers a comprehensive strategy for breast cancer prediction. The use of CNNs for image datasets is a forward-thinking approach, tapping into the potential of deep learning for medical image analysis.

While the study provides a range of machine learning algorithms for numerical data, a deeper dive into the performance metrics of each model would offer clearer insights into their respective strengths and weaknesses. The integration of numerical data derived from digitized images adds another layer of depth to the analysis, though the study could benefit from elaborating on the

extraction and preprocessing of this data. Future research could explore the potential of combining the predictions from both image and numerical data for an even more robust diagnostic tool.

Tapak et.al. (2019) used machine learning to predict breast cancer outcomes, covering survival and metastasis. Their study aimed to aid clinicians with better tools for patient care. The research evaluated various techniques, seeking the best predictors for these outcomes. While the study's focus on survival and metastasis gives a comprehensive view, clearer insights could come from deeper performance metric analysis of each model. Predicting metastasis is crucial for treatment decisions. Though valuable, the study could benefit from a larger dataset for validation. Future work might improve accuracy by combining clinical and genetic data.

Neoadjuvant chemotherapy (NAC) is a cornerstone of breast cancer treatment, and predicting patient response to it is crucial for optimizing outcomes. The study by Tahmasebbi *et. al.* (2019) integrates machine learning with multiparametric magnetic resonance imaging (mpMRI) of the breast to predict early responses to NAC and subsequent survival outcomes. The research aims to provide a more accurate tool for clinicians, enabling personalized treatment plans based on predicted responses.

The integration of machine learning with mpMRI offers a cutting-edge approach to breast cancer management. The use of mpMRI provides detailed imaging data, which, when combined with machine learning, has the potential to revolutionize the prediction of NAC response. The study's focus on early prediction is crucial, as timely interventions can significantly influence patient outcomes. While the research offers a promising approach, it would be beneficial to validate the findings with a larger and more diverse patient cohort. Future studies could explore the integration of other imaging modalities or genetic data to further enhance the predictive power of the model.

Gass et al. (2018) studied triple-negative breast cancer's response marker. Valuable insights were gained, but the retrospective design introduces potential biases, affecting reliability. Findings need validation through prospective studies. While G3 tumor focus is important, a broader exploration of tumor grades would provide a more complete understanding of the disease, potentially revealing overlooked insights.

Li and Chen (2018) offered a detailed comparison of diverse machine learning models tailored for breast cancer prediction. Their meticulous use of multiple datasets and performance metrics lends credibility to their findings. However, while they highlighted the superior performance of the random forest model, the study could benefit from exploring more advanced machine learning techniques and sophisticated feature engineering methods. Such exploration could pave the way for even more accurate predictions and potentially uncover nuances missed by traditional models.

Cain et al. (2018) explored the potential of machine learning models in predicting pathologic response using MRI features. Their innovative approach of utilizing dynamic contrast-enhanced MRI is commendable. However, the study could reach new heights by integrating a myriad of imaging modalities and amalgamating clinical data. Such integration could potentially refine prediction accuracy and provide a more holistic view of patient health.

Hameed et al. (2018) introduced the Learning Vector Quantization Neural Network Technique for breast cancer prognosis. Their novel approach, while promising, warrants a more granular analysis of the model's performance metrics. Furthermore, validation on a diverse array of datasets is crucial to ascertain the model's robustness and wide-scale applicability. The study could also benefit from comparing this technique with other established machine learning methods to determine its relative efficacy. Also, the study's focus on a single machine learning technique without comparing it to other established methods might limit its comprehensiveness. A

comparative analysis with other machine learning techniques would provide a clearer picture of its relative efficacy.

Kim, Oh, and Ahn (2018) carved a niche with their avant-garde machine learning-based method aimed at pinpointing prognostic biomarker genes crucial for cancer prognosis. Their trailblazing approach, which harnesses the capabilities of generative adversarial networks (GANs) and the PageRank algorithm, is a testament to the potential of machine learning in healthcare. However, a looming question remains: How generalizable is this method across a spectrum of cancer types and datasets? The study could benefit from cross-validation using diverse cancer datasets to ensure its findings are not specific to a particular subset of data.

Gray et al. (2018) conducted an independent validation of the PREDICT tool, an online tool designed for prognostication and treatment benefits for early-stage breast cancer patients. Their validation, conducted on a voluminous dataset from the Scottish Cancer Registry, lends immense credibility to their findings. However, to ensure the tool's universal applicability and robustness, further validation across varied populations is imperative. The study could also delve deeper into the specific factors that contribute to the tool's accuracy and identify areas for refinement.

Wang et al. (2018) have made significant strides with their computational methodology, designed to identify genes showcasing pronounced expression differences between original human tumors and their subsequent xenografts in mouse models. The study's innovative focus on post-transplantation gene expression differences is a fresh perspective in the field. However, to truly unlock the study's potential, a deeper dive into the underlying molecular mechanisms is essential. Additionally, validation on a diverse range of xenograft models would enhance the study's findings and ensure they are not specific to a particular model.

Zeng et al. (2018) have ushered in a new era with their model, which automates the identification of local recurrences in breast cancer patients by leveraging electronic health records (EHRs). Their groundbreaking approach, which marries natural language processing with machine learning, is a testament to the future of healthcare. However, to truly gauge the model's efficacy, its performance in diverse datasets and its real-world applicability need thorough exploration. The study could also benefit from a detailed analysis of the challenges faced in integrating natural language processing with clinical data and potential solutions to these challenges.

Zhao et al. (2018) embarked on a meticulous secondary analysis of data sourced from the METABRIC consortium. Their research underscores the immense potential of machine learning in predicting survival outcomes in breast cancer patients. However, to truly harness the study's potential, further validation on diverse datasets is crucial. Additionally, an exploration of alternative dimensional reduction techniques could provide more nuanced insights and potentially uncover patterns missed by traditional methods.

Dimitriou et al. (2018) showcased a machine learning framework tailored for stage II colorectal cancer patients. Their data-driven approach is commendable, but validation on larger and more diverse datasets would enhance the study's findings. The study could also benefit from a detailed analysis of the specific factors that contribute to the model's accuracy and identify areas for refinement.

Fu et al. (2018) pioneered a machine learning approach tailored to detect lymphedema among breast cancer survivors. Their real-time detection methodology is groundbreaking and holds the promise of timely interventions, drastically improving patient outcomes. However, to truly unlock the study's potential, further validation on diverse datasets is essential. The study could also delve deeper into the challenges faced in real-time detection and potential solutions to these challenges.

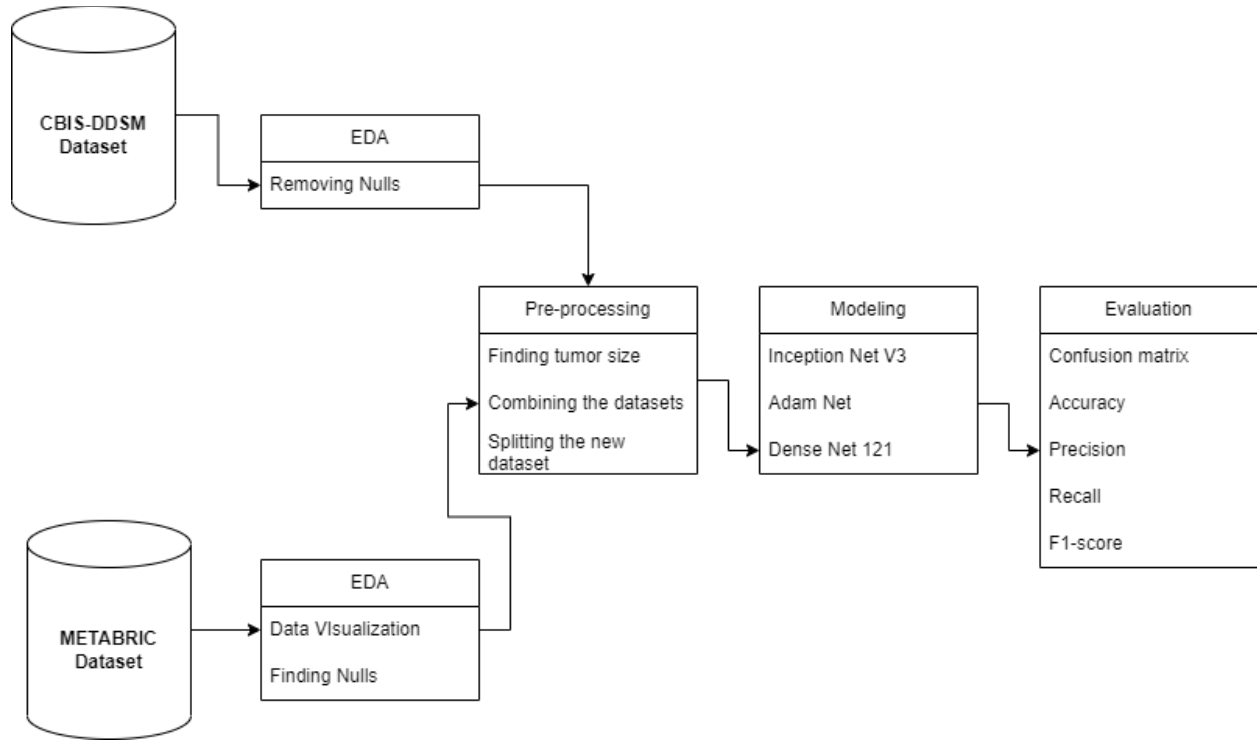
Bai et al. (2018) delved deep into the influence of peripheral lymphocytes on the prognosis of breast cancer patients. Their study, which leverages the Support Vector Machine (SVM) method for prognosis prediction, is groundbreaking. However, to truly harness the study's potential, further validation on diverse datasets is crucial. Additionally, an exploration of alternative machine learning techniques could provide more nuanced insights and potentially uncover patterns missed by traditional methods.

### **2.3 Conclusion**

The literature review highlights the complex characteristics of breast cancer and the progression of predictive methodologies from conventional approaches to more sophisticated computer methods. One notable observation is to the promising prospects of multi-omics data, specifically the amalgamation of gene expression data with mammography pictures, which presents a holistic viewpoint on breast cancer. Nevertheless, the process of integration presents certain difficulties, which in turn require the implementation of advanced algorithms to ensure the efficient utilization of data.

### 3 Methodology

This section of the report presents the methodology adopted for the study. It thoroughly explains the steps taken in order to answer the research question posed. Figure 3.1 below depicts the steps incorporated in the methodology.



**Figure 3.1: Methodology flow**

#### 3.1 Data Collection

The foundation of this research is built upon two pivotal datasets: CBIS-DDSM and METABRIC.

##### 3.1.1 CBIS-DDSM

The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is a subset of a much larger Digital Database for Screening Mammography (DDSM) dataset (Lee *et. al.*, 2017). The CBIS-DDSM dataset contains a diverse collection of mammographic images, including normal, benign, and

malignant cases with verified pathology information. With a total of 1,566 participants, this dataset provides a comprehensive representation of breast cancer cases (Lee *et. al.*, 2017). In addition to the mammogram images, CBIS-DDSM also includes metadata attributes that offer valuable insights into each image. These attributes range from basic patient information, such as birth date and accession number, to technical details such as the number of columns in the image and the date of content creation (Lee *et. al.*, 2017). This wealth of metadata enhances the analysis process by providing contextual information and facilitating the categorization and understanding of each mammogram. The dataset consists of multiple files that contain mammogram images, in-depth information about the patients, the files, regions of interest, and processed images. List of files in the dataset are listed in Table 3.1 below.

<b>Dataset File</b>	<b>Significance</b>
<b>dicom_info.csv</b>	Contains image file names
<b>mass_case_description_train_set.csv</b>	Training set for breast images containing mass
<b>mass_case_description_test_set.csv</b>	Testing set
<b>calc_case_description_train_set.csv</b>	Training set for breast images containing calcification
<b>calc_case_description_test_set.csv</b>	Testing set

**Table 3.1: Dataset files and their contents**

There are in total, 10239 image files present in the CBIS belonging to different classes. The image files include ROI images that denote the tumor location, and cropped images that are subsets of the original images corresponding to the presence of tumor.

### **3.1.2 METABRIC**

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset provides patient-specific gene expressions and outcomes (Curtis *et. al.*, 2012). This dataset focuses

specifically on breast cancer patients and contains a wealth of information for each individual record. Some of the most important attributes in the dataset are tabled below (Table 3.2).

<b>Attribute</b>	<b>Description</b>
<b>Cancer type details</b>	Details the type of Breast Cancer
<b>Age at diagnosis</b>	The age at which the disease is diagnosed
<b>HER2 status</b>	Change in the HER2 receptor
<b>Tumor Histological Subtype</b>	Subtype of the cancer
<b>Tumor Stage</b>	Stage of the tumor
<b>Tumor Size</b>	Size of the tumor
<b>Overall Survival Status</b>	Status of the patient (deceased or alive)

**Table 3.2: Attributes in METABRIC dataset**

The dataset consists of 1980 samples with over 693 attributes. One key attribute in METABRIC is "death\_from\_cancer," which indicates whether the patient succumbed to the disease (Curtis *et. al.*, 2012). This is the attribute that will be used as a pivot for predicting the outcome of the patients based on the mammographic images present in the CBIS. By combining the image data from CBIS-DDSM with the patient outcomes from METABRIC, researchers can gain a comprehensive understanding of the disease and its effects.

In summary, the CBIS-DDSM dataset provides a rich collection of mammographic images accompanied by detailed metadata attributes, enabling in-depth analysis and understanding of each mammogram. The METABRIC dataset, on the other hand, focuses on patient-specific data and provides their outcomes, allowing for the correlation of image data with tangible patient outcomes (Curtis *et. al.*, 2012). By leveraging these datasets in conjunction, researchers can paint a

comprehensive picture of breast cancer, from the initial detection through to the ultimate patient outcomes. This integration of image data and patient outcomes facilitates a holistic and informed approach to breast cancer research.

### 3.2 Exploratory Data Analysis

The expedition of data analysis commenced with an Exploratory Data Analysis (EDA). The initial phase of EDA involved a cursory inspection of the CBIS-DDSM dataset by loading its CSV file into the pandas dataframe. This step provided a snapshot of the data's structure and content (shown in figure 3.2).

	file_path	image_path	AccessionNumber	BitsAllocated	BitsStored	BodyPartExamined	Columns	ContentDate	ContentTime	ConversionType
0	DDSM/dicom/1.3.6.1.4.1.9590.100.1.2.12930... CBIS- DDSM/1.3.6.1.4.1.9590.100.1.2.129308...	DDSM/1.3.6.1.4.1.9590.100.1.2.129308...	NaN	16	16	BREAST	351	20160426	131732.685	WSD
1	DDSM/dicom/1.3.6.1.4.1.9590.100.1.2.24838... CBIS- DDSM/1.3.6.1.4.1.9590.100.1.2.248386...	DDSM/1.3.6.1.4.1.9590.100.1.2.248386...	NaN	16	16	BREAST	3526	20160426	143829.101	WSD
2	DDSM/dicom/1.3.6.1.4.1.9590.100.1.2.26721... CBIS- DDSM/1.3.6.1.4.1.9590.100.1.2.267213...	DDSM/1.3.6.1.4.1.9590.100.1.2.267213...	NaN	16	16	BREAST	1546	20160503	111956.298	WSD
3	DDSM/dicom/1.3.6.1.4.1.9590.100.1.2.38118... CBIS- DDSM/1.3.6.1.4.1.9590.100.1.2.381187...	DDSM/1.3.6.1.4.1.9590.100.1.2.381187...	NaN	16	16	BREAST	97	20160503	115347.770	WSD
4	DDSM/dicom/1.3.6.1.4.1.9590.100.1.2.38118... CBIS- DDSM/1.3.6.1.4.1.9590.100.1.2.381187...	DDSM/1.3.6.1.4.1.9590.100.1.2.381187...	NaN	8	8	Left Breast	3104	20160503	115347.770	WSD

**Figure 3.2: Snapshot of the 'dicom\_info.csv' file in CBIS-DDSM dataset (Source: Notebook)**

A pivotal aspect of EDA is the identification and handling of null or missing values<sup>2</sup>. A thorough scan of the dataset revealed the presence and distribution of null values across columns, highlighting potential gaps in the data. These nulls are removed from the dataset by filling in 'missing' as values. This is done in the 'SeriesDescription' attribute in the dicom\_info.csv file.

<sup>2</sup> <https://www.analyticsvidhya.com/blog/2021/08/exploratory-data-analysis-and-visualization-techniques-in-data-science/>

Figure 3.3 below shows the presence of null values in the CBIS-DDSM dataset.

```
file_path 0
image_path 0
BitsAllocated 0
BitsStored 0
BodyPartExamined 0
ConversionType 0
HighBit 0
LargestImagePixelValue 0
Laterality 566
Modality 0
PatientID 0
PatientName 0
PatientOrientation 0
PhotometricInterpretation 0
PixelRepresentation 0
SamplesPerPixel 0
SecondaryCaptureDeviceManufacturer 0
SecondaryCaptureDeviceManufacturerModelName 0
SeriesDescription 566
SeriesInstanceUID 0
SmallestImagePixelValue 0
SpecificCharacterSet 0
dtype: int64
```

**Figure 3.3: Presence of null values in the dicom\_info.csv file (Source: Notebook)**

To further understand the balance or imbalance of the dataset, a bar plot visualized the distribution of classes in the cropped images dataset. This step was crucial in determining if subsequent class balancing techniques would be required during the preprocessing phase.

### 3.3 Pre-processing

#### 3.3.1 Finding The Common Attribute

The objective of the study required the combination of two distinct datasets: CBIS-DDSM and METABRIC. The decision to merge these datasets was driven by the absence of patient outcomes in the CBIS-DDSM dataset, a crucial piece of information that METABRIC provides. To successfully merge these datasets, a common attribute was necessary, which was identified as 'tumor size'.

### **3.3.2 Determining tumor size in the CBIS-DDSM dataset**

Interestingly, this dataset does not explicitly provide tumor sizes. However, it does offer ROI (Region of Interest) mask images that can be leveraged to deduce the dimensions of the tumor. Upon meticulous examination of the dataset's images, it was discovered that the resolution was 96 dots per inch (dpi). This dpi value indicates that each inch of the image encompasses 96 pixels per square inch of the image. By utilizing unit conversion, this measurement can be translated into square millimeters, allowing for the determination of tumor size.

Moving on to the interpretation of ROI mask images, these images have a binary nature. Tumors are distinctly represented using white pixels, while the rest of the image is rendered in black. To calculate the tumor's area, the white pixels in these ROI mask images are counted and subsequently converted this count into square millimeters. This process helps to obtain a more practical tumor size.

After deducing the tumor sizes from the CBIS-DDSM dataset, the next step involves integrating them with the METABRIC dataset. This integration required categorizing the tumor sizes from METABRIC into distinct bins. These bins were determined using quantiles, specifically: the minimum tumor size, 0.25, 0.5, 0.75, and the maximum tumor size. By having the tumor sizes categorized, it became feasible to extract the corresponding patient statuses, providing valuable insights for further analysis.

The process of annotation and image utilization followed the binning process. The most prevalent class within each bin was identified, and the cropped images from the CBIS-DDSM dataset were annotated based on these identified classes (see Figure 3.4). These annotated images formed the foundation for the survival prediction, which stood as the primary objective of this research endeavor.

	tumor_size	tumor_size_window	death_from_cancer_encoded
0	22.0	Q2	0.0
1	10.0	Q1	0.0
2	15.0	Q1	1.0
3	25.0	Q3	0.0
4	40.0	Q4	1.0
...	...	...	...
1899	25.0	Q3	0.0
1900	20.0	Q2	1.0
1901	25.0	Q3	1.0
1902	25.0	Q3	-1.0
1903	20.0	Q2	-1.0

1904 rows x 3 columns

**Figure 3.4: Binning the tumor size in METABRIC**

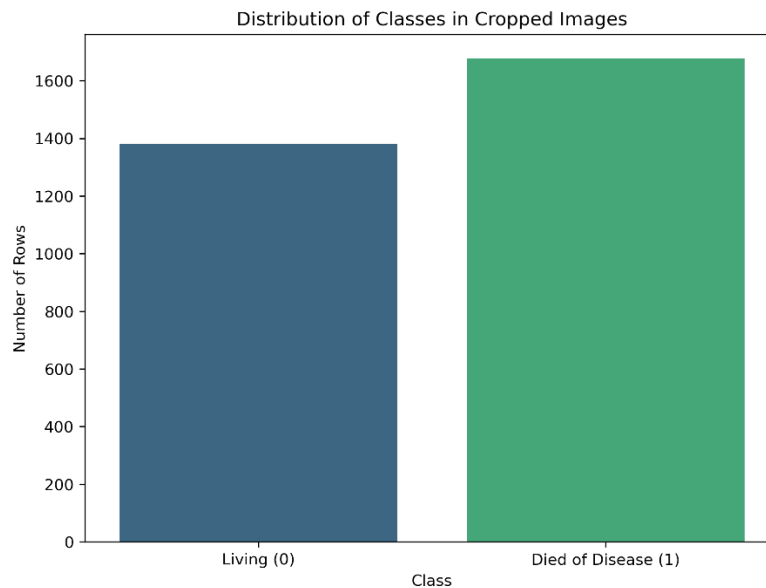
By combining the CBIS-DDSM and METABRIC datasets and leveraging the information obtained through tumor size determination, interpretation of ROI mask images, and integration with METABRIC, valuable insights can be gained for survival prediction and further analysis.

### 3.3.3 Resampling of the classes

After this combination of CBIS-DDSM and METABRIC, the final distribution of the classes is obtained. It is observed from the distribution that the number of images belonging to the class that represents deceased patient are higher compared to those alive. Hence this calls for resampling of the classes.

The resampling of the classes is performed by under sampling the majority class. Hence the final distribution of the images involves there are 1380 images belonging to each of the classes making the combined count to be 2760 images.

Figure 3.5 below shows the distribution of classes before resampling.



**Figure 3.5: Distribution of Classes**

### 3.3.4 Splitting the Data

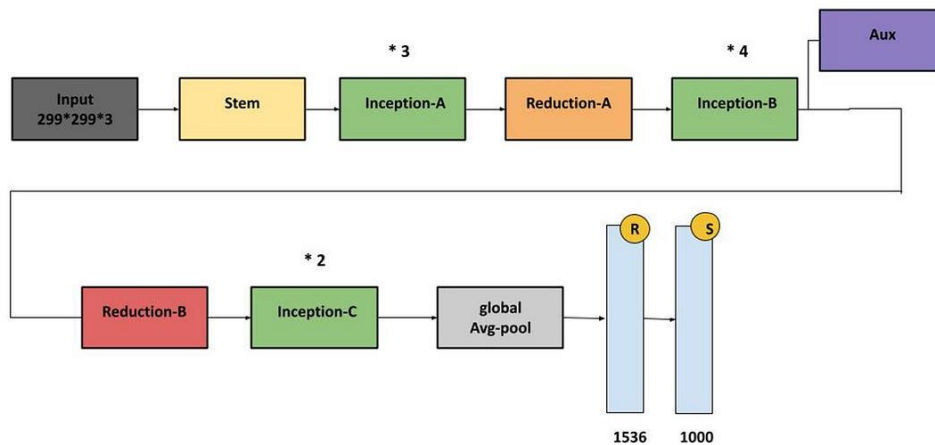
Once a class-balanced dataset is created it is subjected to splitting. This is done so that one part of the data will be used for model training and the remaining part will be used for testing them. For the study, the data is split into a ratio of 80:20 meaning that 80% of the images will be used for training the models and remaining 20% will be used for the testing and evaluation purposes.

### 3.4 Modelling

Once the data representing the life status corresponding to each cropped image in the dataset is obtained, the data is then subjected to modelling. The following are the models implemented in this study. They are discussed in great detail along with the theory and workings of each model thoroughly presented.

### 3.4.1 Inception Net

The Inception Network, often known as GoogLeNet, is a highly esteemed deep convolutional neural network architecture that was built by the research team at Google (Szegedy *et. al.*, 2017). A notable advancement of the network is the Inception Module, which facilitates the network's ability to select from varying convolution sizes inside a single layer (Szegedy *et. al.*, 2017). The ability to adapt and adjust enables the extraction of features at different scales in a manner that is both efficient and effective. The Inception v3 iteration incorporated various innovations, including factorized convolutions and the integration of batch normalization, resulting in enhanced performance and increased accuracy. The GoogLeNet architecture has made a substantial contribution to the field of computer vision by advancing the limits of image categorization and object recognition tasks (Szegedy *et. al.*, 2017). Figure 3.6 shows the architecture of the Inception Net V3.



**Figure 3.6: Architecture of the Inception Net V3 model (Source: [www.medium.com](http://www.medium.com))**

Following are the main blocks of the Inception V3 Architecture.

#### 3.4.1.1 Stem block

The stem component, serving as the initial feature extraction module inside the Inception v3 architecture, assumes a pivotal role (Szegedy *et. al.*, 2017). The task of converting the input data into a format that may be effectively processed by the later Inception modules falls under its responsibility. The stem is comprised of multiple components that collaborate in order to accomplish this objective.

The stem initiates with a sequence of convolutional layers that utilize diverse filter sizes. In general, the convolutional layers in this context often commence with larger filters, such as those with dimensions of  $7 \times 7$ , and subsequently go towards smaller filters, such as those measuring  $3 \times 3$ . This configuration enables the stem to effectively capture the fundamental patterns inherent in the incoming data.

After the application of convolutional layers, it is common to utilize max-pooling layers in order to decrease the spatial dimensions of the feature maps. The aforementioned reduction is advantageous due to its ability to decrease the computing demands of the future layers.

To enhance the stability of activations and accelerate the training procedure, it is possible to employ batch normalization subsequent to the convolutional layers.

#### 3.4.1.2 Inception Block

Moving on from the stem, the Inception modules assume a prominent role inside the Inception v3 framework. The modules have been specifically developed to concurrently record information at many scales, hence facilitating a thorough analysis of the input data (Szegedy *et. al.*, 2017).

In each Inception module, parallel convolutions are employed, utilizing distinct filter sizes including  $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ . The utilization of a parallel convolutional technique facilitates the

network's ability to capture spatial information of varying extents, hence enhancing its capacity to comprehensively comprehend the input data.

In addition to the convolutions, pooling operations, specifically max-pooling, are concurrently executed within each Inception module.

The concurrent convolutions and pooling operations yield separate outputs, which are subsequently merged by concatenation. This process generates a unified feature map that encompasses information from many scales. The integration of these characteristics augments the network's capacity to identify and comprehend intricate patterns within the input data (Szegedy *et. al.*, 2017).

#### 3.4.1.3 Auxiliary Classification Block

To ensure effective gradient flow during training, auxiliary classifiers are introduced in the middle of the network. Unlike traditional classifiers placed at the end, these auxiliary classifiers are strategically positioned in intermediate layers.

Each auxiliary classifier consists of an average pooling layer, a convolutional layer, a fully connected layer, and a SoftMax layer. These classifiers contribute to the total loss during training, facilitating the propagation of gradients through the network (Szegedy *et. al.*, 2017).

However, during inference or when using the model for predictions, these auxiliary classifiers are discarded, and only the main path of the network is utilized.

#### 3.4.1.4 Final Layers

Finally, the architecture concludes with a series of layers that produce the final predictions. Instead of employing fully connected layers, Inception v3 utilizes global average pooling to reduce the

spatial dimensions of the feature maps. This approach reduces the number of parameters and helps prevent overfitting.

To further address overfitting, a dropout layer is incorporated. This layer randomly sets a fraction of input units to 0 at each update during training, encouraging the network to learn more robust and generalizable features.

The final layer of the architecture is a SoftMax layer that converts the raw output scores from the preceding layers into a probability distribution over the different classes. This distribution represents the model's prediction probabilities for each class.

Implementation: Figure below shows the implementation of the Inception model.

```
# Initialize the InceptionV3 model with input shape and without the top layer
base_model = InceptionV3(input_shape=(150, 150, 3), include_top=False)

# Freeze the layers in the base model
for layer in base_model.layers:
    layer.trainable = False

# Add custom layers
x = layers.Flatten()(base_model.output)
x = layers.Dense(128, activation='relu')(x)
x = BatchNormalization()(x)
x = layers.Dense(64, activation='relu')(x)
x = BatchNormalization()(x)
x = layers.Dropout(0.2)(x)
x = layers.Dense(2, activation='softmax')(x)

# Create the final model
Incep_model = Model(base_model.input, x)

# Display the model's architecture
Incep_model.summary()

# Compile the model
Incep_model.compile(optimizer=Adam(lr=0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
```

**Figure 3.7: Implementation of Inception Net V3 (Source: Notebook)**

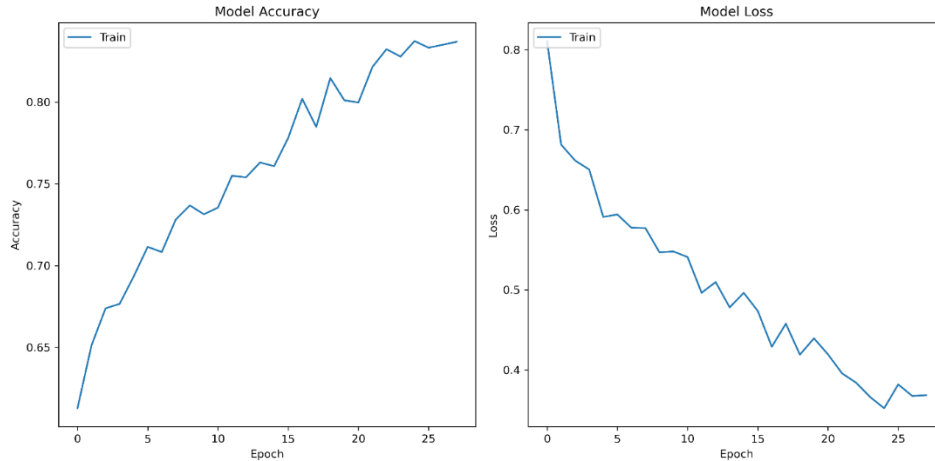
The given code initializes the InceptionV3 model specifically designed for image inputs of dimensions 150x150 pixels and 3 color channels representing red, green, and blue (RGB). In order to facilitate specific adjustments, the uppermost layer of the model, which is generally a fully connected layer, is omitted. Following this, all layers of the basic model are "frozen", indicating

that their weights remain unaltered during the training process. This process guarantees the preservation of the pre-existing knowledge within the model.

On top of the initial model, multiple customized layers are incorporated. The feature maps undergo a process of flattening, resulting in a one-dimensional vector. Next, the model incorporates two dense layers that are fully connected, accompanied by ReLU activations. These layers are interspersed by batch normalization layers, which serve the purpose of stabilizing and expediting the training process. To address the issue of overfitting, a dropout layer is incorporated into the model architecture. This layer randomly deactivates 20% of its input units during the training process. The ultimate layer consists of a dense layer that utilizes softmax activation, which is specifically designed to categorize inputs into one of two distinct groups.

After establishing the architecture, the model is compiled by employing the Adam optimizer with a predetermined learning rate of 0.0001. The selected loss function is `categorical_crossentropy`, which is well-suited for multi-class classification tasks. The model's performance is assessed using accuracy as the evaluation metric.

The model is then trained for 50 epochs and a batch size of 16. The training process is supported by callbacks to avoid overfitting and obtaining the best model. The training accuracy and the loss per epoch are depicted in figure 3.8 below.



**Figure 3.8: Training performance of Inception Net**

### 3.4.2 Adam net

The AdamNet architecture refers to a sequential convolutional neural network (CNN) that has been specifically created for the purpose of image categorization tasks and employs an Adam optimizer for weight adjustment (Krizhevsky, Sutskever and Hinton, 2012). The process exhibits a linear progression, wherein data is transmitted from the input layer to the output layer without any instances of divergence or convergence.

The initial layer of the network consists of a Conv2D layer including 32 filters, each having dimensions of 3x3. The model incorporates the Rectified Linear Unit (ReLU) activation function, which imparts non-linearity to the system. The input layer of the model requires images with dimensions of 150x150 pixels and 3 channels, which are commonly used to represent RGB colour information.

The subsequent layer after the input layer is a MaxPooling2D layer, which utilizes a 2x2 filter and stride. The function of this layer is to conduct max pooling, which leads to a reduction in the spatial dimensions of the feature maps. This process facilitates the identification and extraction of the

most significant characteristics, while simultaneously minimizing the processing demands for succeeding layers.

```

Model: "sequential"
-----
Layer (type)                Output Shape                Param #
-----
conv2d_94 (Conv2D)          (None, 148, 148, 32)      896
max_pooling2d_4 (MaxPooling (None, 74, 74, 32)      0
2D)
conv2d_95 (Conv2D)          (None, 72, 72, 64)        18496
max_pooling2d_5 (MaxPooling (None, 36, 36, 64)      0
2D)
conv2d_96 (Conv2D)          (None, 34, 34, 128)       73856
max_pooling2d_6 (MaxPooling (None, 17, 17, 128)      0
2D)
flatten_1 (Flatten)         (None, 36992)              0
dense_3 (Dense)             (None, 512)                18940416
dropout_1 (Dropout)        (None, 512)                0
dense_4 (Dense)             (None, 2)                  1026
-----
Total params: 19,034,690
Trainable params: 19,034,690
Non-trainable params: 0

```

**Figure 3.9: Adam Net model summary**

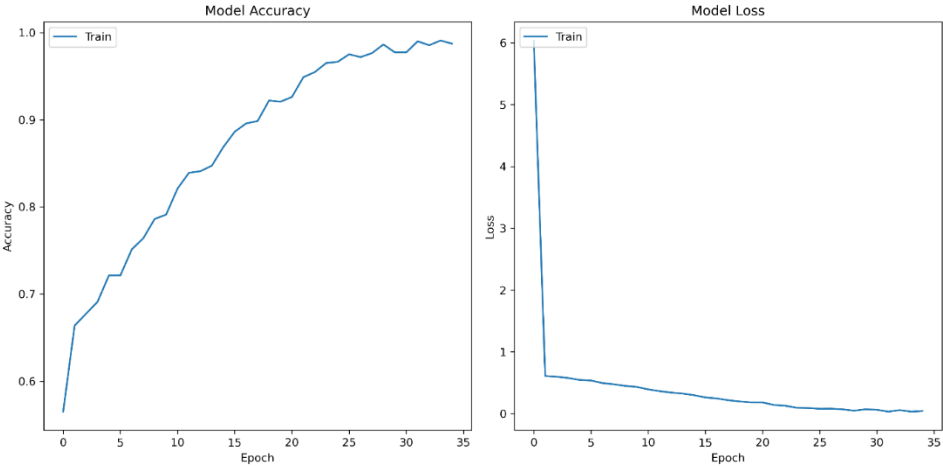
The intermediate convolutional layers consist of two extra Conv2D layers in the architecture. The second layer is composed of 64 filters with dimensions of 3x3 and use the Rectified Linear Unit (ReLU) activation function. It effectively collects intricate characteristics from the aggregated feature maps of the preceding layer. Subsequently, a MaxPooling2D layer is implemented with a filter size of 2x2 and a stride value.

The third convolutional layer is composed of 128 filters with dimensions of 3x3. Additionally, it employs the Rectified Linear Unit (ReLU) activation function. As the network progresses, there is typically an augmentation in the quantity of filters, so enabling the network to effectively capture increasingly complex patterns and information. The architectural design culminates in a concluding layer of MaxPooling2D.

The Flattening layer is employed to transform the three-dimensional feature maps into a one-dimensional vector. The inclusion of this step is necessary due to the requirement of fully connected layers, such as Dense layers, to receive input in a flattened, one-dimensional format.

The fully connected layers consist of a Dense layer with 512 neurons, which is applied after the flattening process. The last layer of the neural network integrates the learned characteristics from preceding layers in order to generate conclusive choices. Furthermore, a Dropout layer is implemented with a dropout rate of 0.5. During each update, Dropout randomly deactivates 50% of its input units, hence aiding in the prevention of overfitting.

The model is then trained for 50 epochs, 1 batch size of 16, and event call-backs. The training accuracy and loss are depicted in figure 3.10 below.



**Figure 3.10: Adam Net Training Performance**

### 3.4.3 Dense Net

DenseNet is a convolutional network that stands out from traditional convolutional networks due to its unique dense connectivity pattern. In DenseNet, each layer receives input from all preceding layers, as opposed to receiving input only from the immediately preceding layer. This dense

connectivity pattern offers several advantages over traditional CNNs (Girdhar, Sinha and Gupta, 2022).

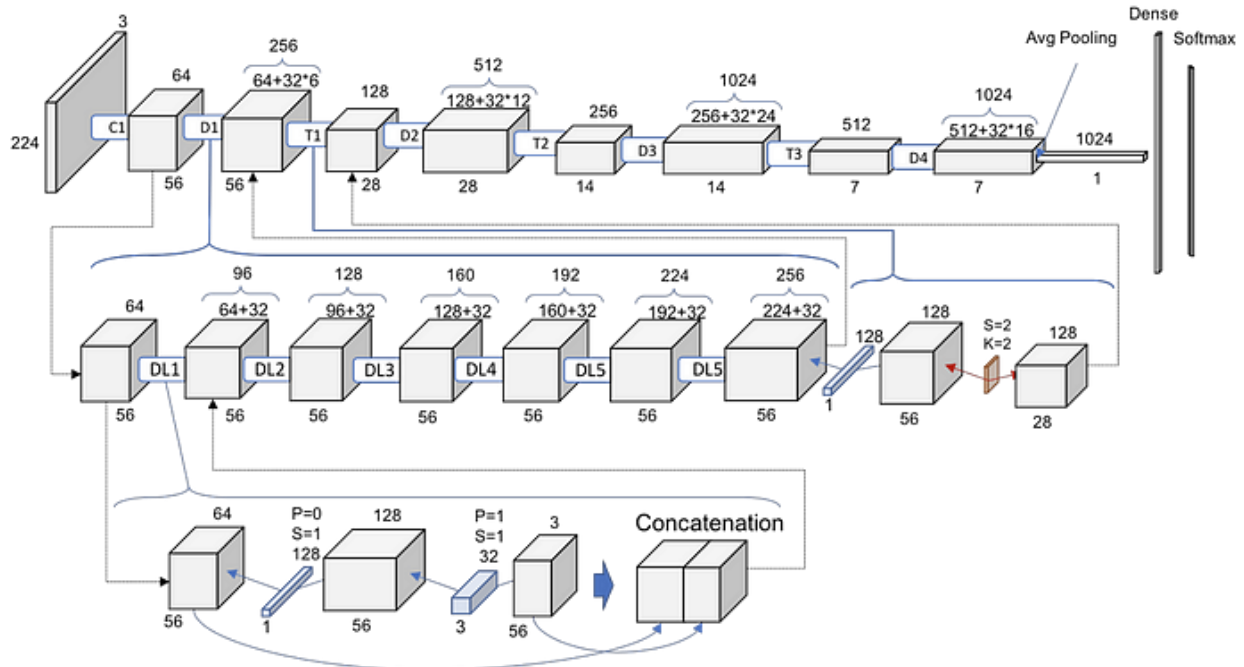
One advantage is parameter efficiency. By receiving additional inputs from all preceding layers, DenseNets allow for feature reuse, which means they can have fewer parameters than traditional CNNs while still achieving superior performance. This parameter efficiency is a significant advantage in deep learning, where reducing the number of parameters can help alleviate computational and memory constraints (Girdhar, Sinha and Gupta, 2022).

Another advantage of DenseNet is improved gradient flow during training. The direct connections from any layer to all subsequent layers ensure a smooth flow of gradients, making it easier to train deeper networks. This is particularly important in deep learning, as training deep networks can be challenging due to the vanishing gradient problem. DenseNet's dense connections help mitigate this problem and enable more effective training of deep networks (Girdhar, Sinha and Gupta, 2022).

DenseNet is also inherently multi-scale, as each layer receives features from all preceding layers. This multi-scale nature allows DenseNets to capture features at various levels of abstraction, enhancing their ability to extract meaningful information from input data (Nandhini and Ashokkumar, 2022).

Furthermore, DenseNets tend to overfit less on tasks with limited data. This is attributed to the improved flow of gradients and the reduced number of parameters. By reducing overfitting, DenseNets can generalize better to unseen data and improve performance on tasks with limited training samples (Nandhini and Ashokkumar, 2022).

Figure 3.11 below shows the architecture of DenseNet121, a specific variant of DenseNet with 121 layers.



**Figure 3.11: DenseNet121 architecture (Source: [www.towardsdatascience.com](http://www.towardsdatascience.com))**

The architecture consists of several key components:

**Initial Convolution:** The network starts with a convolution layer that pre-processes input data before it enters the dense blocks.

**Dense Blocks:** Central to DenseNet, these blocks ensure every layer receives input from all preceding layers, enhancing information flow. Each layer generates  $k$  output feature-maps, termed the growth rate.

**Transition Layers:** Positioned between dense blocks, these layers use pooling to adjust feature-map sizes, streamlining the network's complexity.

**Bottleneck Layers:** For computational efficiency, a bottleneck layer with 1x1 convolutions precedes each 3x3 convolution in the dense blocks.

**Final Layers:** Post the last dense block, a global average pooling condenses feature maps. A subsequent softmax classifier then delivers the final predictions.

One crucial hyperparameter for DenseNets is the growth rate, denoted as  $k$ . The growth rate determines the number of new feature maps produced by each layer within a dense block. In a dense block with  $L$  layers, the total number of feature-maps will be  $k_0 + k * L$ , where  $k_0$  represents the number of channels in the input layer.

DenseNet121 offers several advantages over other deep architectures like ResNet. Firstly, due to its dense connections, DenseNet121 often requires fewer parameters while achieving similar or even better performance. This parameter efficiency can be advantageous in scenarios with limited computational resources.

Secondly, the dense connections in DenseNet121 ensure that features extracted in earlier layers can be reused in later layers. This leads to more efficient feature extraction and can contribute to improved overall performance.

Lastly, the dense connections in DenseNet121 facilitate better gradient flow during training, making the network easier to train. This can eliminate the need for careful initialization and make the training process more stable and efficient.

## Implementation:

The implementation of the DenseNet121 for the curated dataset is shown in figure 3.12 below.

```
# Load the DenseNet121 model with pre-trained weights
base_model = DenseNet121(weights='imagenet', include_top=False)

# Freeze the layers from the base model
for layer in base_model.layers:
    layer.trainable = False

# Add custom layers
x = base_model.output
x = GlobalAveragePooling2D()(x)
x = layers.Dense(128, activation='relu')(x)
x = BatchNormalization()(x)
x = layers.Dense(64, activation='relu')(x)
x = BatchNormalization()(x)
x = layers.Dropout(0.2)(x)
predictions = Dense(2, activation='softmax')(x) # Assuming you have 2 classes

# Create the final model
densenet_model = Model(inputs=base_model.input, outputs=predictions)

# Compile the model
densenet_model.compile(optimizer=Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
```

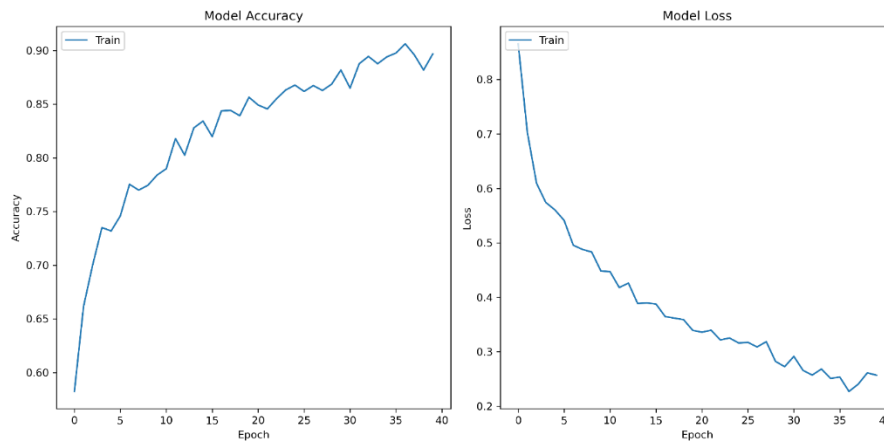
**Figure 3.12: Implementation of DenseNet121 (Source: Notebook)**

The code initiates the DenseNet121 model. This model is loaded with weights pre-trained on the 'imagenet' dataset, ensuring it already possesses knowledge of various image features. To make custom modifications, the top layer, typically a fully connected one, is excluded. All layers in this base model are then "frozen", ensuring their weights remain static during training, preserving the model's pre-trained knowledge.

Building on this foundation, several custom layers are appended. The feature maps are passed through a Global Average Pooling layer, reducing their dimensions while retaining significant information. Subsequent layers include two dense (fully connected) layers with ReLU activations, interspersed with batch normalization layers to stabilize and expedite training. A dropout layer is also integrated, nullifying 20% of its inputs randomly during training to combat overfitting. The final layer is a dense one with softmax activation, tailored for binary classification.

Post architecture configuration, the model is compiled using the Adam optimizer with a set learning rate of 0.0001.

For the training phase, callbacks are defined: an early stopping mechanism monitoring the loss to halt training if no improvement is observed for three consecutive epochs, and a model checkpoint to save the model with the best performance. The model is then trained on the training data for a maximum of 50 epochs with a batch size of 16, leveraging these callbacks. The training's progression, in terms of accuracy and loss across epochs, can be visualized in a subsequent figure 3.13.



**Figure 3.13: Training Performance of the DenseNet121 model (Source: Notebook)**

## 3.5 Evaluation

### 3.5.1 Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It presents the actual vs. predicted classifications. The matrix consists of four values:

True Positives (TP): Correctly predicted positive values.

True Negatives (TN): Correctly predicted negative values.

False Positives (FP): Incorrectly predicted positive values.

False Negatives (FN): Incorrectly predicted negative values.

### **3.5.2 Accuracy**

It is the ratio of correctly predicted instances to the total instances. It provides a general measure of the model's performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

### **3.5.3 Precision**

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It answers the question: Of all the positive labels we've predicted, how many were actually positive?

$$\text{Precision} = \frac{TP}{TP+FP}$$

### **3.5.4 Recall (Sensitivity)**

Recall is the ratio of correctly predicted positive observations to all the actual positives. It answers the question: Of all the actual positive labels, how many did we correctly predict?

$$\text{Recall} = \frac{TP}{TP+FN}$$

### **3.5.5 F1-Score**

The F1-Score is the weighted average of Precision and Recall. It tries to find the balance between precision and recall and is particularly useful when the class distribution is uneven. F1-

$$\text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

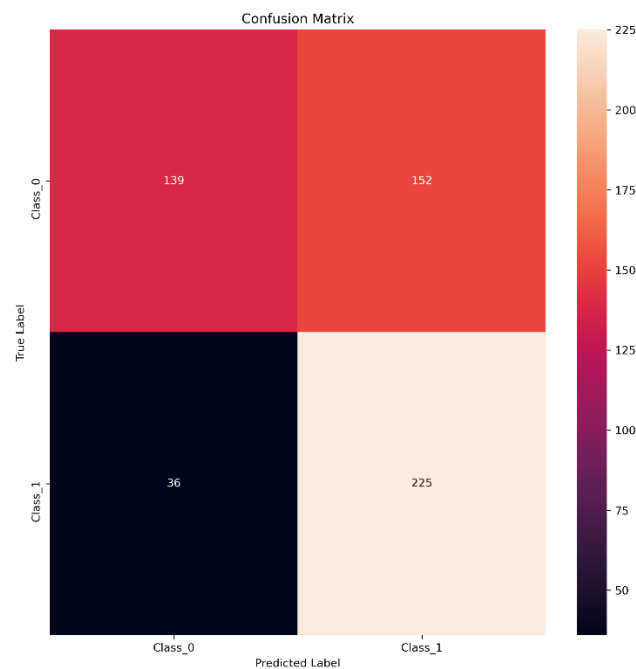
## 4 Results

This section of the report details the findings of the study undertaken. The results obtained from the implementation of the models is discussed below.

### 4.1 Results for the Inception Net

#### 4.1.1 Confusion Matrix

Inception Net appears to struggle in differentiating between Class\_0 and Class\_1. This is evident from the high number of false positives for Class\_0 which is 152, meaning that it wrongly predicted instances as Class\_1 when they were actually Class\_0. This indicates a potential bias towards predicting Class\_1, which could be a limitation of the model. Figure 4.1 shows the confusion matrix for the inception net model.



**Figure 4.1: Confusion matrix for Inception Net (Source: Notebook)**

### 4.1.2 Classification Report

With Inception Net, an accuracy of 66% has been, indicating that this model correctly predicts the class labels for two-thirds of the samples in the test set. However, this accuracy suggests that there is room for improvement. The architecture of Inception Net, with its multiple filters and layers, might not be optimally tuned for this specific dataset, leading to misclassifications. Classification report for the model is shown in figure 4.2 below.

```
Classification Report:
              precision    recall  f1-score   support

   Class_0     0.79      0.48      0.60      291
   Class_1     0.60      0.86      0.71      261

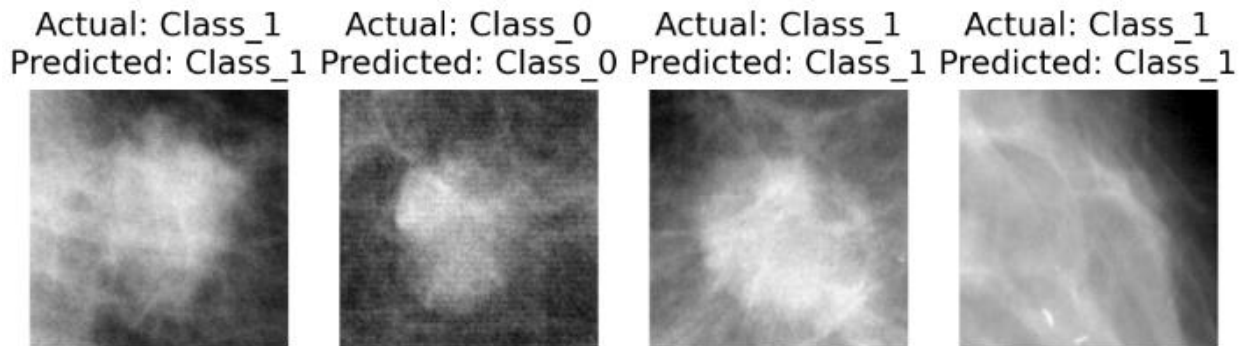
 accuracy              0.66      552
 macro avg           0.70      0.67      0.65      552
 weighted avg        0.70      0.66      0.65      552
```

**Figure 4.2: Classification report for Inception Net (Source: Notebook)**

The performance of Inception Net for Class\_0 that denotes the patient has survived shows that the precision is at 79%, meaning that when the model predicts an instance as Class\_0, it is correct 79% of the time. However, the recall for Class\_0 is only 48%, indicating that the model is only capturing 48% of the actual Class\_0 instances. This means that over half of the true Class\_0 instances are being missed. One possible reason for this discrepancy could be a bias in the model towards predicting Class\_1, resulting in a higher number of false negatives for Class\_0.

Moving on to Class\_1 which denotes that the patient has succumbed to the cancer, the precision for this class is 60%, suggesting that the model's predictions for Class\_1 are less reliable compared to Class\_0. However, the recall for Class\_1 is high at 86%, indicating that the model is capturing a significant majority of the actual Class\_1 instance.

Figure 4.3 below shows the application of Inception net on some of the images in test set.



**Figure 4.3: Subset of test images and their corresponding prediction by Inception Net**

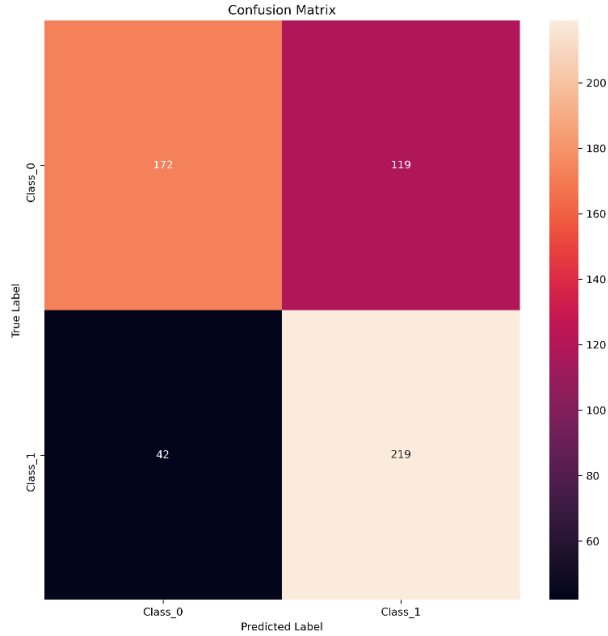
**(Source: Notebook)**

## **4.2 Results for Adam Net**

### **4.2.1 Confusion Matrix**

Adam Net demonstrates better performance compared to Inception Net, particularly in accurately predicting Class\_0 instances. However, it still exhibits a significant number of false positives for Class\_0 amounting to 119. This suggests that although it performs better overall, there is still room for improvement in terms of accurately identifying Class\_0 instances.

The confusion matrix for the model is shown in figure 4.4.



**Figure 4.4: Confusion matrix for Adam Net (Source: Notebook)**

### 4.2.2 Classification Report

This model achieves an accuracy of 71%, which is higher than Inception Net, but still leaves room for improvement. The simpler architecture of Adam Net, as indicated by the provided code, might be better suited for this dataset than the more complex Inception Net. However, further optimization is required to enhance its performance. The classification report for the model is shown in figure 4.5.

```

Classification Report:
              precision    recall  f1-score   support

   Class_0      0.80      0.59      0.68      291
   Class_1      0.65      0.84      0.73      261

 accuracy              0.71      552
 macro avg              0.73      0.72      0.71      552
 weighted avg           0.73      0.71      0.70      552

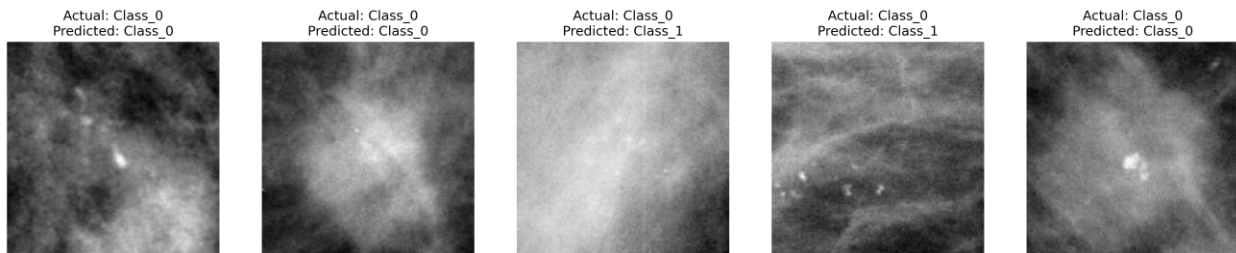
```

**Figure 4.5: Classification report for Adam Net (Source: Notebook)**

Analyzing the performance of Adam Net for Class\_0, we find a precision of 80%, indicating that the model's predictions for Class\_0 are quite reliable. However, the recall for Class\_0 is only 59%, suggesting that the model is missing a significant portion of the true Class\_0 instances. Similar to Inception Net, this could be due to a bias towards predicting Class\_1, resulting in missed Class\_0 instances.

Regarding Class\_1, the precision for Adam Net is 65%, signifying moderately reliable predictions. The recall for Class\_1 is 84%, implying that the model is capturing most of the actual Class\_1 instance. Again, the model's architecture or training data might be favoring features of Class\_1, leading to a higher recall for this class.

Figure 4.6 below shows the application of Adam net on some of the images in test set.



**Figure 4.6: Testing on test data (Source: Notebook)**

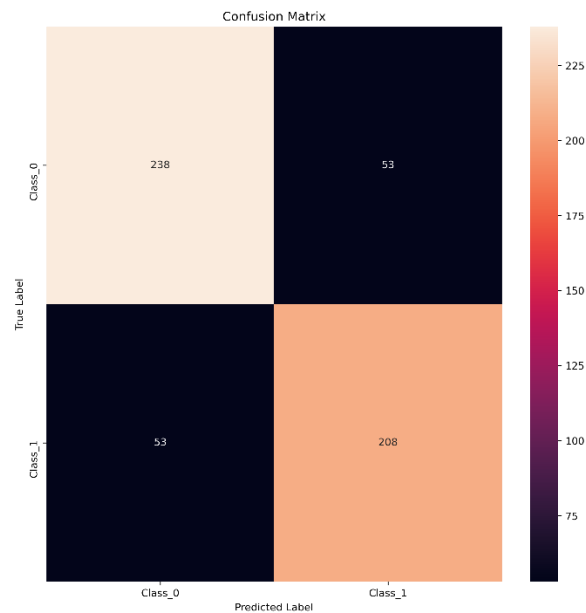
### 4.3 Results for Dense Net

Dense Net outperforms both Inception Net and Adam Net in distinguishing between the two classes.

#### 4.3.1 Confusion Matrix

It achieves the highest number of true negatives and true positives, indicating a more balanced and accurate performance. This suggests that Dense Net might be a more reliable model for predicting

both Class\_0 and Class\_1 instance. Figure 4.7 below depicts the confusion matrix for the Dense Net model.



**Figure 4.7: Confusion matrix for DenseNet121 (Source: Notebook)**

### 4.3.2 Classification Report

This model achieves the highest accuracy among the three, with an accuracy of 81%. The unique architecture of DenseNet, with its dense connections, ensures efficient learning and feature reuse, which can lead to better performance on certain datasets. Classification report for the model is shown in figure 4.8.

```
Classification Report:
      precision    recall  f1-score   support

 Class_0       0.82     0.82     0.82     291
 Class_1       0.80     0.80     0.80     261

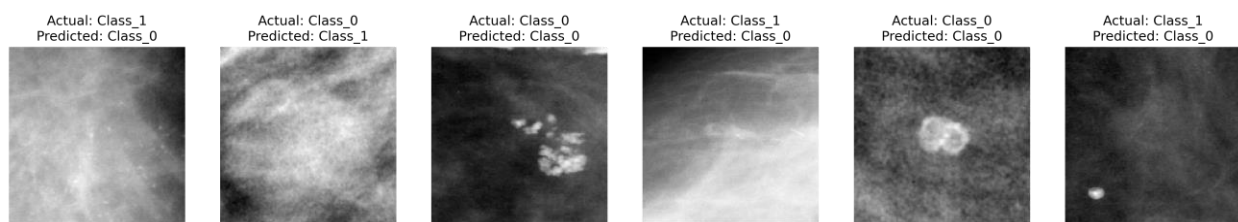
 accuracy              0.81     552
 macro avg       0.81     0.81     0.81     552
 weighted avg    0.81     0.81     0.81     552
```

**Figure 4.8: Classification report for DenseNet121**

Looking at Class\_0 for Dense Net, we observe a precision of 82%, indicating highly reliable predictions. The recall for Class\_0 is also at 82%, suggesting a balanced performance, capturing a significant majority of the true Class\_0 instances.

For Class\_1, the precision remains at 80%, indicating highly reliable predictions. The recall for Class\_1 is 80%, which again suggests a balanced performance. DenseNet's architecture, with its dense blocks and efficient feature propagation, might be ensuring that the model effectively learns features of both classes, resulting in a balanced performance.

Figure 4.9 below shows the application of DenseNet on some of the images in test set.



**Figure 4.9: Testing DenseNet121 on Test Data (Source: Notebook)**

#### 4.4 Conclusion

In conclusion, when comparing the performance of Inception Net, Adam Net, and Dense Net, it is evident that Dense Net offers the highest accuracy. Its unique architecture allows for efficient learning and feature reuse, leading to better performance on this particular dataset.

## 5 Conclusion and Future Work

The core aim of this study was to pioneer a predictive model for assessing the survival probabilities of breast cancer patients. To achieve this, a meticulous integration of multi-omics data from the CBIS-DDSM and METABRIC datasets was undertaken, resulting in a unique dataset that provided tailored prognostic insights derived from mammographic images. Unlike conventional approaches that mainly focused on breast cancer detection using these images, this innovative initiative aimed to predict survival outcomes using the same images, thus forging new ground in this domain.

While previous studies have made significant progress in the detection of breast cancer, the realm of survivability prediction has largely remained unexplored. Our research not only fills this critical gap but also presents a transformative approach to managing breast cancer. Through the utilization of the DenseNet121 architecture, we achieved an impressive accuracy of 81%, highlighting the potential of this model as a reliable tool for predicting survival. This tool has the power to shape personalized treatment regimens, providing healthcare professionals with a nuanced understanding of individual patient trajectories.

Nevertheless, it is crucial to recognize that in the realm of healthcare, the stakes are exceptionally high. Although an accuracy of 81% is commendable, it may not be sufficient for full automation in clinical settings. Therefore, in the interim, this model can be used synergistically alongside systems based on the METABRIC dataset, ensuring a holistic and robust approach to survivability prediction. By integrating these systems, clinicians can gain invaluable insights that facilitate the formulation of tailored treatment strategies, including radiotherapy and chemotherapy, optimized to meet the unique needs of each patient.

## 5.1 Future Work

**Predicting Chances of Survival:** Future advancements should aim to estimate survival probabilities over specific durations, rather than just binary outcomes. This nuanced approach would enable healthcare professionals to tailor treatments more effectively.

**Advanced Model Architectures:** Exploring cutting-edge deep learning architectures and using ensemble techniques could potentially enhance prediction accuracy.

**Feature Engineering:** A deeper dive into datasets, especially METABRIC, might reveal overlooked features that can boost model performance.

**Clinical Trials:** To truly validate the model's efficacy, it's essential to test it in real-world clinical scenarios.

**Feedback Loop:** Incorporating feedback from medical professionals can help continuously refine the model, keeping it updated with the latest research and treatment trends.

## 6 References

1. Arecco L, Blondeaux E, Bruzzone M, Ceppi M, Latocca MM, Marrocco C, Boutros A, Spagnolo F, Razeti MG, Favero D, Spinaci S, Condorelli M, Massarotti C, Goldrat O, Del Mastro L, Demeestere I, Lambertini M. Safety of fertility preservation techniques before and after anticancer treatments in young women with breast cancer: a systematic review and meta-analysis. *Hum Reprod.* 2022 May 3;37(5):954-968.
2. Bai, F., Wei, C., Zhang, P., Bi, D., Ge, M., Chen, Q., Jia, Y., Lu, Y. and Wu, K. (2018). Use of peripheral lymphocytes and support vector machine for survival prediction in breast cancer patients. *Translational Cancer Research*, 7(4).
3. Bakherad, H., Ghasemi, F., Hosseindokht, M. et al. Nanobodies; new molecular instruments with special specifications for targeting, diagnosis and treatment of triple-negative breast cancer. *Cancer Cell Int* 22, 245 (2022).
4. Bhattarai, S., Klimov, S., Aleskandarany, M.A., Burrell, H., Wormall, A., Green, A.R., Rida, P., Ellis, I.O., Osan, R.M., Rakha, E.A. and Aneja, R. (2019). Machine learning-based prediction of breast cancer growth rate in vivo. *British Journal of Cancer*, 121(6), pp.497-504.
5. Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G. and Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. *Cancer medicine*, 9(9), pp.3234-3243.
6. Cain, E.H., Saha, A., Harowicz, M.R., Marks, J.R., Marcom, P.K. and Mazurowski, M.A. (2019). Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast cancer research and treatment*, 173, pp.455-463.

7. Coughlin, S.S. and Ekwueme, D.U. (2009). Breast cancer as a global health concern. *Cancer Epidemiology*, 33(5), pp.315–318.
8. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y. and Gräf, S., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), pp.346-352.
9. Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M.S. and Bergh, J., 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research*, 13(11), pp.3207-3214.
10. Dimitriou, N., Arandjelović, O., Harrison, D.J. and Caie, P.D. (2018). A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *NPJ digital medicine*, 1(1), p.52.
11. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), pp.115-118.
12. Fu, M.R., Wang, Y., Li, C., Qiu, Z., Axelrod, D., Guth, A.A., Scagliola, J., Conley, Y., Aouizerat, B.E., Qiu, J.M. and Yu, G. (2018). Machine learning for detection of lymphedema among breast cancer survivors. *Mhealth*, 4.
13. Gao, Y., Li, S., Jin, Y., Zhou, L., Sun, S., Xu, X., Li, S., Yang, H., Zhang, Q. and Wang, Y., 2022. An Assessment of the Predictive Performance of Current Machine Learning–

Based Breast Cancer Risk Prediction Models: Systematic Review. *JMIR Public Health and Surveillance*, 8(12), p.e35750.

14. Gass, P., Lux, M.P., Rauh, C., Hein, A., Bani, M.R., Fiessler, C., Hartmann, A., Häberle, L., Pretscher, J., Erber, R. and Wachter, D.L. (2018). Prediction of pathological complete response and prognosis in patients with neoadjuvant treatment for triple-negative breast cancer. *BMC cancer*, 18, pp.1-8.
15. Girdhar, N., Sinha, A. and Gupta, S., 2022. DenseNet-II: An improved deep convolutional neural network for melanoma cancer detection. *Soft Computing*, pp.1-20.
16. Gray, E., Marti, J., Brewster, D.H., Wyatt, J.C., Hall, P.S. and SATURNE Advisory Group (2018). Independent validation of the PREDICT breast cancer prognosis prediction tool in 45,789 patients using Scottish Cancer Registry data. *British journal of cancer*, 119(7), pp.808-814.
17. Hameed, W.A., Das, R. and Jaiswal, J. (2018). Breast Cancer Prognosis Using Learning Vector Quantization Neural Network Technique. *International Journal of Engineering & Technology*, 7(4.10), pp.922-924.
18. Hassan, S., 2022. Breast Cancer Screening and Diagnostic Advancements. *Pakistan BioMedical Journal*, pp.02-02.
19. Huang, K., Zhang, J., Yu, Y., Lin, Y. and Song, C. (2022). The impact of chemotherapy and survival prediction by machine learning in early Elderly Triple Negative Breast Cancer (eTNBC): a population based study from the SEER database. *BMC geriatrics*, 22(1), p.268.
20. Hwang, E.S., Lichtensztajn, D.Y., Gomez, S.L., Fowble, B. and Clarke, C.A., 2013. Survival after lumpectomy and mastectomy for early stage invasive breast cancer: the effect of age and hormone receptor status. *Cancer*, 119(7), pp.1402-1411.

21. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E. and Forman, D., 2011. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2), pp.69-90.
22. Kerlikowske, K., Zhu, W., Tosteson, A.N., Sprague, B.L., Tice, J.A., Lehman, C.D., Miglioretti, D.L. and Breast Cancer Surveillance Consortium\*, 2015. Identifying women with dense breasts at high risk for interval cancer: a cohort study. *Annals of internal medicine*, 162(10), pp.673-681.
23. Kim, M., Oh, I. and Ahn, J. (2018). An improved method for prediction of cancer prognosis by network learning. *Genes*, 9(10), p.478.
24. Kumar, D.N., Chaudhuri, A., Aqil, F., Dehari, D., Munagala, R., Singh, S., Gupta, R.C. and Agrawal, A.K., 2022. Exosomes as emerging drug delivery and diagnostic modality for breast cancer: recent advances in isolation and application. *Cancers*, 14(6), p.1435.
25. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
26. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M. and Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1), pp.1-9.
27. Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., Miglioretti, D.L. and Breast Cancer Surveillance Consortium, 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11), pp.1828-1837.
28. Li, Y. and Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math*, 7(4), pp.212-216.

29. Long, F., Ma, H., Hao, Y., Tian, L., Li, Y., Li, B., Chen, J., Tang, Y., Li, J., Deng, L. and Xie, G., 2023. A novel exosome-derived prognostic signature and risk stratification for breast cancer based on multi-omics and systematic biological heterogeneity. *Computational and Structural Biotechnology Journal*, 21, pp.3010-3023.
30. Malik, V., Kalakoti, Y. and Sundar, D., 2021. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *Bmc Genomics*, 22, pp.1-11.
31. Massicano, A.V., Marquez-Nostra, B.V. and Lapi, S.E., 2018. Targeting HER2 in nuclear medicine for imaging and therapy. *Molecular Imaging*, 17, p.1536012117745386.
32. Nandhini, S. and Ashokkumar, K., 2022. An automatic plant leaf disease identification using DenseNet-121 architecture with a mutation-based henry gas solubility optimization algorithm. *Neural Computing*
33. Perou, C.M., Sørlie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A. and Fluge, Ø., 2000. Molecular portraits of human breast tumours. *nature*, 406(6797), pp.747-752.
34. Siu, A.L., 2016. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Annals of internal medicine*, 164(4), pp.279-296.
35. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S. and Thorsen, T., 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19), pp.10869-10874.

36. Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A., 2017, February. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).
37. Varn, F.S., Ung, M.H., Lou, S.K. and Cheng, C., 2015. Integrative analysis of survival-associated gene sets in breast cancer. *BMC medical genomics*, 8(1), pp.1-16.
38. Wang, D., Li, J.R., Zhang, Y.H., Chen, L., Huang, T. and Cai, Y.D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes*, 9(3), p.155.
39. Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y. and Zheng, Y. (2022). The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. *JMIR medical informatics*, 10(2), p.e33440.
40. Yang, H.S., Kwon, S., Lee, S., Lee, S. and Kim, J.Y. (2022). Development of Breast Cancer Prognosis Prediction Model Based on Clinical Features Including CEA and CA15-3 Serum Levels. *Journal of Health Informatics and Statistics*, 47(1), pp.35-47.
41. Yu, K.H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L. and Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7(1), p.12474.
42. Zhong, X., Luo, T., Deng, L., Liu, P., Hu, K., Lu, D., Zheng, D., Luo, C., Xie, Y., Li, J. and He, P. (2020). Multidimensional machine learning personalized prognostic model in an early invasive breast cancer population-based cohort in china: algorithm validation study. *JMIR Medical Informatics*, 8(11), p.e19069.

43. Zhao, M., Tang, Y., Kim, H. and Hasegawa, K. (2018). Machine learning with k-means dimensional reduction for predicting survival outcomes in patients with breast cancer. *Cancer informatics*, 17, p.1176935118810215.
44. Zeng, Z., Espino, S., Roy, A., Li, X., Khan, S.A., Clare, S.E., Jiang, X., Neapolitan, R. and Luo, Y. (2018). Using natural language processing and machine learning to identify breast cancer local recurrence. *BMC bioinformatics*, 19(17), pp.65-74.