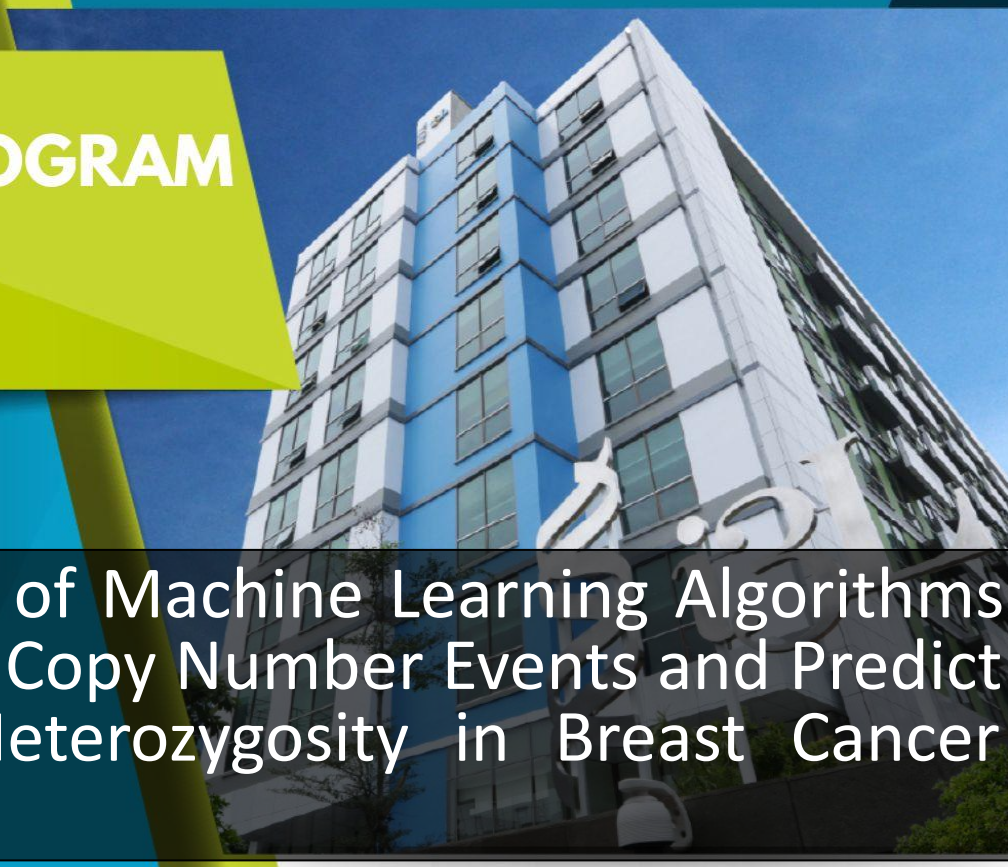
Indonesia International Institute for Life Sciences

# ENRICHMENT PROGRAM REPORT



# Utilization of Machine Learning Algorithms to Classify Copy Number Events and Predict Loss of Heterozygosity in Breast Cancer Patients

STUDY PROGRAM
**Bioinformatics**

JUAN LORELL
21010107

Chih-Yang Wang, Ph.D.
(FIELD SUPERVISOR)
Muammar Sadrawi, Ph.D.
(EP SUPERVISOR)

INDONESIA INTERNATIONAL INSTITUTE FOR LIFE SCIENCES (i3L)
2024

**RESEARCH REPORT**

# Utilization of Machine Learning Algorithms to Classify Copy Number Events and Predict Loss of Heterozygosity in Breast Cancer Patients

By
Juan Lorell
21010107

Submitted to

i3L – Indonesia International Institute for Life-Sciences
School of Life Sciences

in-partial fulfilment of the enrichment program for the Bachelor of Science in
Bioinformatics

Research Project Supervisor: Muammar Sadrawi, B.S., M.S., Ph.D.
Research Project Field Supervisor: Chih-Yang Wang, B.S., M.S., Ph.D.

Jakarta, Indonesia
2024

**INSTITUT BIO SCIENTIA INTERNASIONAL INDONESIA**

Jl. Pulomas Barat Kav. 88 Jakarta Timur 13210 Indonesia
+6221 295 67888, +6221 295 67899, +6221 296 17296
www.i3l.ac.id

## Certificate of Approval

Student : Juan Lorell

Cohort : 2021

Title of Enrichment Program project : Utilization of Machine Learning Algorithms to Classify Copy Number Events and Predict Loss of Heterozygosity in Breast Cancer Patients

We hereby declare that this EP project is from student's own work. The EP Report has been read and presented to i3L's Examination Committee. The EP has been found to be satisfactory and accepted as part of the requirements needed to obtain an i3L bachelor's degree.

Approved by,

EP Advisor

(Muammar Sadrawi, B.S., M.S., Ph.D.)
Date:15/01/2025

Assessor

(Puspa Setia Pratiwi, S.Kom., IM SME, Ph.D.)
Date: 20/01/2025

# COPYRIGHT NOTICE

# STATEMENT OF ORIGINALITY

Submitted to
Indonesia International Institute for Life Sciences (i3L)

I, Juan Lorell, do herewith declare that the material contained in my EP Report entitled:

"Utilization of Machine Learning Algorithms to Classify Copy Number Events and Predict Loss of Heterozygosity in Breast Cancer Patients"

Is an original work performed by me under the guidance and advice of my EP advisor Muammar Sadrawi B.S., M.S., Ph.D. have read and do understand the definition and information on the use of source and citation style published by i3L. By signing this statement, I unequivocally assert that the aforementioned thesis conforms to published information.

i3L has my permission to submit an electronic copy of my thesis to a commercial document screening service with my name included. If you check NO, your name will be removed prior to submission of the document screening.

☑ Yes

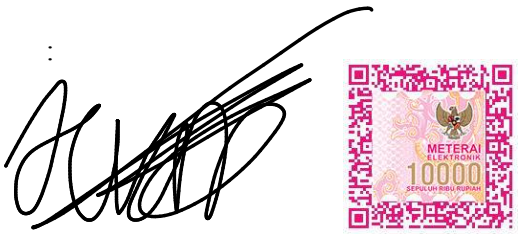
☐ No

Student Name : Juan Lorell

Student ID :21010107

Study Program :Bioinformatics

Signature :




Date : 20 December 2024

# ABSTRACT

Copy number events, specifically copy number aberrations, are occurrences that play an important part in the development of certain cancer types. However, elucidating the correct type of copy number aberration has always been a debated topic with the use of both *in silico* and *in vitro* techniques having their own disadvantages. Machine learning may prove to be an excellent avenue to explore as recent advancements have made it easier to build, study, and apply in the medical field. With it, determining a gene's specific copy number type may be possible to elucidate thus allowing better understanding for possible target therapy. Using pre-established and validated software, elucidated copy number segmentation value was inputted for a regiment of machine learning algorithm. The top five models were selected for hyperparameter tuning with cross validation with the end goal of Ensembl voting while genomics data was visualized to ensure better clarity for data interpretation. Analysis resulted in a clear pipeline for copy number analysis for future data entry to increase the trustworthiness of the model. However, future studies can look into the use of next generation sequencing data, which can offer more coverage of the genome at the cost of higher computational burden.

Keywords: *Breast Cancer*, *Machine Learning*, *Copy Number Analysis*, *Loss of Heterozygosity*

# ACKNOWLEDGEMENTS

First of all, I would like to thank the Lord Almighty for all his blessings and guidance during the whole of my EP Internship in Taipei medical University (TMU). Only by His grace Am I able to complete the report and project in a consistent and timely manner.

I would like to thank my family who has supported me during my time in i3L. Without their help and advice, both financially and emotionally, I would not have been able to experience what I have now. I also would like to express my gratitude to my high school friends; Michael, Gaby, and Yuki who have been with me and helped me when I was feeling down since before i3L. I hope that in the future, our friendship and bonds will be stronger.

Many thanks and gratitude for the immense help and support given to me during the internship by Assoc Prof. Chih-Yang Wang and the Wang lab as without their input and help during this internship, the end result of this report would not be as it is now. I would like to also express gratitude to Assoc Prof. Hsin-Hyi Chen who is a collaborator and pseudo supervisor of the project. Thank you especially to Sachin Kumar who helped me during my transition from Indonesia to Taiwan and guided me throughout my internship, Nguyen Ngouc Trung who became my bench mate when we were still stuck in the storage area of the lab, E Chern Wong who often helped me in buying lunch as for the life of me, I can't speak Chinese, and Gleb Shamrin who taught me the theory behind wet lab whilst also buying me breakfast. Lastly, thank you very much for the eye-opening insight to everyone in TMU that I met as I got to experience a lot here in Taiwan.

However, my lecturers in Indonesia, especially my advisor Muammar Sadrawi, Ph.D's contribution during my study should not be underestimated. Thank you very much from the bottom of my heart for the last 3 years or so in guiding me in both class work and research projects. Thank you to Prof. Dr.rer.nat. Arli Aditya Parikesit who guided me in almost all my research projects in i3L; Rizky Nurdiansyah M.Si. for being my academic advisor for the first two years and also teaching me a lot about genomics analysis thought process which helped a lot in the project; David Agustriawan, Ph.D. who introduced me to cancer dry lab; Nanda Rizqi Pradana M.Stat. and Puspa Setia Pratiwi, Ph.D. for giving me the basics for every statistical analysis so I could keep up with my lab mates work.

Lastly, thank you very much to everyone in BI21 and cohort 21 as well as some of the seniors I had the pleasure of meeting in this short time for the amazing and eye-opening time in i3L. I would like to especially thank Mukti Subagja as without his kindness in letting me borrow his laptop, I would not have been able to work on this project as smoothly. Thank you also again to Michael for helping for proofreading.

# TABLE OF CONTENTS

# LIST OF FIGURES, TABLES, AND ILLUSTRATIONS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **TMU** | Taipei medical University |
| **TCGA** | The Cancer Genome Atlas |
| **BRCA** | Breast Invasive Carcinoma |
| **CNA** | Copy Number Aberration/Alteration |
| **DNA** | Deoxyribose Nucleic Acid |
| **CNVs** | Copy Number Variations |
| **LoH** | Loss of Heterozygosity |
| **pitNETS** | Pituitary Neuroendocrine Tumors |
| **APT** | Affymetrix Power Tools |
| **GEO** | Gene Omnibus Express |
| **LRR** | Log R Ratio |
| **BAF** | Beta Allele Frequency |
| **MLP** | Multi-layer Perceptron |
| **IDC** | Invasive Ductal Carcinoma |
| **ILC** | Invasive Lobular Carcinoma |
| **ER** | Estrogen Receptor |
| **PR** | Progesterone Receptor |
| **HER2** | Human Epidermal Growth Factor 2 |
| **TNBC** | Triple Negative Breast Cancer |
| **SMOTE** | Synthetic Minority Oversampling Technique |
| **ENN** | Edited Nearest Networks |
| **GBM** | Gradient Boost Model |
| **XGB** | Extreme Gradient Boost |

# I. INTRODUCTION

## 1.1. Background

There are 33 cancer types of cancer as classified by the TCGA (The Cancer Genome Atlas) according to the location and tumor type (National Cancer Institute, 2022). Out of those 33, breast cancer or breast invasive carcinoma (BRCA) ranks second among the leading number of cancer diagnoses in the world, with lung cancer being the first (Giaquinto et al., 2022). Current research on breast cancer centers around discovering therapy medication to remove the tumor, however, these medications cause adverse events and in some cases, cause the development of resistance (Grimm, 2023). Specific studies have documented the importance of the morphology of the tumor alongside the variance in treatment methods which caused these adverse events (Masood, 2016). However, with advancements in the field of genomics, potential gene mutations and differential expression of key genes can be elucidated to discover a more effective and safer therapeutic approach (Bennett et al., 2022).

One of the most researched events in cancer is copy number aberrations/alteration (CNAs). This event is a common somatic mutation considered as a pathogenic type of copy number variants (CNVs) that plays a vital role in tumorigenesis and cancer progression (Mallory et al., 2020; Zeira & Raphael, 2020). CNAs are considered a type of somatic mutation as it changes the structure of DNA through amplification or deletion causing change in gene expression which leads to tumorigenesis (Tan et al., 2022).

One of the classification of a CNA event is the loss of heterozygosity (LoH) event where a particular section of the genome becomes homozygous due to a multitude of reasons such as mitotic recombination and loss of chromosome (Naeim et al., 2018). In breast cancer, LoH events can be seen in tumor suppressor genes like BRCA1, where Santana dos Santos et al. (2022) found that LoH is a significant factor to determine the pathogenicity of a breast cancer tumor. This finding is further supported by other studies such as the ones done by Lebok et al. (2015) and Deryusheva et al. (2017) in which a particular gene was reported to be particularly susceptible to LoH or like Tsyganov et al. (2022) where a certain loci is susceptible.

Some research has been done on the utilization of machine learning algorithms to analyze copy numbers for prediction in the field of cancer, such as the ones by Mu and Wang (2021), Rajpal et al. (2023), and Young et al. (2024) among many others. However, those studies focus on all copy number events while specific events are rarely targeted. This can be seen by two researchers that use machine learning algorithms that target LoH specifically. In the study by Pyke et al. (2022), they aimed to predict LoH in the gene HLA in pan-cancer while the other by Lin et al. (2024) the tumor type Pituitary neuroendocrine tumors (pitNET) is the subject for the prediction. However, research regarding LoH on breast cancer has not utilized machine learning algorithms yet while breast cancer is an important cancer to study with their high occurrence rate.

To fill in this gap of knowledge, this study aims to create a pipeline that can predict possible loss of heterozygosity directly from raw microarray data utilizing deep learning algorithms. This model would then be compared against known processing algorithms to determine the effectiveness of the algorithm. To achieve an algorithm with high accuracy, data from current tools were used as the input data which will be used to train the data. The algorithm was trained for several hundreds of iterations using several layers that can be used to find the correlation in the raw data before the weight is validated using another dataset altogether to avoid bias. This is done in the hopes that a novel model would be created that can show and target specific copy number events.

## 1.2. Objective

- Create a pipeline suitable to produce copy number data that could be used for preclinical study and comparison of samples from Affymetrix SNP 6.0 Human microarray data
- Create a novel machine learning model that could classify copy number events with high balanced accuracy

## 1.3. Hypothesis

- It is possible to create a suitable pipeline for sample comparison in preclinical study
- A novel algorithm can be made from copy number data which offers good balanced accuracy

# II. LITERATURE REVIEW

## 2.1. Cancer

As defined by the National Cancer Institute (2021), cancer is an uncontrollable growth of cells in the body which could metastasize to other cells in a single organism. Cancer as a disease has caused millions of fatalities with an increasing number of people being diagnosed every year. This is backed by a news article by the World Health Organization/WHO (2024), which reports that 20 million people were diagnosed and 9.4 million people died. Throughout the years, researchers have constantly endeavored in finding a cure to the disease, leading to many new discoveries and findings.. These findings led to newer definitions of cancer such as one stated in a review article by Brown et al. (2023), proposed a new definition of cancer which is "Cancer is a disease of uncontrolled proliferation by transformed cells subject to evolution by natural selection".

Using this definition, it can be seen that there are many causes of cancer progression and initiation. Those causes have been categorized and grouped into what is known as the hallmarks of cancer in a review paper by Hanahan & Weinberg (2000). Those hallmarks are sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing/accessing vasculature, activating invasion and metastasis. This was later updated in 2011 and 2022 in which the current hallmarks were updated as seen in **Figure 1** (Hanahan, 2022). In the end however, cancer hallmarks work by mainly affecting the molecular pathway of tumor suppressor and oncogenes (Ostroverkhova et al., 2023).

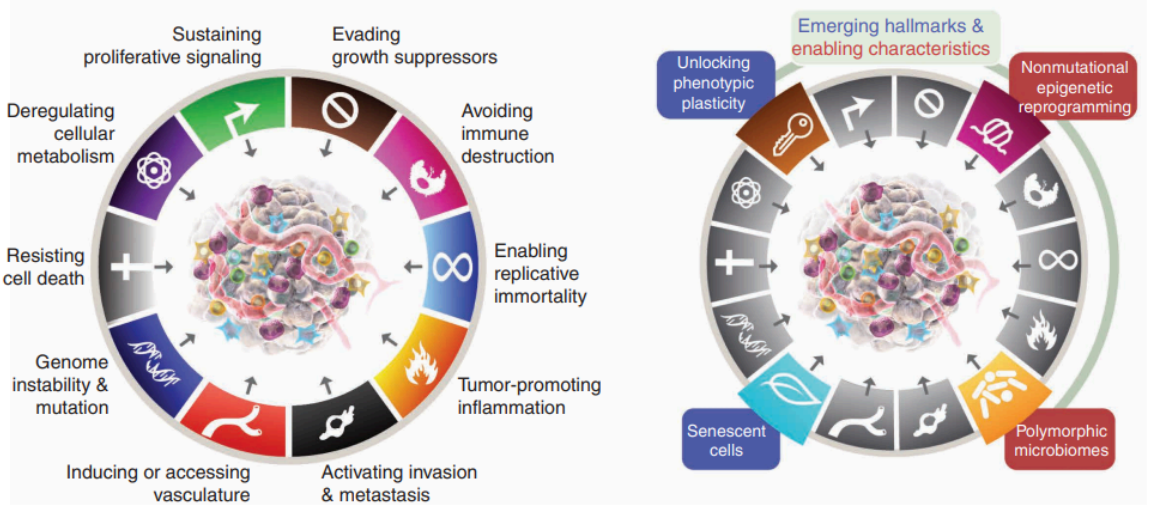



**Figure 1**. Illustration on the current hallmarks of cancer taken from "Hallmarks of Cancer: New Dimension" by Hanahan (2022).

## 2.2. Breast Cancer

As mentioned before, breast cancer is one of the many different types of cancer that has been categorized by the TCGA project. Like in most other types of cancer, histopathological analysis of breast cancer has always been considered as the best way to diagnose and categorize the type of breast cancer found in the patient (Zeiser et al., 2021). The histopathological type of the cancerous tumor is a critical criteria in determining diagnostic and prognostic evaluation of patients which is why WHO's blue book constantly updates the classification according to current development. As of the writing of this manuscript in 2024, five iterations/editions of the blue book have been released where the updates on histopathology can be seen fully in the review article by Cserni (2020). Although there are many histopathological types, most are exceedingly rare with 75% of breast cancer diagnoses being Invasive Ductal Carcinoma (IDC) and ~10% being Invasive Lobular Carcinoma (ILC) (Yoon et al., 2023; Liu

et al., 2024). Due to being the most common type of breast cancer, the TCGA code for is called breast invasive carcinoma rather than any of the other histological name classification.

Aside from diagnosis methods through direct investigation of the histological pattern of the tumor, breast cancer also has tests specific to it such as imaging through mammography and molecular subtyping. Mammography is the practice of detecting the presence of tumor cells in patients who have shown no outwards/detectable symptoms using low-energy x-rays to differentiate fatty and fibroglandular breast tissue according to their absorbance rate (din et al., 2022; Ritse, 2023). Molecular subtyping on the other hand, is a series of tests using biopsy/sample from the patient to verify the presence of hormone receptors estrogen (ER), progesterone (PR), and human epidermal growth factor 2 (HER2) (Orrantia-Borunda et al., 2022). The resulting classification of both analyses is different as mammography is used to determine the malignancy of the tumor while molecular subtyping is used to determine what type of chemotherapy drug is effective. The end result for both is also different as early detection from mammography can be used as evidence for tumor resection. Meanwhile, molecular subtyping requires further support from other tests before a suitable therapy plan can be made.

However, not all breast cancer subtypes are easily handled using targeted therapy as multiple factors play a part in allowing medical specialists to design the most suited therapy plan for the patient. As can be seen in **Figure 2**, breast cancer can be classified using the presence of the hormone receptor. Different subtypes of breast cancer require their own specific treatment according to their molecular subtype with some having better prognosis than others (Charan et al., 2020). The subtype with the poorest prognosis is the triple negative breast cancer (TNBC) or sometimes referred to as the basal subtype. The TNBC subtype, is the most aggressive subtype where none of the previously mentioned hormone receptors are present which means that detection can only be done using imaging and Immunohistochemistry (Dass et al., 2021).

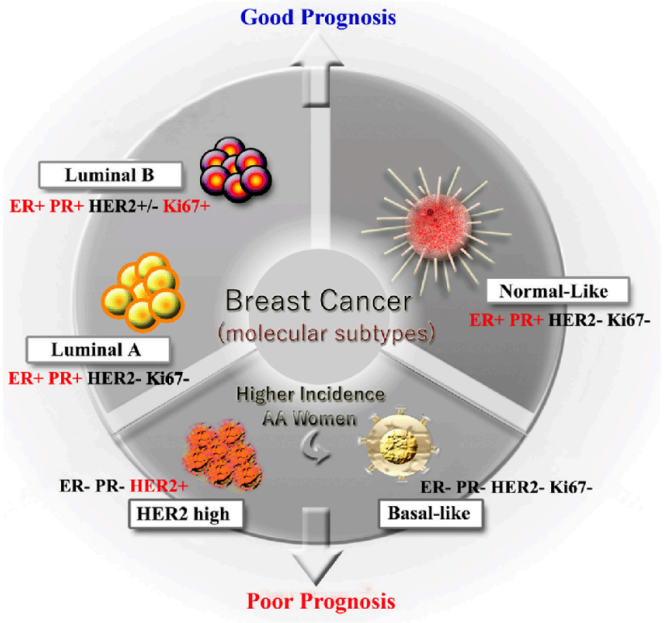



**Figure 2.** Classification of breast cancer subtypes. Take from Charan et al., (2020) Titled "*Molecular and Cellular Factors Associated with Racial Disparity in Breast Cancer*".

With the presence of hard-to-treat variants, research in breast cancer needs to be intensified as some types require more personalized treatment compared to others. This is where the updated hallmarks of cancer play a part as they have confirmed that genetics do play an important role in tumorigenesis. As such, efforts have been made to study the genomic landscape, leading to results

such as the discovery and importance of BRCA 1 and 2, crucial tumor suppressor genes with direct causative effect towards the development of breast cancer (Mehrgou & Akouchekian, 2016). Aside from those, genes such as TP53 CDH1, PTEN/STK11, and CHEK2 have been extensively researched for their potential as biomarkers or treatment targets (Walsh et al., 2017). More newly discovered genes such as CACNG4, PKMYT1, EPYC, and CHRNA6 have been proposed as potential prognosis biomarkers or as therapeutic targets by studies such as one done by Golestan et al. (2024). However, there is still a need for more accurate and specific markers that can assist in each individual's case.

.

### 2.3. Chromosomal Instability: Copy Number alterations

As it eventually comes back to the human genome, exploration of the genomic ecosystem and how the molecular function is affected becomes a particularly interesting and fascinating focus of study. Looking back at **Figure 1**, one of the newly added is “Genomic Instability & Mutation” which indicates that instability of gene expression and mutations in the genomic sequence would lead to progression of cancer. These genomic errors induce changes in the activity of certain genes and their downstream products thus causing irregular levels of pathway activity. This change of expression can be attributed to alteration events such as sequence mutation or copy number aberrations, errors which number at around 3000 base pair mutations alongside a hundred or so copy number changes in a cancer genome when compared to a normal sample (Macconaill & Garraway, 2010).

While mutations and copy numbers might seem similar, a stark difference lies in their heredity. Mutations can be passed down and may happen to be common in a population while copy numbers are unique to every individual with the numerous variations having unknown effects. A difference can also be found in that, as mentioned before, it is a somatic type of mutation where it will occur only after conception while normal mutation could be somatic or germline (Oota, 2020). What is known is that there are errors found in the human genome with each and every event, be it deletion or duplication, contributing to evolutionary traits, disease and/or microbiome interaction (Pös et al., 2021).

These unique changes result in different expression levels in genes on an individual level, but copy numbers in larger quantities results in the development of aneuploidy. Aneuploidy is a commonly found condition in cancer in which cells contain an abnormal amount of chromosomes. However, in the context of cancer, aneuploidy can also be used to describe the loss of the longer arm of a chromosome. The state of aneuploidy grants similar effects to that of continuous genetic mutations in cancer, namely increasing genetic variety. Variations in copy numbers lead to differing responses towards external stimuli and add another layer on the genetic diversity of tumors. Those variations include deletion, inversion or duplication of the specific allele as can be seen in **Figure 3** . This diversity promotes cancer progression while also assisting in cancer immune evasion (Ben-David & Amon, 2019, Lakhani et al., 2023).

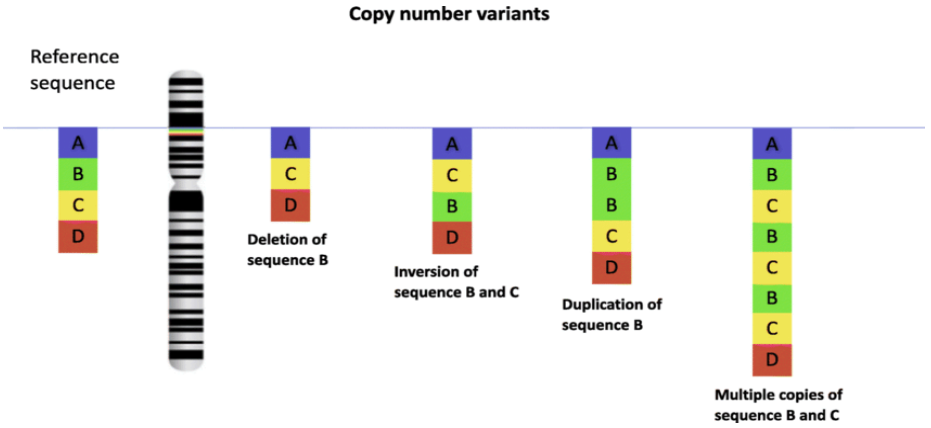



**Figure 3**. Variations of Copy Number Events as Described by Chirwani & Campbell, (2020) in "*Genetics for paediatric radiologists"*.

### 2.3.1. Copy Number in Breast Cancer

As mentioned before, aneuploidy plays an important part in cancer progression. As far back as 2010, a paper by Navin et al., (2010) already knew that aneuploidy in breast cancer is almost certain with half of breast cancer patients showing signs of aneuploidy. The common nature of aneuploidy in breast cancer is supported by newer studies such as one done by Pfister et al., (2018) and Lakhani et al. (2024). Aside from Aneuploidy, another state caused by chromosomal instability is the variation of copy numbers. Copy number variations or CNV are losses or gains of genomic fragments with a length of 50 bp up to several mbs, this is adjusted from the previous classification of only being around 1kb (Pfister et al., 2018). CNVs are common alterations and can be found in most parts of the human genome with various effects from no effect at all to being the direct cause of cancer development and progression (Murakami et al., 2020). A benign example of CNV is the variation in AMY1 gene which is more highly expressed in populations that have a higher amount of starch in their diet.On the other hand, an example of a more lethal CNV is the deletion of the BRCA1/2 Gene which directly contributes to the development of breast cancer. Other CNVs common in breast cancer are amplifications of TERT or CDK4, and deletions on chromosomes 17, 19, and 20 (Sablin et al., 2024; Mirzaei & Petreaca, 2022; Hakkaart et al., 2022).

### 2.3.2. Loss of Heterozygosity

The aberrations and increase in genomic variance caused by factors such as aneuploidy and CNV would lead to the occurrence of loss of heterozygosity events. Loss of heterozygosity or LoH is an event where a heterozygous pair of a gene or chromosome becomes homozygous. This is generally caused by failed separation during mitosis, error in homologous recombination or just deletion in a segment of a chromosome (Chambliss & Marzinke, 2020). LoH is often associated with the suppression of tumor suppressor genes and the two hit models for cancer development. The two hit model is one where a heterozygous tumor suppressor code with one being deactivated and the other activated becomes a homozygous pair and deactivates the tumor suppressor gene (Pös et al., 2021). In the context of breast cancer, a germline mutation in 1 copy of the BRCA1/2 genes would be vulnerable to somatic LOH in the corresponding pair, leading to inactivation of the gene & initiation of tumor growth (Kim & Suyama, 2022). LoH also affects the behaviour and susceptibility of the tumor. Certain genes confer resistance or modulate treatment sensitivity meaning LoH in these genes would affect how the tumor reacts to any administered medications. LoH would also affect the behaviour of a tumor such as the correlation between LoH and aggressiveness in PitNET and immune evasion through suppression of HLA expression (Yang et al.,2022; Lin et al., 2024; Santos et al., 2022).

## 2.4. Overview of Machine Learning

Artificial intelligence does not have a specified term, but the most commonly accepted one is the utilization of computers to imitate human intelligence with a branch of the studied called machine learning where the aim is to utilize algorithms for analyzing large amounts of data (França et al., 2021; Sheikh et al., 2023). The increase in technology allows the rapid advancement in machine learning. Whereas before the current iteration of machine learning was implemented, they were just basic algorithms that were made to solve simple basic problems based on predetermined results. This then evolved to what we now know with it being used in basic everyday tasks such as image recognition to something very complex such as predicting molecular structures of proteins. The aim of AI is now to be able to do the work of humans more efficiently and in the hopes that it will perform better than us.

Over the years, a lot of new algorithms are developed or old algorithms are being improved upon thus allowing there to be a lot of algorithms to choose from. Each algorithm fits a certain type of work. Because of the many applications of different machine learning and AI models, the way they are implemented is also infinite in nature as every possible problem could be solved using AI. In everyday life, the use of AI has now grown exponentially with many uses in real life. However, the main four categories: supervised; unsupervised; semi-supervised; and reinforcement learning, have their application where they are more suited (Sarker, 2021).

### 2.4.1. Supervised Learning

In how they are processed, machine learning algorithms are divided into four major categories which are mentioned above. According to **Figure 4** by (Sarker, 2021), the major 3 are: supervised; unsupervised; and reinforcement learning. Seeing as Semi-supervised is not really used, it will not be discussed. But in general it acts as an intermediary between supervised and unsupervised.

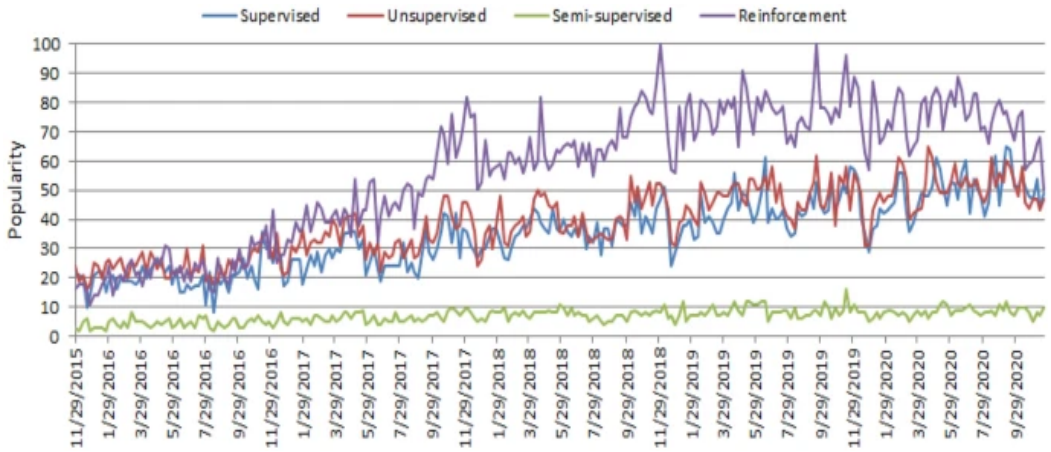



**Figure 4**. Representation popular machine learning category from 2015–2020 by Sarker (2021).

As for supervised and unsupervised, the difference between them is that unsupervised learning does not need a guide and allows the algorithm to make its own conclusion (Alloghani et al., 2020). For supervised learning, the model is guided by using labels or identifiers to get the result the human wants. In terms of complexity, unsupervised learning is very complex; however, it is oftentimes not very accurate and may give unexpected results which may cause complications in highly sensitive matter (Naeem et al., 2023). Meanwhile, reinforced learning is a type of machine learning where the program trains in an unknown,

ever changing, environment and is allowed to learn through trial and error (Naeem et al., 2020).

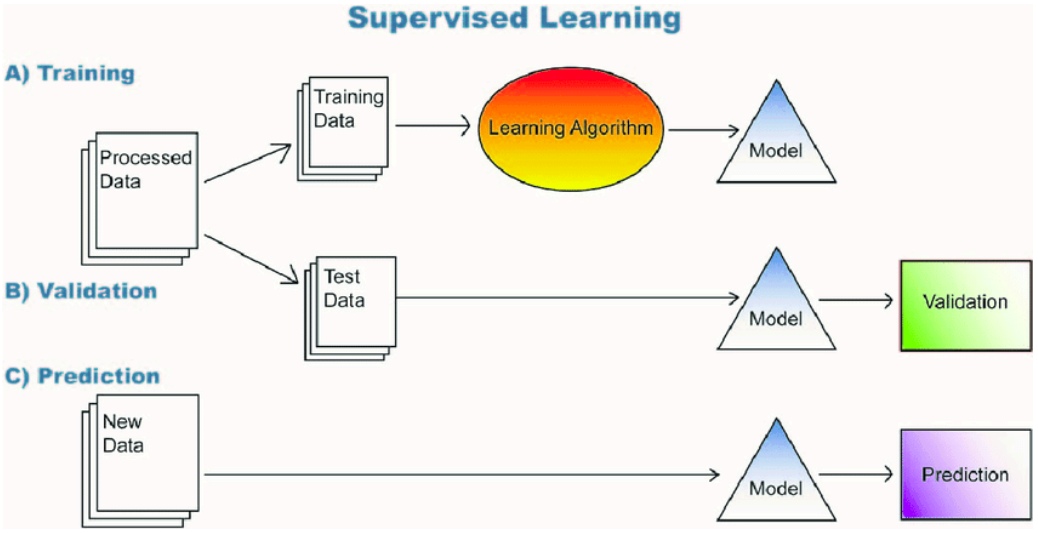



**Figure 5**. Representation of how supervised learning is done by Kanevsky et al., (2016)

In general, for highly complex aim and data, unsupervised learning is generally recommended while the aim for supervised learning is to get something as accurate as possible. Building a complex model using unsupervised learning takes time and computational power as a large number of data is needed (Bantan et al. 2020). Meanwhile, the ever changing environment for reinforced learning is very hard to apply for some data, especially with some boundaries that are not able to be translated to code (Ding & Dong, 2020). This is especially true if the sample data is very small in the beginning. As such, it is oftentimes a good idea to first propose a supervised model as a baseline for further studies. This would also allow more comprehensive data filtering to filter out bad quality data first so that no time would be wasted training and testing the model. The data could then be processed or trained according to the need, but supervised learning itself is divided into 2 major types. The first is classification, which is using class labels or features as an input and then finding the distinct value which allows categorization of class (Kornyo et al., 2023). The second is called regression where the algorithm creates a model to fit several variables together using a linear line (or plane) to then produce an equation to a value of question (Hox & Maas, 2005). The use case of regression is oftentimes to predict a value with high accuracy and certainty. However, the difficulty comes in tuning and then finding which algorithm will work best with, oftentimes, strict input necessities. This is different compared to categorical algorithms which can accommodate all data types and is more versatile, but with the trade off of not being able to do precise things.

### 2.4.2. Machine Learning in Breast Cancer Data

Machine learning and AI in itself for medical related work recently received public attention with the emergence of COVID 19. This new boom brings news to the development of medicine, especially in the terms of imaging. In particular for breast cancer, one of the ways for testing is for mammography. Recently, there have been efforts by Kyono et al., (2020); Prodan et al., (2023); and Hanis et al., (2022) to try using machine learning imaging for easier analysis of mammography data. However, mammography is mainly used for diagnosis and not prognosis (Reeves & Kaufman,

2024). As such, it is important for other methods to be explored, particularly in terms of predictive and prognosis testing which can be done through analysis of genes and then utilizing personalized medicine.

In the case of personalized medication, what is needed is high accuracy compared to complexity as the data is highly sensitive. As mentioned before, large amounts of data needed for unsupervised training would take a lot of time. This wasted time can be used to train multiple supervised machine learning models. This is why a lot of publications on breast cancer use supervised machine learning such as Mustapha et al., (2020) which trains Wisconsin data using a variety of supervised machine learning method; while there is also the use of multi omics data of breast cancer to predict therapy response by Sammut et al., (2023). This shows the efforts done to further breast cancer prognosis and prediction for future cases.

### 2.4.3. Data Stratification

Data stratification is something that has to be dealt with in the case of machine learning for biological data as generally, the data is imbalanced. There are a lot of ways to handle this, but the main method is either under sampling by removing majority data or over sampling by adding minority data (Gnip et al., 2021). By doing this, the data would not have a huge difference in each category's sample. It is also the most recommended way to transform the data so that overfitting i.e testing data that is too similar to the oneness being used to create the machine learning model.

# III. MATERIALS & METHODS

## 3.1. Pipeline Overview

The in-silico analysis uses the system provided by the host institution (Taipei medical University, Professor Chih-Yang Wang's lab) with the specifications of i9-11900 and 128 GB of DDR4 ram. The OS used was the Linux distro Ubuntu 22.04 LTS, which processed the Affymetrix CEL format files using Affymetrix Power Tools 1.19 (APT) and PennCNV through the terminal (Wang et al., 2007; Diskin et al., 2008; Wang et al., 2008; Thermo Fisher Scientific., 2016; Qiao et al., 2023). This was followed with R coding using the latest version of 4.4.1 for running the ASCAT 3.1.3 and CINdex 1.32 package alongside all dependencies that came with it (Van Loo et al., 2010; Song et al., 2017). However, for machine learning prediction, a windows system was used to facilitate the Python environment through the use of Anaconda Navigator, specifically the Jupyter Notebook (Anaconda Software Distribution, 2024).

A step-by-step overview can be seen in **Figure 6**. In general, data is collected from Gene Expression Omnibus (GEO) followed by translation of raw microarray read in LogR Ratio (LRR) and Beta-Allele Frequency (BAF) using the recommended steps in (https://penncnv.openbioinformatics.org/en/latest/user-guide/affy/). This is followed by ASCAT segmentation which is followed up by visualization in CINdex and prediction using machine learning.

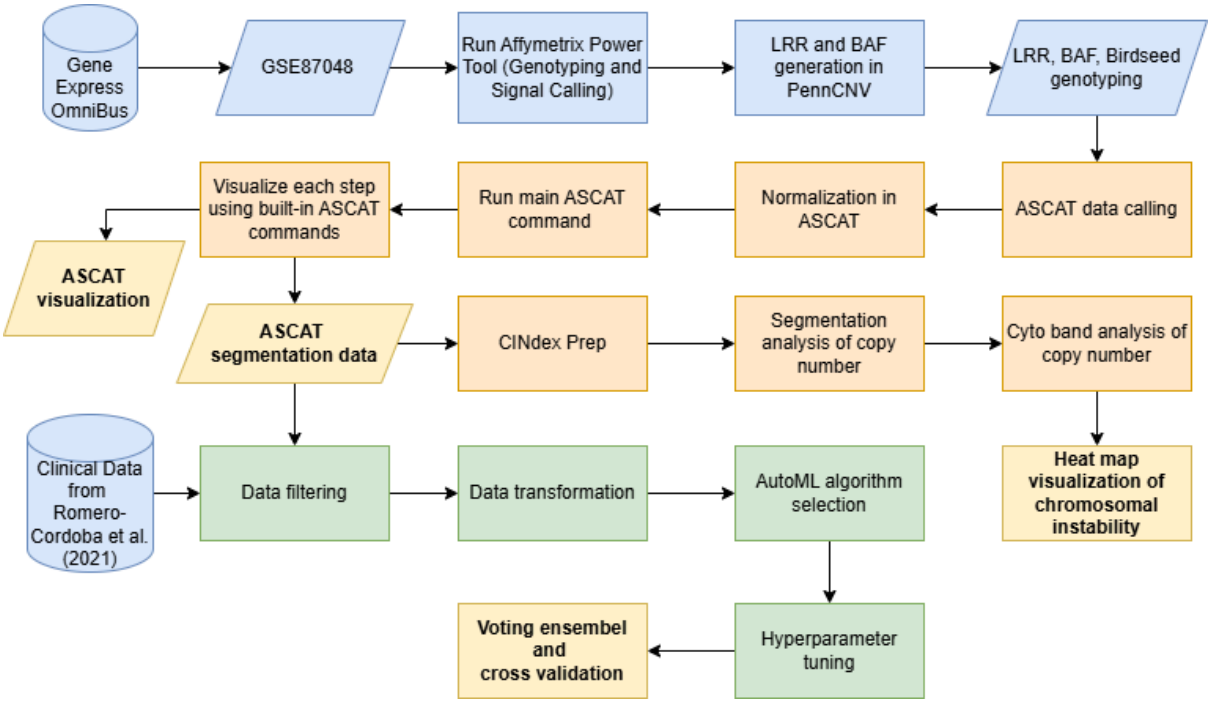



**Figure 6**. Overview of the project's pipeline

## 3.2. Data Collection

Data was collected from the GEO accession GSE87048 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87048) which were submitted in 17th of September 2016 by Romero-Cordoba et al. (2021). Data from GSE87048 contains 100 varying breast cancer data from mexican-Hispanic population with both tumor and peripheral blood samples being arrayed in Affymetrix Genome-Wide Human SNP 6.0 Array. Clinical data, meanwhile, can be found in the paper attached to the GSE entry.

In short, the sample (Tumor and peripheral blood) was extracted using QIAamp DNA Blood Maxi Kit (Qiagen, Valencia, CA). Each sample was digested using NspI and StyI enzymes (New England Biolabs), followed by T4 DNA ligation (New England Biolabs) and amplification. The

sample was then purified using magnetic beads (Agencourt) before labeling with biotin with the final step being hybridization to the array.

### 3.3. CEL Data Cleaning and Preprocessing

Data cleaning and preprocessing was performed using APT and PennCNV using the conditions mentioned above. Raw data from GEO was gathered as mentioned in section 3.2 in linux. Compressed CEL files were first decompressed followed by the execution of the penn-affy protocol provided by PennCNV. CEL was first preprocessed using an algorithm provided by APT to generate genotyping files and quantile normalization that will later be used to create the LRR and BAF file (Pitea et al., 2020). This was done using the APT tool's apt-probeset-genotype and apt-probeset-summarize which gives a birdseed and summary result respectively. The birdseed file will be used later while it will be the input of PennCNV to calculate the LRR and BAF values.

### 3.4. ASCAT

Cleaned data, as described above, will be used as input for ASCAT. Importing was done using the basic R command to import CSV files containing the needed data to the environment. The data was then processed according to the guidelines directed by the author with some modifications. Data type conversion was done to GenomicRanges followed by normalization by the in-house ASCAT algorithm followed by segmentation of the data according to the adjacent normal tumor data. After that, the main ASCAT algorithm was run to produce the copy number variant data as well as the ploidy of the sample. For a full view of the code, please visit the GitHub repository (https://github.com/Darkam1101/Copy_number).

Classification of CNV was done based on the traditional assumptions that the sample has a diploid genome (Gardina et al., 2008). They are classified into 4 general categories which are: normal, loss, gain, and LoH. For normal, nMajor and nMinor must total to 1 and 1 under the assumption that the total copy number is equal. Loss is categorized as 1 and 0 or 0 and 0 where all n value is gone . Gain is where nMajor and nMinor is more than 1 while LoH is when nMinor = 0 and nMajor > 0.

### 3.5. CINdex

CINdex was run in the same environment as ASCAT. The CINdex package was first imported from Bioconducter before ASCAT segmentation output was converted into GenomicRanges data type. Preparation was also done by importing reference data for the Human genome using the assembly HG19 which includes cytoband location, gene annotation, reference genome, and clinical data using the sub-types. The package was then run according to the author's vignette, please refer to the previously given GitHub repository for code used.

### 3.6. Machine Learning

Raw Segmentation data from ASCAT was concatenated for each of the samples and then imported to python. In total, only 73 samples out of 100 were used due to missing clinical data. Each data was then processed using one hot encoding while missing data followed by transformation of Boolean data to integers. Testing was then done using a multitude of algorithms through the use of the AutoML package LazyPredict (https://github.com/shankarpandala/lazypredict). The results of the first round of testing were then trained using the top 5 algorithm while partly modifying LazyPredicts preprocessing for the same results. Each algorithm was then visualized using a confusion matrix for better clarity of their performance. After that, data imbalance was fixed by applying the SMOTE algorithm and two other variants i.e SMOTE+ENN and SMOTE+Tomek. The same process was repeated using the best performing balancer algorithm and the top 5 models based on the F1 score were trained manually with hyper-parameter tuning before Ensemble voting was applied. The scoring

was then visualized using a confusion matrix to get a better view on how accurate each model is (balanced accuracy >0.8). For further information regarding the code, please see the GitHub repository given before.

# IV. RESULTS AND DISCUSSION

## 4.1. Exploration of the Data Set

The data set used from Romero-Cordoba et al. (2021). In total, There is 100 data that is Inputted into ASCAT. The distribution of the cancer types before data cleaning can be seen in **Figure 7A** while after it has been cleaned by removing data without clinical data can be seen in **Figure 7B**. In total, around 2500 or so data points were deleted with data from the segmentation data. Both **Figures 7A and B** are consistent with past research where both luminal subtypes are common while basal is the least common as they have the highest and lowest prognosis, respectively. This indicates that the distribution and resulting data would have a slight bias towards luminal A samples, particularly in the machine learning parts.

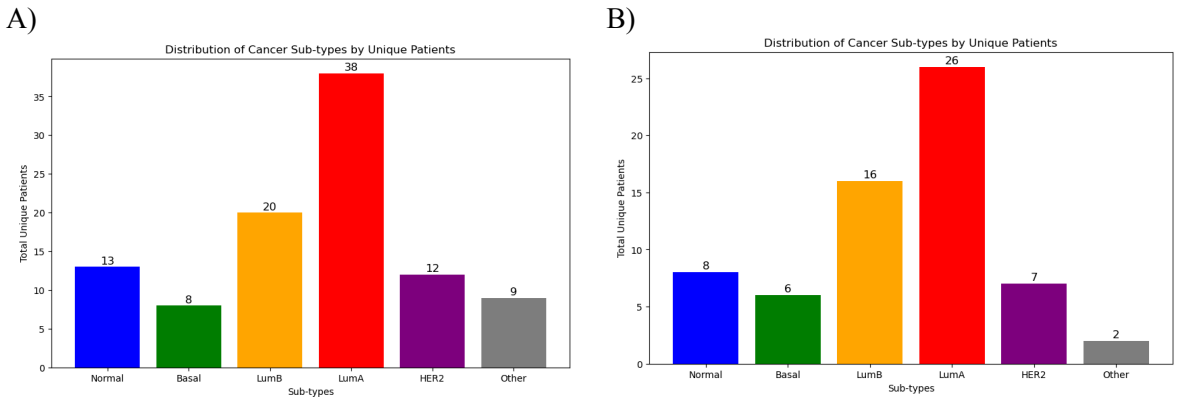



**Figure 7.** A) Distribution of subtypes before data cleaning. B) Distribution of data after it has been cleaned

## 4.2. Copy Number Events from ASCAT

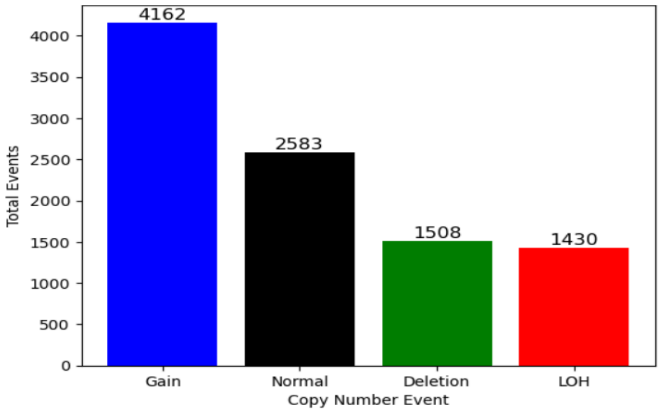



**Figure 8**. Distribution of CNV events after being processed through ASCAT and classified according to GISTIC classification.

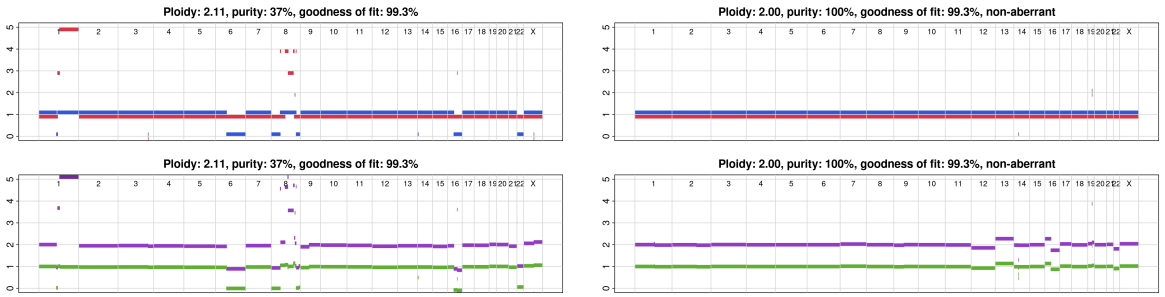
The ASCAT algorithm is an algorithm that can calculate the ploidy and CNV for each sample that is inputted into the algorithm. The data is derived from the LogR and BAF data generated according to the method explained above. LogR and BAF data that is inputted can actually be used for other algorithms that can also predict CNV such as OncoSNP, GenoCNA, and GISTIC (Pitea et al., 2018). Those algorithms have their own specific way of classifying CNV events. But in general, ASCAT is the one used the most due to its ability to detect LoH events, aberrant tumor cell fraction, copy neutral events, and ploidy with it being the most accepted way to analyze copy number data (Favero et al., 2015; Shahrouzi et al., 2024).

During data calling and processing, ASCAT data can be visualized in each step of the process as the algorithm is not a continuous process, but multiple codes that need to be run. The results from ASCAT come in two forms, visualized plots and segmentation data in the form of a .txt file. The segmentation was plotted manually in python as seen in **Figure 8** where the classification was done as described in the method section 3. 4.

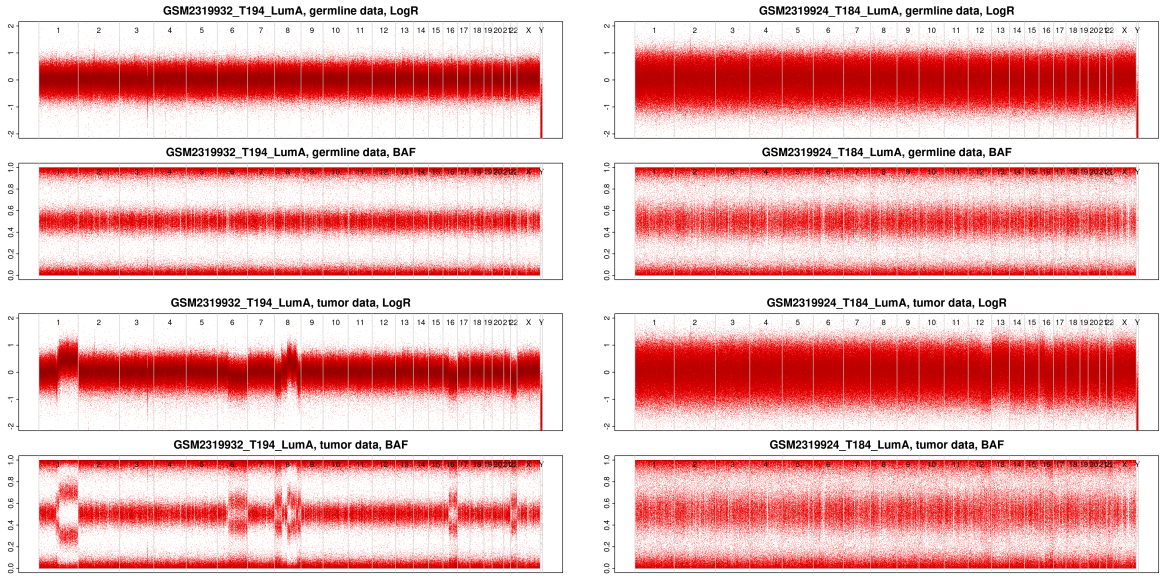
Meanwhile, an example of the visualized results of ASCAT can be seen in **Figure 9** where luminal A subtype tumor and non-aberrant tumor sample can be seen in **Figure 9**. In the figures, it is possible to compare one instance of an aberrant sample with a non-aberrant tumor of the same subtype. In **Figure 9A,** the figure showed the segmentation data before and after it was rounded which snows where specifically the CNV happens. However, using the raw signaling from LRR and BAF value in **Figure 9B** is possible albeit complicated. This is however incomplete without the sunrise plot in **Figure 9C** as in this plot, it shows the probability of where the ploidy is with the darker the color representing higher accuracy.

From comparison in **Figure 9A**, raw segmentation data was rounded up by the algorithm to make a better graph that when compared shows the condition of the chromosomes. In this case, we can see some deletion/loss events and amplification/gain events. Gain events are seen when the total reading of the chart is not equal to the ploidy value while loss events are the reverse. In this case, the event mainly covers chromosome 1, 6, 8, minor parts in 14, 18, and 22. This correlates with the possibility that some tumor suppressor gene is likely to be found in chromosomes 6, 8, 18, and 22.

A)




B)



C)

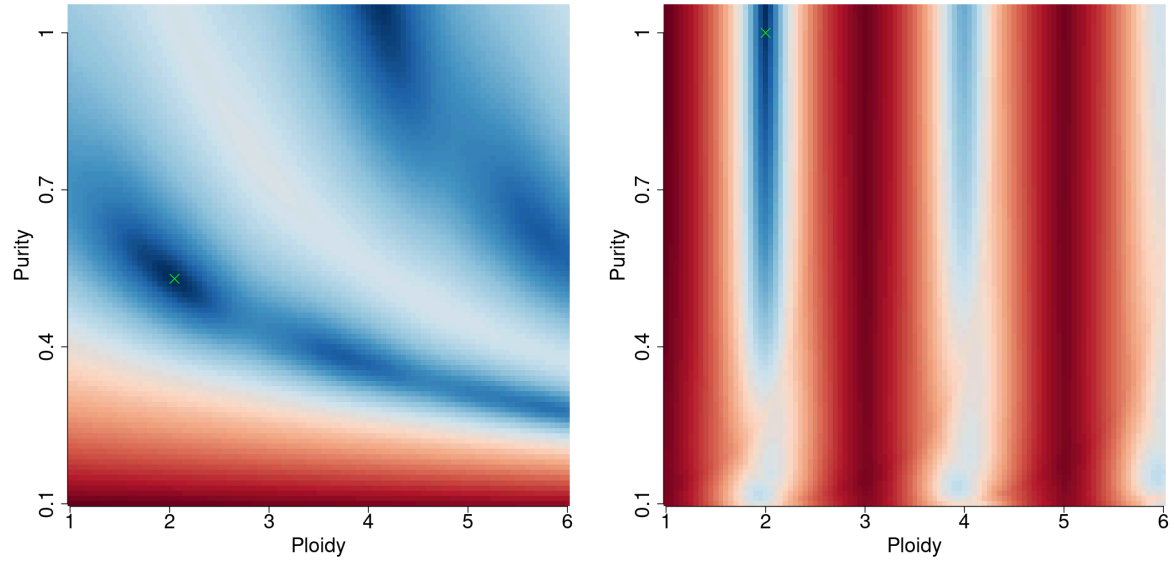



**Figure 9**. Visualization of ASCAT from luminal A sample/sample T194 (left) and non-aberrant luminal A sample/sample T184 (right): A) Segmentation data from Raw (up) and rounded (down) ; B) Data of tumor (up) and normal/germline (down); and C) Sunrise plot of the probability of the ploidy

An example is the whole deletion in chromosome 22 where the tumor suppressor CHK2 is found. As it is a tumor suppressor, loss of one of the alleles as in LoH of CHK2 would affect many downstream activation pathways. According to **Figure 10** by Boonen et al., (2022) would affect several other tumor suppressor genes in breast cancer, thus allowing tumorigenesis by causing several things. One of the pathways is BRCA1 where it is connected to DNA repair. According to Li et al., (2020), the pathway is the one responsible for maintaining the stability of the genome during repair. If they are affected, possible CNV may occur again, thus lowering prognosis.

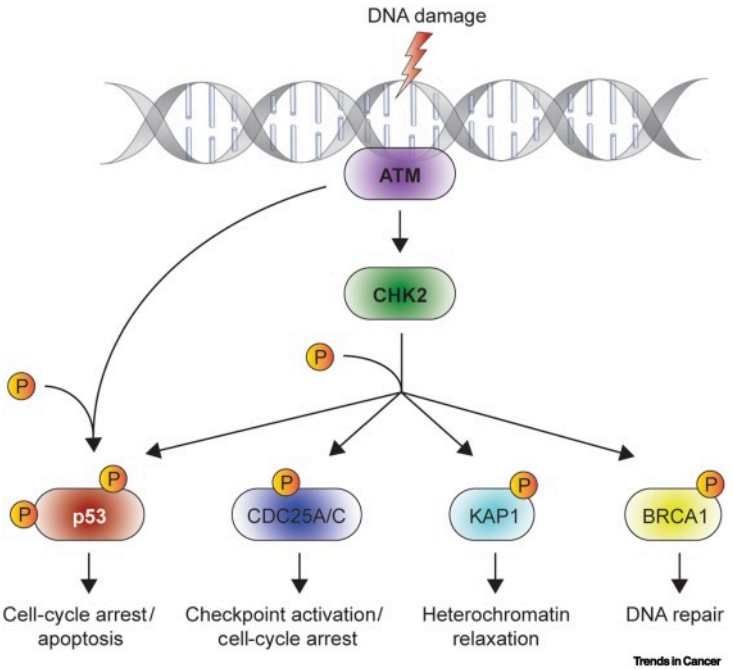



**Figure 10**. Downstream pathway of CHK2 by Boonen et al., (2022) in "CHEK2 variants: linking functional impact to cancer risk"

Aside from luminal A, there are other types of subtypes being analyzed as mentioned before. The result of the segmentation can be seen in **Figure 11**. From what can be seen HER2 and basal have the most subtypes. However, more CNV can be seen in the HER2 subtypes while ploidy in basal is a lot more which could affect the expression of the whole genome. But it should be noted that, copy number affects everyone differently as they are catered towards individuals. Only general conclusions can be made as each sample's analysis is only relevant to that particular sample due to the nature of CNV itself.

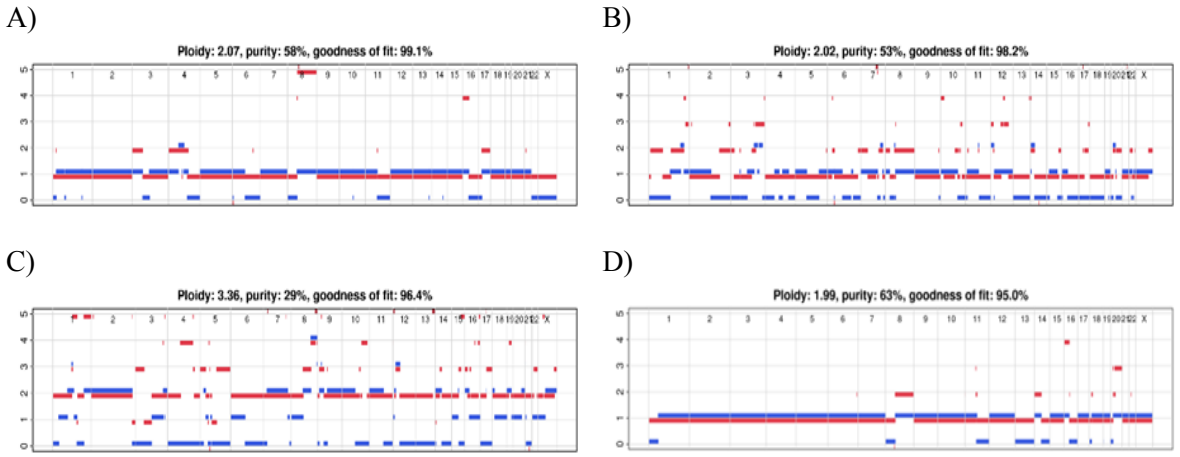



**Figure 11**. Visualization of CNV data from sample A) T186, subtype luminal B; B) T193, subtype HER2; C) T174, subtype basal; D)T29, subtype normal-like.

As seen above, the use of copy number is very relevant in predicting and knowing how the genome might be affected at that particular time. This offers prognostic value and if early enough done, might be useful for prediction of breast cancer pathogenicity.

### 4.3. Chromosomal Instability Visualization from CINdex

After ASCAT was done, the data produced was trained for machine learning and CINdex. CINdex is a tool found in anaconda to visualize the chromosomal instability from any copy number segmentation data. The instability of the chromosome can be directly tied to and is the cause of several diseases. From CINdex, it is possible to see which particular chromosome and their particular region is highly unstable. The software produces visualization for copy number segmentation using their own proprietary classification of copy number events. However, the results for normalized value is null and for each different cut off is the same. This in turn affected the visualization for the cytoband of each chromosome where all normalized plots are empty or only some being hard to interpret in the cytoband level. The cause of the missing value in the figure is most likely caused by the translation data done by APT where half of the probe's signals were deleted. This caused only some parts of the genome to be analyzed by ASCAT.

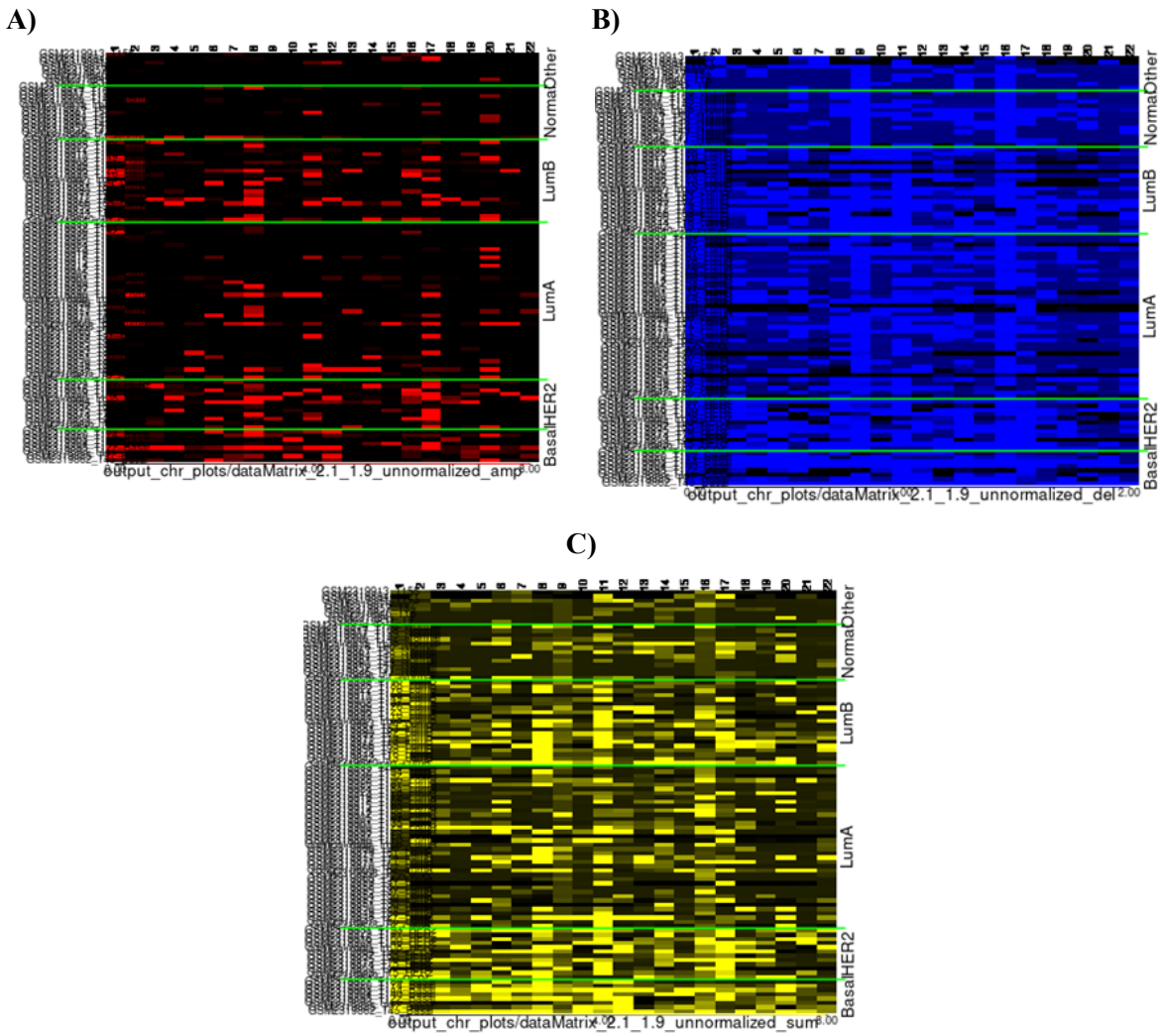



**Figure 12.** Visualization of CNV using CINdex across the genome with threshold (gain = 2.1 and loss = 1.9). A) unnormalized amplification events; B) unnormalized deletion event; C) unnormalized sum event.

As mentioned before, the results for normalized values are blank. However, it is possible to interpret them using the unnormalized value as seen in **Figure 12** . From what can be seen, most copy number events are present in chromosome 11 and 8 with chromosome 17 being a close third. This gain result of chromosome 8 and 11, especially in ER+ subtypes (luminal A and B) is consistent with previous studies as written in a literature review by Shahrouzi et al., (2024). Unfortunately, noise in deletion events leads to bias in interpretation which is also backed up by the results of ASCAT as seen in **Figure 8**. One of the possible reasons for this is because while the data from ASCAt was categorized using GISTIC's categorization, CINdex has their own way of categorizing the events.

### 4.4. Machine learning Training

The segmentation data previously obtained from ASCAT then was imported into the python environment. The segmentation data used was the raw segmentation as the values of the copy number were not rounded. For ASCAT and CINdex, as mentioned before, there was no data that was excluded from the analysis. Meanwhile for machine learning, the rows that were cleaned first were the ones that didn't have any clinical data according to the clinical data file as given by Romero-Cordoba et al. (2021) and also had more than 5 features of missing data (**Supplementary Figure 5)**. The sample ID removed total 35, with the exact ID being:

GSM2319837, GSM2319839, GSM2319841, GSM2319846, GSM2319847,
GSM2319852, GSM2319855, GSM2319862, GSM2319863, GSM2319865,
GSM2319868, GSM2319873, GSM2319876, GSM2319883, GSM2319890,
GSM2319891, GSM2319892, GSM2319896, GSM2319913, GSM2319915,
GSM2319916, GSM2319918, GSM2319919, GSM2319920, GSM2319921,
GSM2319931, GSM2319934, GSM2319836, GSM2319843, GSM2319845,
GSM2319851, GSM2319860, GSM2319899, GSM2319905, GSM2319935

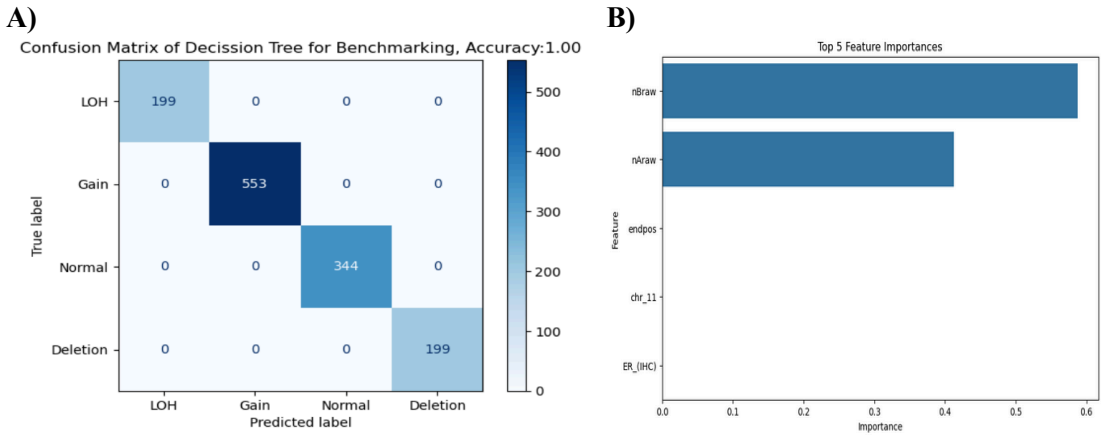



**Figure 13**. A) Result of decision tree for benchmarking; B) Feature importance analysis of the decision tree

After data cleaning and imputation, training was performed using the basic decision tree to get a benchmark on the data. The result is 100% accuracy as depicted in **Figure 13A**. This was due to the presence of the 'nAraw' and 'nBraw'. This was confirmed using the feature importance test as seen in **Figure 13B** which would indicate that the data was not learning at all. This is unfortunately not what is needed which means that the features need to be removed.

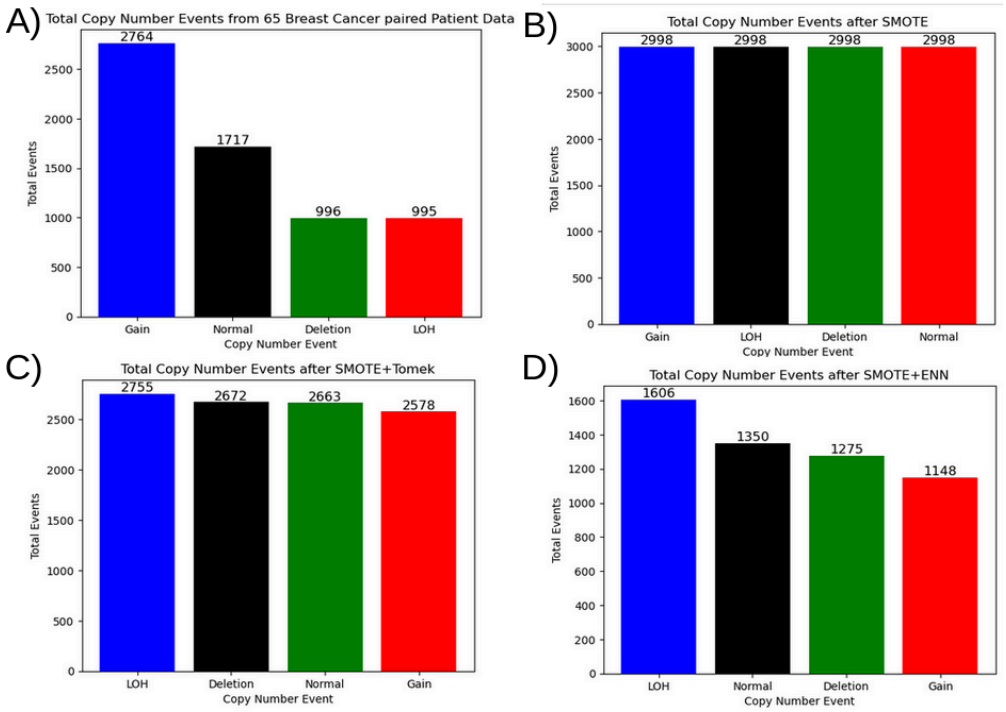



**Figure 14**. Distribution of data points used for machine learning in: A) pre-stratification of data (7057); B) SMOTE data stratification (11992); C) SMOTE+TOMEK data stratification (10668); D) SMOTE+ENN data stratification (5379).

As can be seen in **Figure 14** the data was imbalanced as was common for genomic data, it is imperative that the data be stratified for machine learning.Although it is possible to generate the data using machine learning, it is not recommended to do so as creation of genomic data without strict controls would actually introduce more bias and interfere with analysis. Despite this, data generation will have to be done in order to balance the data and to replace overlapping data using. While it is possible to use under-sampling (i.e. removing data) if missing data is found, pre-established methods since Batista et al., (2004), support that data-generation as the more appropriate alternative. To achieve this, data generation using SMOTE and its derivative was used to handle this (Pradipta et al., 2021). All variations of the most popular SMOTE were used, which were Tomek link/Tomek and ENN. Tomek links and ENN are both additions to SMOTE where after data was added, under sampling was done. In Tomek, it is done by removing overlapping samples while ENN removes samples that are thought to be noise (Sasada et al., 2020).

**Table 1**. Machine learning training using LazyPredict after removing 'nAraw' and 'nBraw' features.

| Model | F1 Score | | | |
|---|---|---|---|---|
| | No Transformation | SMOTE | SMOTE +TOMEK | SMOTE+ENN |
| ExtraTreesClassifier | 0.68 | 0.50 | 0.75 | 0.95 |
| RandomForestClassifier | 0.69 | 0.54 | 0.76 | 0.94 |
| LGBMClassifier | 0.71 | 0.62 | 0.77 | 0.94 |
| XGBClassifier | 0.72 | 0.61 | 0.78 | 0.93 |
| ExtraTreeClassifier | 0.62 | 0.51 | 0.69 | 0.91 |
| BaggingClassifier | 0.71 | 0.57 | 0.76 | 0.90 |
| DecisionTreeClassifier | 0.66 | 0.55 | 0.73 | 0.90 |
| LabelSpreading | 0.65 | 0.48 | 0.68 | 0.88 |
| LabelPropagation | 0.65 | 0.47 | 0.68 | 0.88 |
| SVC | 0.68 | 0.56 | 0.71 | 0.86 |
| KNeighborsClassifier | 0.65 | 0.56 | 0.69 | 0.84 |
| NuSVC | 0.67 | 0.53 | 0.68 | 0.83 |
| LogisticRegression | 0.59 | 0.55 | 0.63 | 0.76 |
| LinearSVC | 0.58 | 0.54 | 0.61 | 0.75 |
| CalibratedClassifierCV | 0.58 | 0.54 | 0.62 | 0.75 |
| LinearDiscriminantAnalysis | 0.58 | 0.56 | 0.61 | 0.74 |

| | | | | |
|---|---|---|---|---|
| RidgeClassifier | 0.58 | 0.54 | 0.61 | 0.73 |
| RidgeClassifierCV | 0.58 | 0.54 | 0.62 | 0.73 |
| SGDClassifier | 0.55 | 0.52 | 0.58 | 0.72 |
| Perceptron | 0.50 | 0.47 | 0.51 | 0.69 |
| PassiveAggressiveClassifier | 0.48 | 0.46 | 0.51 | 0.67 |
| AdaBoostClassifier | 0.56 | 0.53 | 0.58 | 0.63 |
| QuadraticDiscriminantAnalysis | 0.33 | 0.42 | 0.33 | 0.49 |
| NearestCentroid | 0.46 | 0.43 | 0.45 | 0.57 |
| BernoulliNB | 0.45 | 0.45 | 0.44 | 0.55 |
| GaussianNB | 0.41 | 0.35 | 0.34 | 0.53 |
| DummyClassifier | 0.10 | 0.26 | 0.10 | 0.31 |

As SMOTE+ENN showed the best results, the top 5 models based on F1 score in **Figure 15** were then passed through hyperparameter tuning for better results. The F1 score symbolized the mean value between precision and recall ability of the algorithm (Hicks et al., 2022). The use of it in this case was because with better precision and recall, the features predicted are relevant features and the model could minimize false predictions. The results for each model after hyper parameter tuning can be seen in **Figure 16.** The best model performer according to the figure is random forest with 0.97 after training with 5 times cross validation. This is followed by extra tree, light GBM, XGB, and finally decision tree. The presence of the decision tree at the bottom is not surprising, but what is surprising is that it made the top 5.

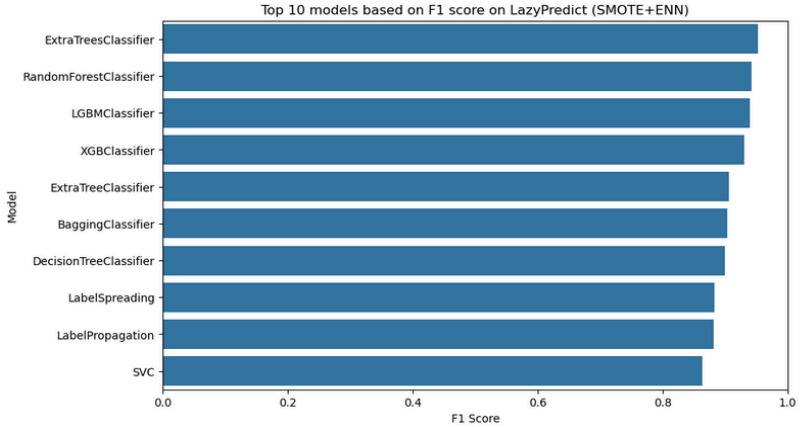



**Figure 15**. F1 score ranking from LazyPredict

Comparing the results from LazyPredict (**Supplementary Tables 4–7**) with the results from **Figure 14** a significant improvement can be seen in some cases. This is consistent with what Talaei Khoei & Kaabouch, (2023) said with hyper parameter tuning.

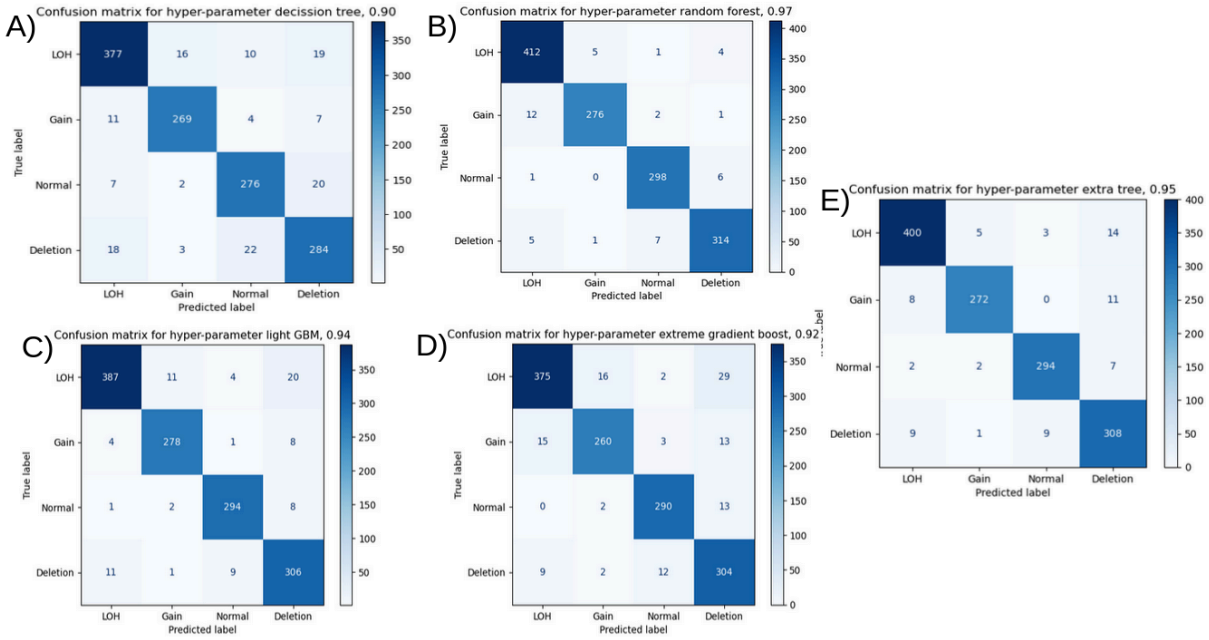



**Figure 16.** Confusion matrices for the algorithm: A) decision tree; B) random forest; C) light GBM; D) XGB; and E) Extra tree

In terms of complexity, the decision tree is one of the more basic algorithms that can be used, However, despite its simplicity, it is robust enough that it can handle a variety of tasks (Mienye & Jere, 2024). Unfortunately, because of this trait, it is not expected that the decision tree will be near the top of the list (**Table 1**). The others however, are not that surprising as the Ensemble algorithm is known to be one of the most powerful collections of algorithms. Ensembl itself uses already defined algorithms and builds upon it, thereby perfecting the model as can be seen in the figure below (Kumar et al., 2022). Whereas light GBM, XGB and random forest are built upon decision trees, the voting algorithm further enhances all of them by taking input from every single algorithm thus making it having the best result (**Figure 17**). This is the reason why Ensemble was used instead of deep learning, as deep learning has a steep learning curve and the time trying to implement it may not be worth the effort (Mohammed & Kora, 2023).

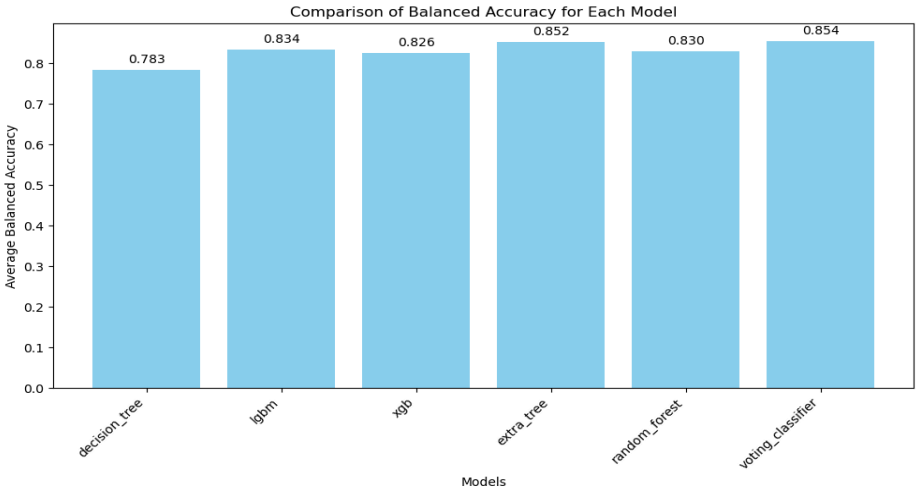



**Figure 17**. Distribution of balanced accuracy during cross validation after voting training: A) decision tree; B) random forest; C) light GBM; D) XGB; and E) Extra tree

### 4.5. Limitations

As with many other studies, there were limitations that hinder the project's results. First was regarding the ASCAT segmentation. It should be noted that the segmentation data for ASCAT starts not from the 0 bp in every chromosome which, unfortunately, the data showed is not the whole genome, even though it should be. The reason is that the copy number probe which makes up half of the SNP 6.0 Human DNA array was deleted by APT. This is consistent with reports by Dennis et al., (2020), where CNV probe calls must be made from overlapping probes to ensure accuracy. Because of that, it could be assumed that APT removed those due to covering specific regions that may introduce bias downstream. There is also the fact that, as with past research and standard practices, that classification of copy number is done with the assumption that the genome is diploid. In cases where ploidy is more than 2, manual re-classification must be done to avoid wrong classification as seen in Masood et al. (2024). This unfortunately was not done to inconclusive support on how to define aneuploidy copy number with several researchers debating what is the best way to classify aneuploid CNV.

Limitations can also be found with CINdex. While the use and existence of such a package is extremely helpful in viewing and comparing data, the long processing time will present a problem when implementing it in real-life. Aside from that, the 3 different levels of threshold which are gain = 2.1 and loss = 1.9; gain = 2.25 and loss = 1.75; and gain = 2.5 and loss = 1.5 produces raw and normalized values that are then plotted. However, the plot for raw values are the same while for normalized values, they are blank (Appendix 1–3). The error in blank results also could be caused by the large coverage by ASCAT and not small regions that may be expected by the algorithms. However, a possible replacement was developed by Oza et al., (2023) that also calculates CIN, but is more focused on specific CIN which does not allow the visualization of the whole genome like CIN.

In machine learning, limitations faced were the data imbalance of CNV variants and the classification of the CNV types. Imbalance data was solved using SMOTE+ENN as described above in the result and discussion. But, with genomic data, it is generally not recommended to generate pseudo-data and has been a controversial topic even though guidelines have been established (Lazic et al., 2020). Due to this, it is very hard to work with machine learning for copy numbers as 65 sources is not enough for good prediction. However, more samples may have the ability to ensure better results and better validity of the model. For CNV classification, there are many different ways of classifying it. One is using the GISTIC way with:

“-2 for homozygous loss (nMinor + nMajor = 0), –1 for hemizygous loss (nMinor + nMajor = 1), 0 for normal (nMinor + nMajor = 2), 1 for three copies (nMinor + nMajor = 3), and 2 for more than three copies (nMinor + nMajor > 3)” (Castro-Mondragon et al., 2022).

However, this was done with further transformation and normalization that isn’t natively supported by ASCAT. There is also unfortunately no consensus on how to process ASCAT data into GISTIC format when done individually. However, a paper by Renault et al. (2017) may solve the problem as they have an integrated pipeline for CNV analysis using both microarray and whole genome/exome sequencing. Although the package's dependencies are largely deprecated, necessitating the use of Docker, managing Docker itself represents a separate task beyond the scope of this report.

# V. SELF REFLECTION

This internship, for me personally, is something of a new experience. Blessed with the opportunity to go abroad, I went to another country known for their work ethics and leading technology in the field of semiconductor which contrast the sometimes old settings of Taipei, where I stayed. However, they are not an english speaking country with almost no english skills, as is commonly in east asian countries, which leads me to picking up some very basic mandarin skills of which I am grateful for.

As I mentioned previously, I stayed in Taipei city where the capital of Taiwan is located. The specific institution I interned In was Taipei medical University or shortened as TMU, which is very heavily focused in medicine, especially in cancer research with them having five hospitals located around Taipei. Due to this, I am very much immersed with the latest research in trying to treat cancer as the specific department I am placed in was the Graduate Institute for Cancer Research and Drug Discovery.

The specific project I am assigned with was breast cancer research with a focus on the genomics side for analysis. As can be read in detail in the literature review section, I am tasked with finding copy number data from *in silico* genomics analysis of cancer patients. As cancer is a topic rarely covered in the Bioinformatics major, I was a bit stuck in the working of the project. However, several courses have prepared me for this such as Genomics, Epigenetics, coding classes such as R and python, and also several projects that are held as part of the class. They mostly helped in the critical thinking and problem solving department as no course actually taught what to do and what to use for copy number. But for the machine learning part, the courses in Python and Ai in Life Sciences certainly helped a lot for the machine learning department.

I also wanted to highlight the usefulness of the course BM3115 Functional Genomics and Proteomics for my project as they certainly gave a lot of context and background, not for this project but the side project assigned to me. Aside from that main project, the PI in TMU often asked for help in analyzing wet lab data, specifically analysis of sequencing data which is preceded by wet lab experiments done in house. Through this course and the Genomics course I was able to have the mindset for both wet lab and dry lab which allows for better interpretation of the data. However the courses did not have anything that can be implemented as Genomics mainly handles DNA while I worked on RNA, and Functional Genomics and Proteomics only gives the theory.

As the courses have not given any concrete methodology that can be replicated, I mostly have to search and do everything by myself and take a lot of time just reading and troubleshooting. This is why compared to i3L students in the department, my project took longer as I have no basis to work on from both i3L and TMU. Because of this, I gained new skills and learned a lot in Genomics analysis for RNA and what analysis I should do while for the copy number itself, I gained both knowledge in how and what to correlate copy number with the possibility of breast cancer prognosis.

Aside from that, the programs and events held by i3L also helped in giving a glimpse of how work would be done and how to hold myself with others. Because of this I was able to work in a timely manner and also able to help some of my lab mates and finish some of the work assigned to me. I am truly grateful for the opportunity to meet; help; and in some cases, teach some on how to do *In silico* analysis for their data.

# VI. CONCLUSION

The results shown in this report suggest and also proves that the use of machine learning is possible in trying to determine copy number events from a data set of 100 patient data. But, the best accuracy reported, that is 97 %, should be taken with a grain of salt as with every machine model that is shown here or that has been utilized in real life. This is because, to implement this in the clinical setting, there is a need to further improve the model whilst also adding more data with cross validation score achieving high numbers. The additions of the data will bring more credibility for the algorithm whilst simultaneously refining the model to achieve better results. This would certainly require a lot of time and effort as the samples collected must be normalized and then be processed through a variety of segmentation algorithms.

Several suggestions on what could be improved is by using a different target such as predicting the value of nAraw and nBraw as those directly contribute to the state of the copy number. Classification of the CNV can also be expanded whilst also taking account different ploidy status of the cancer genome as the classes specified in this report are a very generalized one assuming a diploid genome. Another interesting thing would be to directly predict the signal value given by the microarray, however that is still very far in the future. As such, using the current pipeline built would be a good step in the right direction in analyzing more copy number data in the hopes that they could be used to train new models.

# REFERENCES

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1

Anaconda Software Distribution. (2024). Anaconda. In *Anaconda Inc.* (2.6.2). Anaconda Inc. https://anaconda.com/

Bantan, R. A. R., Ali, A., Naeem, S., Jamal, F., Elgarhy, M., & Chesneau, C. (2020). Discrimination of sunflower seeds using multispectral and texture dataset in combination with region selection and supervised classification methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(11). https://doi.org/10.1063/5.0024017

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Ben-David, U., & Amon, A. (2019). Context is everything: aneuploidy in cancer. *Nature Reviews Genetics, 21(*1), 44–62. https://doi.org/10.1038/s41576-019-0171-x

Bennett, C., Carroll, C., Wright, C., Awad, B., Park, J. M., Farmer, M., Brown, E. (Bryce), Heatherly, A., & Woodard, S. (2022). Breast cancer genomics: Primary and most common metastases. *Cancers*, *14*(13), 3046. https://doi.org/10.3390/cancers14133046

Boonen, R. A. C. M., Vreeswijk, M. P. G., & van Attikum, H. (2022). CHEK2 variants: linking functional impact to cancer risk. *Trends in Cancer*, *8*(9), 759–770. https://doi.org/10.1016/j.trecan.2022.04.009

Brown, J. S., Amend, S. R., Austin, R. H., Gatenby, R. A., Hammarlund, E. U., & Pienta, K. J. (2023). Updating the definition of cancer. *Molecular Cancer Research*, *21*(11). https://doi.org/10.1158/1541-7786.mcr-23-0411

Chambliss, A. B., & Marzinke, M. A. (2020). Applications of molecular techniques in the clinical laboratory. In *Contemporary Practice in Clinical Chemistry* (pp. 337–349). Elsevier. https://doi.org/10.1016/B978-0-12-815499-1.00020-X

Charan, M., Verma, A. K., Hussain, S., Misri, S., Mishra, S., Majumder, S., Ramaswamy, B., Ahirwar, D., & Ganju, R. K. (2020). Molecular and Cellular Factors Associated with Racial Disparity in Breast Cancer. *International Journal of Molecular Sciences*, *21*(16), 5936. https://doi.org/10.3390/ijms21165936

Chawla, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cserni, G. (2020). Histological type and typing of breast carcinomas and the WHO classification changes over time. In *Pathologica* (Vol. 112, Issue 1, pp. 25–41). Pacini Editore S.p.A. https://doi.org/10.32074/1591-951X-1-20

Dennis, J., Walker, L., Tyrer, J., Michailidou, K., & Easton, D. F. (2021). Detecting rare copy number variants from Illumina genotyping arrays with the CamCNV pipeline: Segmentation of *z*

‐scores improves detection and reliability. *Genetic Epidemiology*, *45*(3), 237–248. https://doi.org/10.1002/gepi.22367

Deryusheva, I. v, Tsyganov, M., Garbukov, E. Y., Ibragimova, M. K., Kzhyshkovska, J. G., Slonimskaya, E., Cherdyntseva, N. v, & Litviakov, N. v. (2017). Genome-wide association study of loss of heterozygosity and metastasis-free survival in breast cancer patients. *Experimental Oncology*, *39*(2), 145–150. https://doi.org/10.31768/2312-8852.2017.39(2):145-150

din, N. M. ud, Dar, R. A., Rasool, M., & Assad, A. (2022). Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. In *Computers in Biology and Medicine* (Vol. 149). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2022.106073

Ding, Z., & Dong, H. (2020). Challenges of Reinforcement Learning. In *Deep Reinforcement Learning* (pp. 249–272). Springer Singapore. https://doi.org/10.1007/978-981-15-4095-0_7

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., & Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, *36*(19), e126–e126. https://doi.org/10.1093/nar/gkn556

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207

Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., Szallasi, Z., & Eklund, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, *26*(1), 64–70. https://doi.org/10.1093/annonc/mdu479

França, R. P., Borges Monteiro, A. C., Arthur, R., & Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. In V. Piuri, S. Raj, A. Genovese, & R. Srivastava (Eds.), *Trends in Deep Learning Methodologies* (pp. 63–87). Elsevier. https://doi.org/10.1016/B978-0-12-822226-3.00003-9

Gardina, P. J., Lo, K. C., Lee, W., Cowell, J. K., & Turpaz, Y. (2008). Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP mapping arrays. BMC Genomics, 9(1), 489. https://doi.org/10.1186/1471-2164-9-489

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., Jemal, A., & Siegel, R. L. (2022). Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(6). https://doi.org/10.3322/caac.21754

Gnip, P., Vokorokos, L., & Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. *PeerJ Computer Science*, *7*, e604. https://doi.org/10.7717/peerj-cs.604

Golestan, A., Tahmasebi, A., Maghsoodi, N., Faraji, S. N., Irajie, C., & Ramezani, A. (2024). Unveiling promising breast cancer biomarkers: an integrative approach combining bioinformatics analysis and experimental verification. *BMC Cancer*, *24*(1), 155. https://doi.org/10.1186/s12885-024-11913-7

Grimm, D. (2023). Recent advances in breast cancer research. *International Journal of Molecular Sciences*, *24*(15), 11990. https://doi.org/10.3390/ijms241511990

Hakkaart, C., Pearson, J. F., Marquart, L., Dennis, J., Wiggins, G. a. R., Barnes, D. R., Robinson, B. A., Mace, P. D., Aittomäki, K., Andrulis, I. L., Arun, B. K., Azzollini, J., Balmaña, J., Barkardottir, R. B., Belhadj, S., Berger, L., Blok, M. J., Boonen, S. E., Borde, J., . . . Walker, L. C. (2022). Copy number variants as modifiers of breast cancer risk for BRCA1/BRCA2 pathogenic variant carriers. *Communications Biology, 5*(1). https://doi.org/10.1038/s42003-022-03978-6

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. In *Cancer Discovery* (Vol. 12, Issue 1, pp. 31–46). American Association for Cancer Research Inc. https://doi.org/10.1158/2159-8290.CD-21-1059

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. *Cell*, *100*, 57–70. https://doi.org/https://doi.org/10.1016/S0092-8674(00)81683-9

Hanis, T. M., Islam, M. A., & Musa, K. I. (2022). Diagnostic Accuracy of Machine Learning Models on Mammography in Breast Cancer Classification: A Meta-Analysis. *Diagnostics (Basel, Switzerland)*, *12*(7). https://doi.org/10.3390/diagnostics12071643

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Hox, J. J., & Maas, C. J. M. (2005). Multilevel Analysis. In *Encyclopedia of Social Measurement* (pp. 785–793). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00560-0

Kanevsky, J., Corban, J., Gaster, R., Kanevsky, A., Lin, S., & Gilardino, M. (2016). Big Data and Machine Learning in Plastic Surgery: A New Frontier in Surgical Innovation. *Plastic & Reconstructive Surgery*, *137*(5), 890e–897e. https://doi.org/10.1097/PRS.0000000000002088

Kim, H., & Suyama, M. (2022). Genome-wide identification of copy neutral loss of heterozygosity reveals its possible association with spatial positioning of chromosomes. *Human Molecular Genetics, 32*(7), 1175–1183. https://doi.org/10.1093/hmg/ddac278

Kornyo, O., Asante, M., Opoku, R., Owusu-Agyemang, K., Tei Partey, B., Baah, E. K., & Boadu, N. (2023). Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms. *Computers & Security*, *135*, 103456. https://doi.org/10.1016/j.cose.2023.103456

Kumar, V., Singh Aydav, P. S., & Minz, S. (2022). Multi-view ensemble learning using multi-objective particle swarm optimization for high dimensional data classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 8523–8537. https://doi.org/10.1016/j.jksuci.2021.08.029

Kyono, T., Gilbert, F. J., & van der Schaar, M. (2020). Improving Workflow Efficiency for Mammography Using Machine Learning. *Journal of the American College of Radiology*, *17*(1), 56–63. https://doi.org/10.1016/j.jacr.2019.05.012

Lakhani, A. A., Thompson, S. L., & Sheltzer, J. M. (2023). Aneuploidy in human cancer: new tools and perspectives. *Trends in genetics : TIG, 39(*12), 968–980. https://doi.org/10.1016/j.tig.2023.09.002

Lazic, S. E., Mellor, J. R., Ashby, M. C., & Munafo, M. R. (2020). A Bayesian predictive approach for dealing with pseudoreplication. *Scientific Reports*, *10*(1), 2366. https://doi.org/10.1038/s41598-020-59384-7

Lebok, P., Kopperschmidt, V., Kluth, M., Hube-Magg, C., Özden, C., Taskin, B., Hussein, K., Mittenzwei, A., Lebeau, A., Witzel, I., Wölber, L., Mahner, S., Jänicke, F., Geist, S., Paluchowski, P., Wilke, C. R., Heilenkötter, U., Simon, R., Sauter, G., … Burandt, E. (2015). Partial PTEN deletion is linked to poor prognosis in breast cancer. *BMC Cancer*, *15*(963). https://doi.org/10.1186/s12885-015-1770-3

Li, L., Guan, Y., Chen, X., Yang, J., & Cheng, Y. (2021). DNA Repair Pathways in Cancer Therapy and Resistance. *Frontiers in Pharmacology*, *11*. https://doi.org/10.3389/fphar.2020.629266

Lin, A. L., Rudneva, V. A., Richards, A. L., Zhang, Y., Woo, H. J., Cohen, M., Tisnado, J., Majd, N., Wardlaw, S. L., Page-Wilson, G., Sengupta, S., Chow, F., Goichot, B., Ozer, B. H., Dietrich, J., Nachtigall, L., Desai, A., Alano, T., Ogilive, S., … Tabar, V. (2024). Genome-wide loss of heterozygosity predicts aggressive, treatment-refractory behavior in pituitary neuroendocrine tumors. *Acta Neuropathologica*, *147*(1). https://doi.org/10.1007/s00401-024-02736-8

Liu, G., Kong, X., Dai, Q., Cheng, H., Wang, J., Gao, J., & Wang, Y. (2023). Clinical Features and Prognoses of Patients With Breast Cancer Who Underwent Surgery. *JAMA Network Open*, *6*(8), E2331078. https://doi.org/10.1001/jamanetworkopen.2023.31078

Macconaill, L. E., & Garraway, L. A. (2010). Clinical implications of the cancer genome. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *28*(35), 5219–5228. https://doi.org/10.1200/JCO.2009.27.4944

Mallory, X. F., Edrisi, M., Navin, N., & Nakhleh, L. (2020). Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02119-8

Mann, R. (2023). Conventional Breast Imaging. In M. Iima, S. C. Partridge, & D. le Bihan (Eds.), *Diffusion MRI of the Breast* (pp. 18–39). Elsevier. https://doi.org/10.1016/B978-0-323-79702-3.00002-2

Masood, D., Ren, L., Nguyen, C., Brundu, F. G., Zheng, L., Zhao, Y., Jaeger, E., Li, Y., Cha, S. W., Halpern, A., Truong, S., Virata, M., Yan, C., Chen, Q., Pang, A., Alberto, R., Xiao, C., Yang, Z., Chen, W., & Wang, C. (2024). Evaluation of somatic copy number variation detection by NGS technologies and bioinformatics tools on a hyper-diploid cancer genome. Genome Biology, 25(1). https://doi.org/10.1186/s13059-024-03294-8

Masood, S. (2016). Breast cancer subtypes: Morphologic and biologic characterization. *Women's Health*, *12*(1), 103–119. https://doi.org/10.2217/whe.15.99

Mehrgou, A., & Akouchekian, M. (2016). The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. *Medical Journal of the Islamic Republic of Iran*, *30*, 369.

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees: Concepts, Algorithms, and Applications. *IEEE Access*, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838

Mirzaei, G., & Petreaca, R. C. (2022). Distribution of copy number variations and rearrangement endpoints in human cancers with a review of literature. *Mutation research, 824*, 111773. https://doi.org/10.1016/j.mrfmmm.2021.111773

Murakami, F., Tsuboi, Y., Takahashi, Y., Horimoto, Y., Mogushi, K., Ito, T., Emi, M., Matsubara, D., Shibata, T., Saito, M., & Murakami, Y. (2020). Short somatic alterations at the site of copy number variation in breast cancer. *Cancer Science, 112(*1), 444–453. https://doi.org/10.1111/cas.14630

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014

Mohammed, A. A. (2021). The clinical behavior of different molecular subtypes of breast cancer. *Cancer Treatment and Research Communications*, 29, 100469. https://doi.org/10.1016/j.ctarc.2021.100469

Mu, Q., & Wang, J. (2021). CNAPE: A machine learning method for copy number alteration prediction from gene expression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(1), 306–311. https://doi.org/10.1109/tcbb.2019.2944827

Mustapha, M. T., Ozsahin, D. U., Ozsahin, I., & Uzun, B. (2022). Breast Cancer Screening Based on Supervised Learning and Multi-Criteria Decision-Making. *Diagnostics (Basel, Switzerland)*, *12*(6). https://doi.org/10.3390/diagnostics12061326

Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. *IEEE Access*, *8*, 209320–209344. https://doi.org/10.1109/ACCESS.2020.3038605

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

Naeim, F., Rao, P. N., Song, S. X., & Phan, R. T. (2018). Cancer cytogenetics. In *Atlas of Hematopathology: Morphology, Immunophenotype, Cytogenetics, and Molecular Approaches* (pp. 57–68). Elsevier BV. https://doi.org/10.1016/b978-0-12-809843-1.00003-6

National Cancer Institute. (2021, October 11). *What is cancer?* National Institutes of Health. https://www.cancer.gov/about-cancer/understanding/what-is-cancer

National Cancer Institute. (2022, May 13). *The cancer genome atlas program (TCGA)*. National Institute of Health. https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Månér, S., Zetterberg, A., Hicks, J., & Wigler, M. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Research*, *20*(1), 68–80. https://doi.org/10.1101/gr.099622.109

Oota, S. (2020). Somatic mutations – Evolution within the individual. *Methods*, *176*, 91–98. https://doi.org/10.1016/j.ymeth.2019.11.002

Orrantia-Borunda, E., Anchondo-Nuñez, P., Acuña-Aguilar, L. E., Gómez-Valles, F. O., & Ramírez-Valdespino, C. A. (2022). Subtypes of Breast Cancer. In H. N. Mayrovitz (Ed.), *Breast Cancer* (pp. 31–42). Exon Publications. https://doi.org/10.36255/exon-publications-breast-cancer-subtypes

Ostroverkhova, D., Przytycka, T. M., & Panchenko, A. R. (2023). Cancer driver mutations: Predictions and reality. *Trends in Molecular Medicine*, *29*(7). https://doi.org/10.1016/j.molmed.2023.03.007

Oza, V. H., Fisher, J. L., Darji, R., & Lasseigne, B. N. (2023). CINmetrics: an R package for analyzing copy number aberrations as a measure of chromosomal instability. *PeerJ*, *11*, e15244. https://doi.org/10.7717/peerj.15244

Pfister, K., Pipka, J. L., Chiang, C., Liu, Y., Clark, R. A., Keller, R., Skoglund, P., Guertin, M. J., Hall, I. M., & Stukenberg, P. T. (2018). Identification of Drivers of Aneuploidy in Breast Tumors. *Cell reports, 23(*9), 2758–2769. https://doi.org/10.1016/j.celrep.2018.04.102

Pitea, A., Kondofersky, I., Sass, S., Theis, F. J., Mueller, N. S., & Unger, K. (2018). Copy number aberrations from Affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches? *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bby096

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., & Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*, *44*(5), 548–559. https://doi.org/10.1016/j.bj.2021.02.003

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021). SMOTE for Handling Imbalanced Data Problem : A Review. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–8. https://doi.org/10.1109/ICIC54025.2021.9632912

Prodan, M., Paraschiv, E., & Stanciu, A. (2023). Applying Deep Learning Methods for Mammography Analysis and Breast Cancer Detection. *Applied Sciences*, *13*(7), 4272. https://doi.org/10.3390/app13074272

Pyke, R. M., Mellacheruvu, D., Dea, S., Abbott, C. W., McDaniel, L., Bhave, D. P., Zhang, S. v, Lévy, E., Bartha, G., West, J., Snyder, M., Chen, R., & Boyle, S. M. (2022). A machine learning algorithm with subclonal sensitivity reveals widespread pan-cancer human leukocyte antigen loss of heterozygosity. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-29203-w

Rajpal, S., Rajpal, A., Agarwal, M., Kumar, V., Abraham, A., Khanna, D., & Kumar, N. (2023). XAI-CNVMarker: Explainable AI-based copy number variant biomarker discovery for breast cancer subtypes. *Biomedical Signal Processing and Control*, *84*, 104979. https://doi.org/10.1016/j.bspc.2023.104979

Reeves, R. A., & Kaufman, T. (2023). Mammography. In PubMed. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK559310/

Renault, V., Tost, J., Pichon, F., Wang-Renault, S.-F., Letouzé, E., Imbeaud, S., Zucman-Rossi, J., Deleuze, J.-F., & How-Kit, A. (2017). aCNViewer: Comprehensive genome-wide visualization of absolute copy number and copy neutral variations. PLOS ONE, 12(12), e0189334. https://doi.org/10.1371/journal.pone.0189334

Romero-Cordoba, S. L., Salido-Guadarrama, I., Rebollar-Vega, R., Bautista-Piña, V., Dominguez-Reyes, C., Tenorio-Torres, A., Villegas-Carlos, F., Fernández-López, J. C., Uribe-Figueroa, L., Alfaro-Ruiz, L., & Hidalgo-Miranda, A. (2021a). Comprehensive omic characterization of breast cancer in Mexican-Hispanic women. *Nature Communications*, *12*(1), 2245. https://doi.org/10.1038/s41467-021-22478-5

Romero-Cordoba, S. L., Salido-Guadarrama, I., Rebollar-Vega, R., Bautista-Piña, V., Dominguez-Reyes, C., Tenorio-Torres, A., Villegas-Carlos, F., Fernández-López, J. C., Uribe-Figueroa, L., Alfaro-Ruiz, L., & Hidalgo-Miranda, A. (2021b). Comprehensive omic characterization of breast cancer in Mexican-Hispanic women. *Nature Communications*, *12*(1), 2245. https://doi.org/10.1038/s41467-021-22478-5

Sablin, M., Gestraud, P., Jonas, S. F., Lamy, C., Lacroix-Triki, M., Bachelot, T., Filleron, T., Lacroix, L., Tran-Dien, A., Jézéquel, P., Mauduit, M., Monteiro, J. B., Jimenez, M., Michiels, S., Attignon, V., Soubeyran, I., Driouch, K., Servant, N., Tourneau, C. L., . . . Bièche, I. (2024). Copy number alterations in metastatic and early breast tumours: prognostic and acquired biomarkers of resistance to CDK4/6 inhibitors. *British Journal of Cancer, 131(*6), 1060–1067. https://doi.org/10.1038/s41416-024-02804-6

Sammut, S.-J., Crispin-Ortuzar, M., Chin, S.-F., Provenzano, E., Bardwell, H. A., Ma, W., Cope, W., Dariush, A., Dawson, S.-J., Abraham, J. E., Dunn, J., Hiller, L., Thomas, J., Cameron, D. A., Bartlett, J. M. S., Hayward, L., Pharoah, P. D., Markowetz, F., Rueda, O. M., … Caldas, C. (2022). Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, *601*(7894), 623–629. https://doi.org/10.1038/s41586-021-04278-5

Santana dos Santos, E., Spurdle, A. B., Carraro, D. M., Briaux, A., Southey, M., Torrezan, G., Petitalot, A., Leman, R., Lafitte, P., Meseure, D., Driouch, K., Side, L., Brewer, C., Beck, S., Melville, A., Callaway, A., Revillion, F., Folgueira, M. A. A. K., Parsons, M. T., … Rouleau, E. (2022). Value of the loss of heterozygosity to BRCA1 variant classification. *Npj Breast Cancer*, *8*(1). https://doi.org/10.1038/s41523-021-00361-2

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A Resampling Method for Imbalanced Datasets Considering Noise and Overlap. *Procedia Computer Science*, *176*, 420–429. https://doi.org/10.1016/j.procs.2020.08.043

Schirwani, S., & Campbell, J. (2020). Genetics for paediatric radiologists. *Pediatric Radiology*, *50*(12), 1680–1690. https://doi.org/10.1007/s00247-020-04837-4

Shahrouzi, P., Forouz, F., Mathelier, A., Kristensen, V. N., & Duijf, P. H. G. (2024a). Copy number alterations: a catastrophic orchestration of the breast cancer genome. *Trends in Molecular Medicine*, *30*(8), 750–764. https://doi.org/10.1016/j.molmed.2024.04.017

Shahrouzi, P., Forouz, F., Mathelier, A., Kristensen, V. N., & Duijf, P. H. G. (2024b). Copy number alterations: a catastrophic orchestration of the breast cancer genome. *Trends in Molecular Medicine*, *30*(8), 750–764. https://doi.org/10.1016/j.molmed.2024.04.017

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Artificial Intelligence: Definition and Background* (pp. 15–41). https://doi.org/10.1007/978-3-031-21448-6_2

Song, L., Bhuvaneshwar, K., Wang, Y., Feng, Y., Shih, I.-M., Madhavan, S., & Gusev, Y. (2017). CINdex: A bioconductor package for analysis of chromosome instability in DNA copy number data. *Cancer Informatics*, *16*, 117693511774663. https://doi.org/10.1177/1176935117746637

Talaei Khoei, T., & Kaabouch, N. (2023). Machine Learning: Models, Challenges, and Research Directions. *Future Internet*, *15*(10), 332. https://doi.org/10.3390/fi15100332

Tan, E. S., Knepper, T. C., Wang, X., Permuth, J. B., Wang, L., Fleming, J. B., & Xie, H. (2022). Copy Number Alterations as Novel Biomarkers and Therapeutic Targets in Colorectal Cancer. *Cancers*, *14*(9), 2223. https://doi.org/10.3390/cancers14092223

Thermo Fisher Scientific Inc. (2016). *Affymetrix Power Tools* (1.19.0). Thermo Fisher Scientific Inc.

Tsyganov, M. M., Ibragimova, M. K., Garbukov, E. Yu., Bragina, O. D., Zdereva, E. A., Usynin, E. A., & Litviakov, N. v. (2022). Predictive and prognostic significance of loss of heterozygosity in ABC transporter genes in breast cancer. *Siberian Journal of Oncology*, *21*(5), 34–43. https://doi.org/10.21294/1814-4861-2022-21-5-34-43

van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A. L., & Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(39), 16910–16915. https://doi.org/10.1073/pnas.1009843107

Walsh, M. F., Nathanson, K. L., Couch, F. J., & Offit, K. (2016). Genomic Biomarkers for Breast Cancer Risk. *Advances in Experimental Medicine and Biology*, *882*, 1–32. https://doi.org/10.1007/978-3-319-22909-6_1

Wang, K., Chen, Z., Tadesse, M. G., Glessner, J., Grant, S. F. A., Hakonarson, H., Bucan, M., & Li, M. (2008). Modeling genetic inheritance of copy number variations. *Nucleic Acids Research*, *36*(21), e138–e138. https://doi.org/10.1093/nar/gkn641

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–1674. https://doi.org/10.1101/gr.6861907

World Health Organization. (2024, February 1). *Global cancer burden growing, amidst mounting need for services*. World Health Organization. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

Xu, J., Huang, L., & Li, J. (2016). DNA aneuploidy and breast cancer: a meta-analysis of 141,163 cases. *Oncotarget*, *7*(37), 60218–60229. https://doi.org/10.18632/oncotarget.11130

Yang, Y., Kim, E., & Kim, S. (2022). Insignificant effects of loss of heterozygosity in HLA in the efficacy of immune checkpoint blockade treatment. *Genes & genomics, 44(*4), 509–515. https://doi.org/10.1007/s13258-021-01207-8

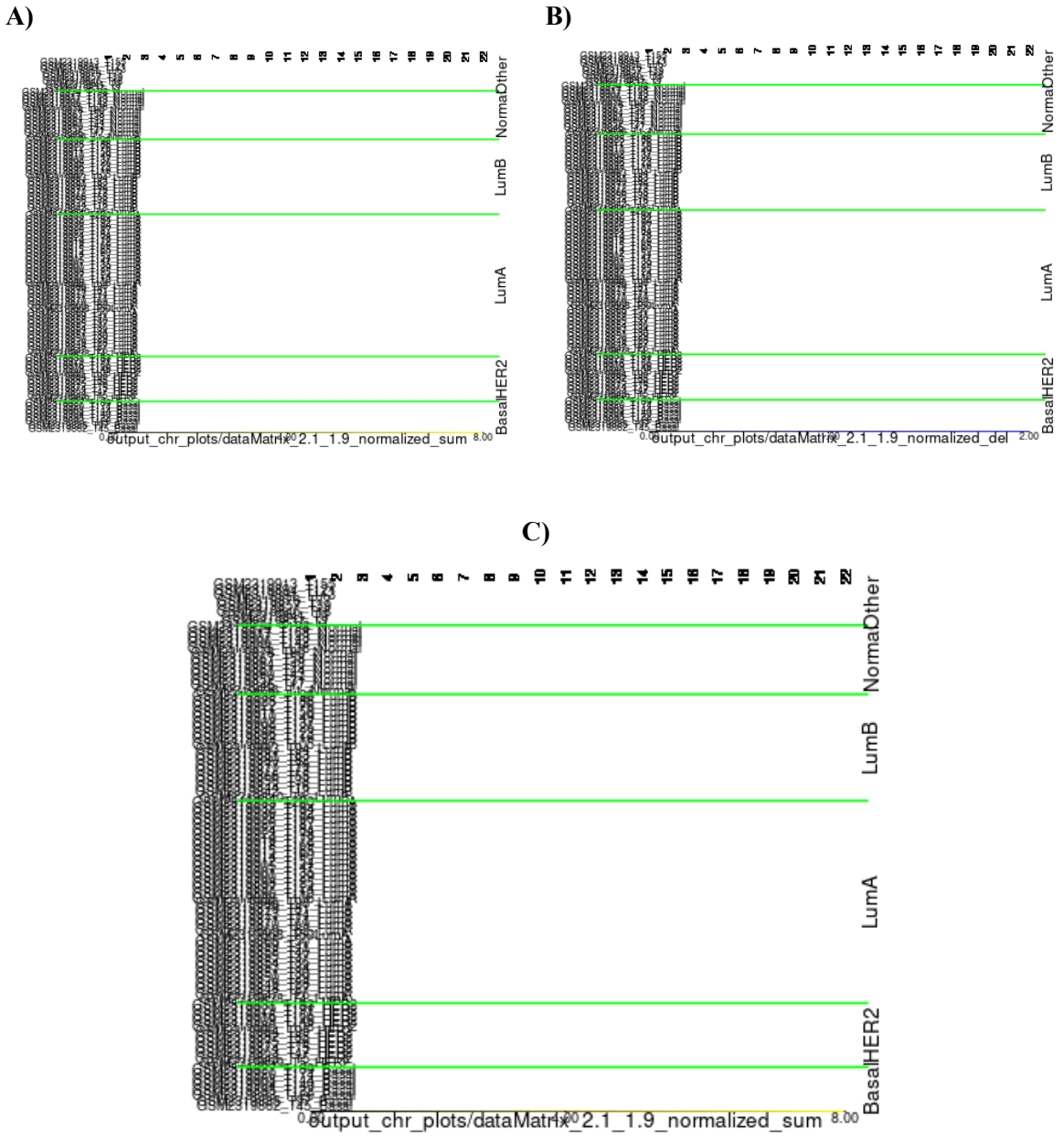
Yoon, T. I., Jeong, J., Lee, S., Ryu, J. M., Lee, Y. J., Lee, J. Y., Hwang, K. T., Kim, H., Kim, S., Lee, S. B., Ko, B. S., Lee, J. W., Son, B. H., Metzger, O., & Kim, H. J. (2023). Survival Outcomes in Premenopausal Patients With Invasive Lobular Carcinoma. *JAMA Network Open*, *6*(11), E2342270. https://doi.org/10.1001/jamanetworkopen.2023.42270

Young, C. C., Eason, K., Garcia, R. M., Moulange, R., Mukherjee, S., Chin, S.-F., Caldas, C., & Rueda, O. M. (2024). Development and validation of a reliable DNA copy-number-based machine learning algorithm (CopyClust) for breast cancer integrative cluster classification. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-62724-6

Zeira, R., & Raphael, B. J. (2020). Copy number evolution with weighted aberrations in cancer. *Bioinformatics*, *36*(Supplement_1), i344–i352. https://doi.org/10.1093/bioinformatics/btaa470

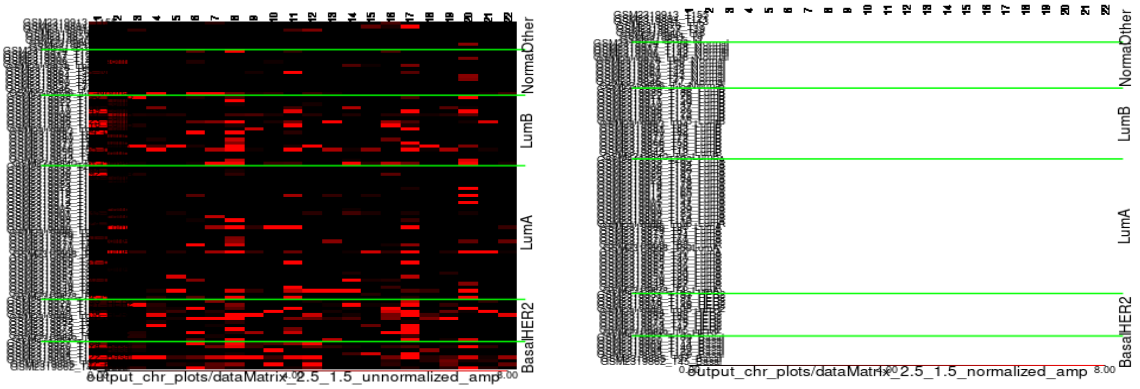
Zeiser, F. A., da Costa, C. A., Roehe, A. V., Righi, R. da R., & Marques, N. M. C. (2021). Breast cancer intelligent analysis of histopathological data: A systematic review. *Applied Soft Computing*, *113*, 107886. https://doi.org/10.1016/j.asoc.2021.107886

# APPENDICES

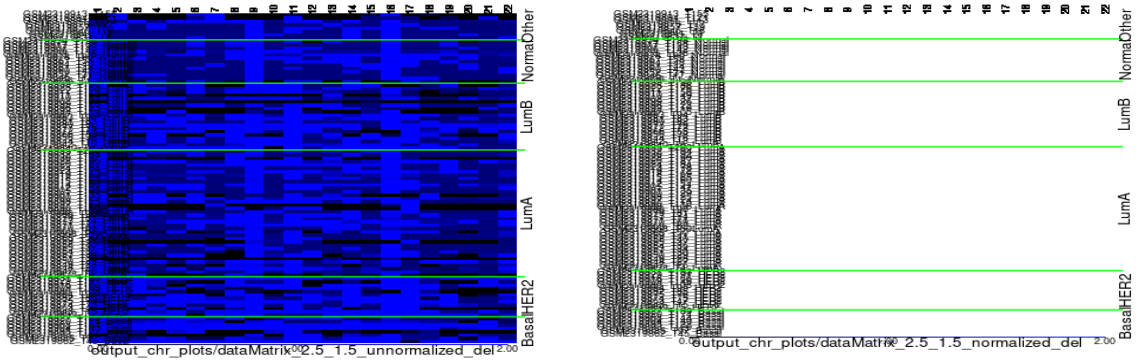



**Supplementary Figure 1**. Visualization of CNV using CINdex across the genome with threshold (gain = 2.1 and loss = 1.9). A) normalized amplification events; B) normalized deletion event; C) normalized sum event.

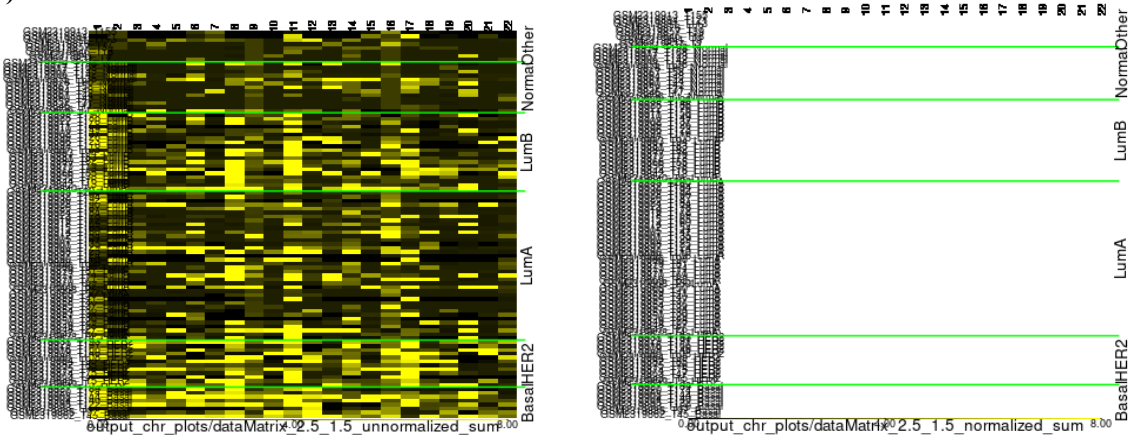
**A)**


NormalOther
LumB
LumA
BasalHER2
output_chr_plots/dataMatrix_2.5_1.5_unnormalized_amp
output_chr_plots/dataMatrix_2.5_1.5_normalized_amp

**B)**

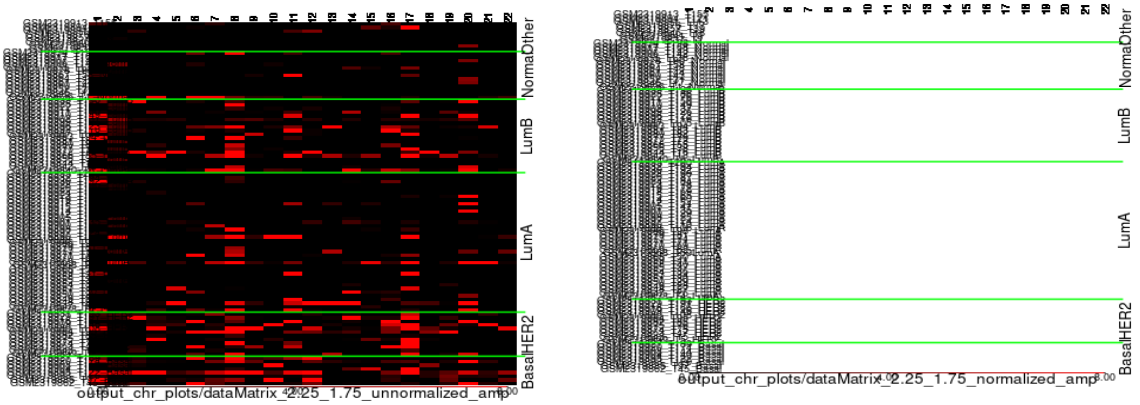

NormalOther
LumB
LumA
BasalHER2
output_chr_plots/dataMatrix_2.5_1.5_unnormalized_del
output_chr_plots/dataMatrix_2.5_1.5_normalized_del

**C)**

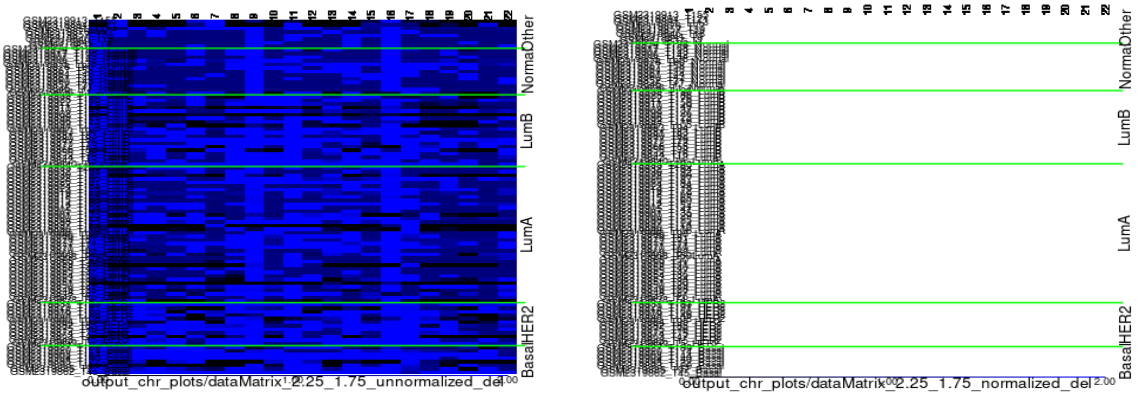

NormalOther
LumB
LumA
BasalHER2
output_chr_plots/dataMatrix_2.5_1.5_unnormalized_sum
output_chr_plots/dataMatrix_2.5_1.5_normalized_sum

**Supplementary Figure 2.** Visualization of CNV using CINdex across the genome with threshold (gain = 2.5 and loss = 1.5). A) unnormalized amplification events; B) unnormalized deletion event; C) unnormalized sum event.

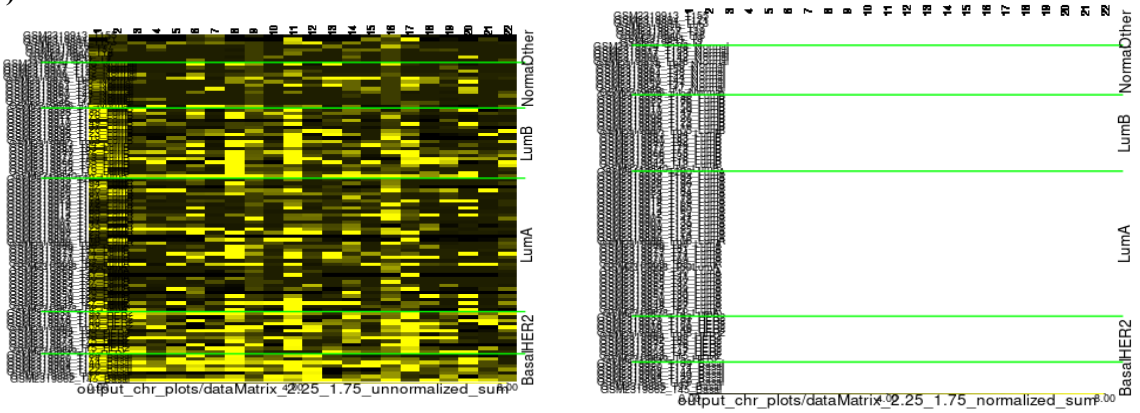
**A)**



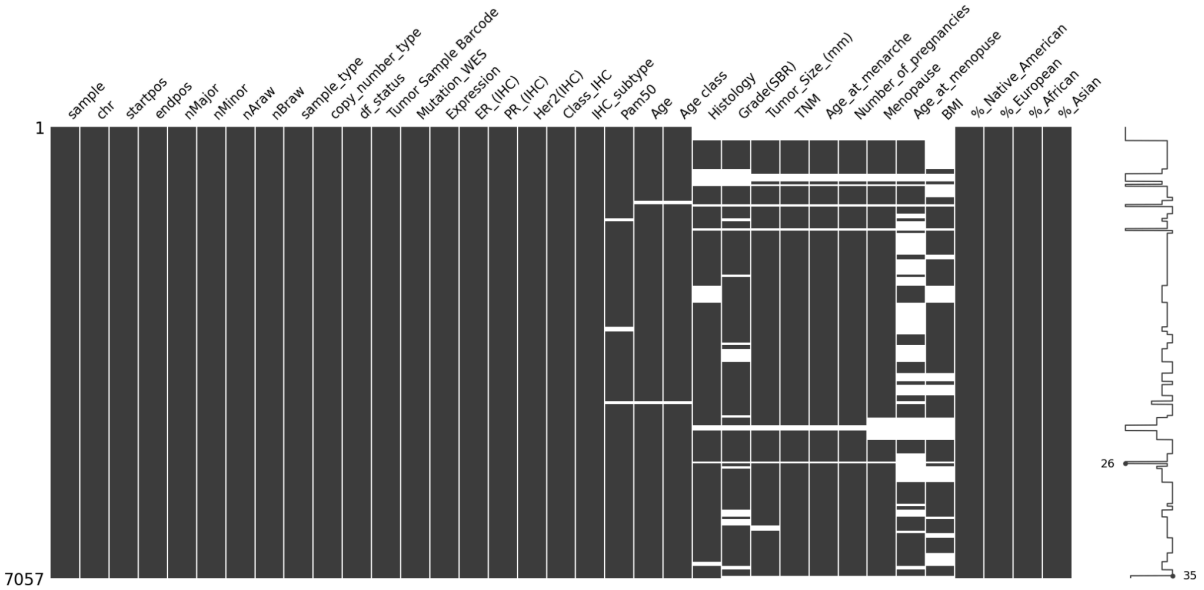

**B)**




**C)**



**Supplementary Figure 3**. Visualization of CNV using CINdex across the genome with threshold (gain = 2.25 and loss = 1.75). A) unnormalized amplification events; B) unnormalized deletion event; C) unnormalized sum event.




**Supplementary Figure 4**. Original Data Result of Missingno package depicting the missing values found in the original dataset.

**Supplementary Table 1**. LazyPredict results before stratification

| Model | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | Time Taken |
|---|---|---|---|---|---|
| LGBMClassifier | 0.63 | 0.55 | None | 0.62 | 0.80 |
| XGBClassifier | 0.62 | 0.54 | None | 0.61 | 0.36 |
| BaggingClassifier | 0.57 | 0.50 | None | 0.57 | 0.22 |
| KNeighborsClassifier | 0.57 | 0.49 | None | 0.56 | 0.03 |
| LinearDiscriminantAnalysis | 0.57 | 0.49 | None | 0.56 | 0.02 |
| DecisionTreeClassifier | 0.55 | 0.48 | None | 0.55 | 0.04 |
| SVC | 0.60 | 0.48 | None | 0.56 | 0.80 |
| RandomForestClassifier | 0.55 | 0.47 | None | 0.54 | 0.43 |
| LogisticRegression | 0.57 | 0.47 | None | 0.55 | 0.05 |
| RidgeClassifier | 0.58 | 0.46 | None | 0.54 | 0.02 |
| RidgeClassifierCV | 0.58 | 0.46 | None | 0.54 | 0.02 |
| CalibratedClassifierCV | 0.57 | 0.46 | None | 0.54 | 0.91 |

| | | | | | |
|---|---|---|---|---|---|
| LinearSVC | 0.57 | 0.46 | None | 0.54 | 0.22 |
| SGDClassifier | 0.53 | 0.45 | None | 0.52 | 0.18 |
| AdaBoostClassifier | 0.56 | 0.45 | None | 0.53 | 0.21 |
| NuSVC | 0.58 | 0.44 | None | 0.53 | 0.99 |
| ExtraTreeClassifier | 0.50 | 0.44 | None | 0.51 | 0.02 |
| ExtraTreesClassifier | 0.50 | 0.43 | None | 0.50 | 0.45 |
| QuadraticDiscriminantAnalysis | 0.40 | 0.43 | None | 0.42 | 0.02 |
| NearestCentroid | 0.42 | 0.43 | None | 0.43 | 0.01 |
| Perceptron | 0.48 | 0.42 | None | 0.47 | 0.02 |
| LabelSpreading | 0.48 | 0.41 | None | 0.48 | 0.60 |
| LabelPropagation | 0.48 | 0.41 | None | 0.47 | 0.39 |
| GaussianNB | 0.35 | 0.41 | None | 0.35 | 0.02 |
| BernoulliNB | 0.45 | 0.40 | None | 0.45 | 0.02 |
| PassiveAggressiveClassifier | 0.46 | 0.40 | None | 0.46 | 0.04 |
| DummyClassifier | 0.43 | 0.25 | None | 0.26 | 0.01 |

**Supplementary Table 2**. LazyPredict results after SMOTE

| Model | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | Time Taken |
|---|---|---|---|---|---|
| XGBClassifier | 0.72 | 0.72 | None | 0.72 | 0.39 |
| LGBMClassifier | 0.71 | 0.71 | None | 0.71 | 0.84 |
| BaggingClassifier | 0.71 | 0.71 | None | 0.71 | 0.52 |
| RandomForestClassifier | 0.69 | 0.69 | None | 0.69 | 0.88 |
| ExtraTreesClassifier | 0.68 | 0.68 | None | 0.68 | 0.89 |
| SVC | 0.67 | 0.67 | None | 0.68 | 2.90 |
| NuSVC | 0.67 | 0.66 | None | 0.67 | 4.21 |
| DecisionTreeClassifier | 0.66 | 0.66 | None | 0.66 | 0.09 |
| LabelPropagation | 0.65 | 0.65 | None | 0.65 | 1.54 |

| | | | | | |
|---|---|---|---|---|---|
| KNeighborsClassifier | 0.65 | 0.65 | None | 0.65 | 0.05 |
| LabelSpreading | 0.65 | 0.65 | None | 0.65 | 2.22 |
| ExtraTreeClassifier | 0.62 | 0.62 | None | 0.62 | 0.03 |
| LogisticRegression | 0.59 | 0.59 | None | 0.59 | 0.10 |
| LinearSVC | 0.58 | 0.58 | None | 0.58 | 1.42 |
| CalibratedClassifierCV | 0.58 | 0.58 | None | 0.58 | 4.47 |
| LinearDiscriminantAnalysis | 0.58 | 0.58 | None | 0.58 | 0.04 |
| RidgeClassifier | 0.58 | 0.58 | None | 0.58 | 0.03 |
| RidgeClassifierCV | 0.57 | 0.58 | None | 0.58 | 0.04 |
| AdaBoostClassifier | 0.56 | 0.56 | None | 0.56 | 0.40 |
| SGDClassifier | 0.55 | 0.55 | None | 0.55 | 0.39 |
| PassiveAggressiveClassifier | 0.48 | 0.49 | None | 0.48 | 0.06 |
| Perceptron | 0.50 | 0.50 | None | 0.50 | 0.05 |
| NearestCentroid | 0.46 | 0.46 | None | 0.46 | 0.02 |
| BernoulliNB | 0.46 | 0.46 | None | 0.45 | 0.03 |
| GaussianNB | 0.44 | 0.44 | None | 0.41 | 0.03 |
| QuadraticDiscriminantAnalysis | 0.42 | 0.42 | None | 0.33 | 0.04 |
| DummyClassifier | 0.25 | 0.25 | None | 0.10 | 0.02 |

**Supplementary Table 3**. LazyPredict results after SMOTE+TOMEK

| Model | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | Time Taken |
|---|---|---|---|---|---|
| XGBClassifier | 0.78 | 0.78 | None | 0.78 | 0.40 |
| LGBMClassifier | 0.77 | 0.77 | None | 0.77 | 0.86 |
| BaggingClassifier | 0.76 | 0.76 | None | 0.76 | 0.45 |
| RandomForestClassifier | 0.76 | 0.76 | None | 0.76 | 0.76 |
| ExtraTreesClassifier | 0.75 | 0.75 | None | 0.75 | 0.75 |
| DecisionTreeClassifier | 0.73 | 0.73 | None | 0.73 | 0.08 |

| | | | | | |
|---|---|---|---|---|---|
| SVC | 0.71 | 0.71 | None | 0.71 | 2.19 |
| ExtraTreeClassifier | 0.69 | 0.69 | None | 0.69 | 0.03 |
| KNeighborsClassifier | 0.69 | 0.69 | None | 0.69 | 0.05 |
| LabelPropagation | 0.69 | 0.68 | None | 0.68 | 1.27 |
| LabelSpreading | 0.69 | 0.68 | None | 0.68 | 1.74 |
| NuSVC | 0.68 | 0.68 | None | 0.68 | 3.31 |
| LogisticRegression | 0.63 | 0.63 | None | 0.63 | 0.10 |
| CalibratedClassifierCV | 0.62 | 0.62 | None | 0.62 | 4.21 |
| RidgeClassifierCV | 0.62 | 0.62 | None | 0.62 | 0.04 |
| RidgeClassifier | 0.61 | 0.61 | None | 0.61 | 0.02 |
| LinearSVC | 0.61 | 0.61 | None | 0.61 | 1.07 |
| LinearDiscriminantAnalysis | 0.61 | 0.61 | None | 0.61 | 0.04 |
| AdaBoostClassifier | 0.58 | 0.58 | None | 0.58 | 0.35 |
| SGDClassifier | 0.58 | 0.58 | None | 0.58 | 0.31 |
| Perceptron | 0.52 | 0.52 | None | 0.51 | 0.05 |
| PassiveAggressiveClassifier | 0.51 | 0.51 | None | 0.51 | 0.07 |
| NearestCentroid | 0.46 | 0.46 | None | 0.45 | 0.02 |
| BernoulliNB | 0.44 | 0.45 | None | 0.44 | 0.03 |
| QuadraticDiscriminantAnalysis | 0.43 | 0.44 | None | 0.33 | 0.03 |
| GaussianNB | 0.40 | 0.40 | None | 0.34 | 0.03 |
| DummyClassifier | 0.25 | 0.25 | None | 0.10 | 0.02 |

**Supplementary Table 4**. LazyPredict results after SMOTE+ENN

| Model | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | Time Taken |
|---|---|---|---|---|---|
| ExtraTreesClassifier | 0.95 | 0.95 | None | 0.95 | 0.27 |
| RandomForestClassifier | 0.94 | 0.94 | None | 0.94 | 0.29 |
| LGBMClassifier | 0.94 | 0.94 | None | 0.94 | 0.83 |

| | | | | | |
|---|---|---|---|---|---|
| XGBClassifier | 0.93 | 0.93 | None | 0.93 | 0.33 |
| ExtraTreeClassifier | 0.91 | 0.90 | None | 0.91 | 0.02 |
| BaggingClassifier | 0.90 | 0.90 | None | 0.90 | 0.17 |
| DecisionTreeClassifier | 0.90 | 0.90 | None | 0.90 | 0.04 |
| LabelSpreading | 0.88 | 0.88 | None | 0.88 | 0.38 |
| LabelPropagation | 0.88 | 0.88 | None | 0.88 | 0.27 |
| SVC | 0.86 | 0.86 | None | 0.86 | 0.39 |
| KNeighborsClassifier | 0.84 | 0.84 | None | 0.84 | 0.04 |
| NuSVC | 0.83 | 0.83 | None | 0.83 | 0.65 |
| LogisticRegression | 0.76 | 0.76 | None | 0.76 | 0.05 |
| LinearSVC | 0.75 | 0.75 | None | 0.75 | 0.62 |
| CalibratedClassifierCV | 0.75 | 0.75 | None | 0.75 | 2.47 |
| LinearDiscriminantAnalysis | 0.74 | 0.74 | None | 0.74 | 0.02 |
| RidgeClassifier | 0.73 | 0.73 | None | 0.73 | 0.02 |
| RidgeClassifierCV | 0.73 | 0.73 | None | 0.73 | 0.02 |
| SGDClassifier | 0.72 | 0.72 | None | 0.72 | 0.13 |
| Perceptron | 0.69 | 0.69 | None | 0.69 | 0.03 |
| PassiveAggressiveClassifier | 0.67 | 0.68 | None | 0.67 | 0.04 |
| AdaBoostClassifier | 0.63 | 0.64 | None | 0.63 | 0.18 |
| QuadraticDiscriminantAnalysis | 0.54 | 0.58 | None | 0.49 | 0.02 |
| NearestCentroid | 0.57 | 0.57 | None | 0.57 | 0.01 |
| BernoulliNB | 0.55 | 0.55 | None | 0.55 | 0.02 |
| GaussianNB | 0.53 | 0.51 | None | 0.48 | 0.02 |
| DummyClassifier | 0.31 | 0.25 | None | 0.15 | 0.01 |