

Multimodal Data Analytics For Predicting The Survival Of Cancer Patients

Supervisor:

Prof. Rashid Ali

Department of Computer Engineering

Z.H.C.E.T, A.M.U

Presented By:

Hasan Shaikh

Department of Computer Engineering

Z.H.C.E.T, A.M.U





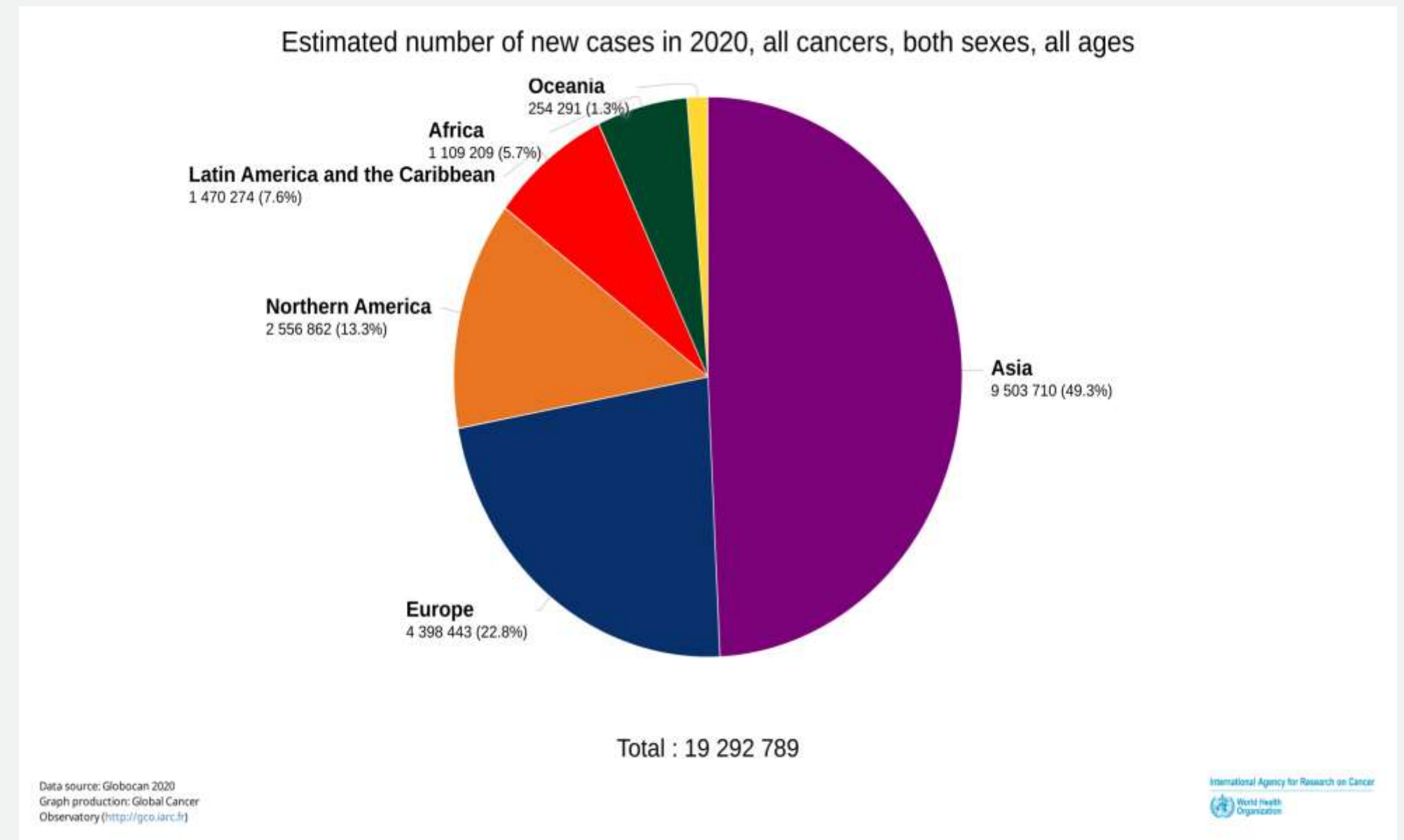
OVERVIEW

1. Introduction
2. Literature Review
3. Research Gap
4. Objective of the Work
5. Design of Cancer Survival prediction system
6. Implementation of cancer survival prediction system
7. Experiments and Results
8. Conclusion & Future Work
9. References

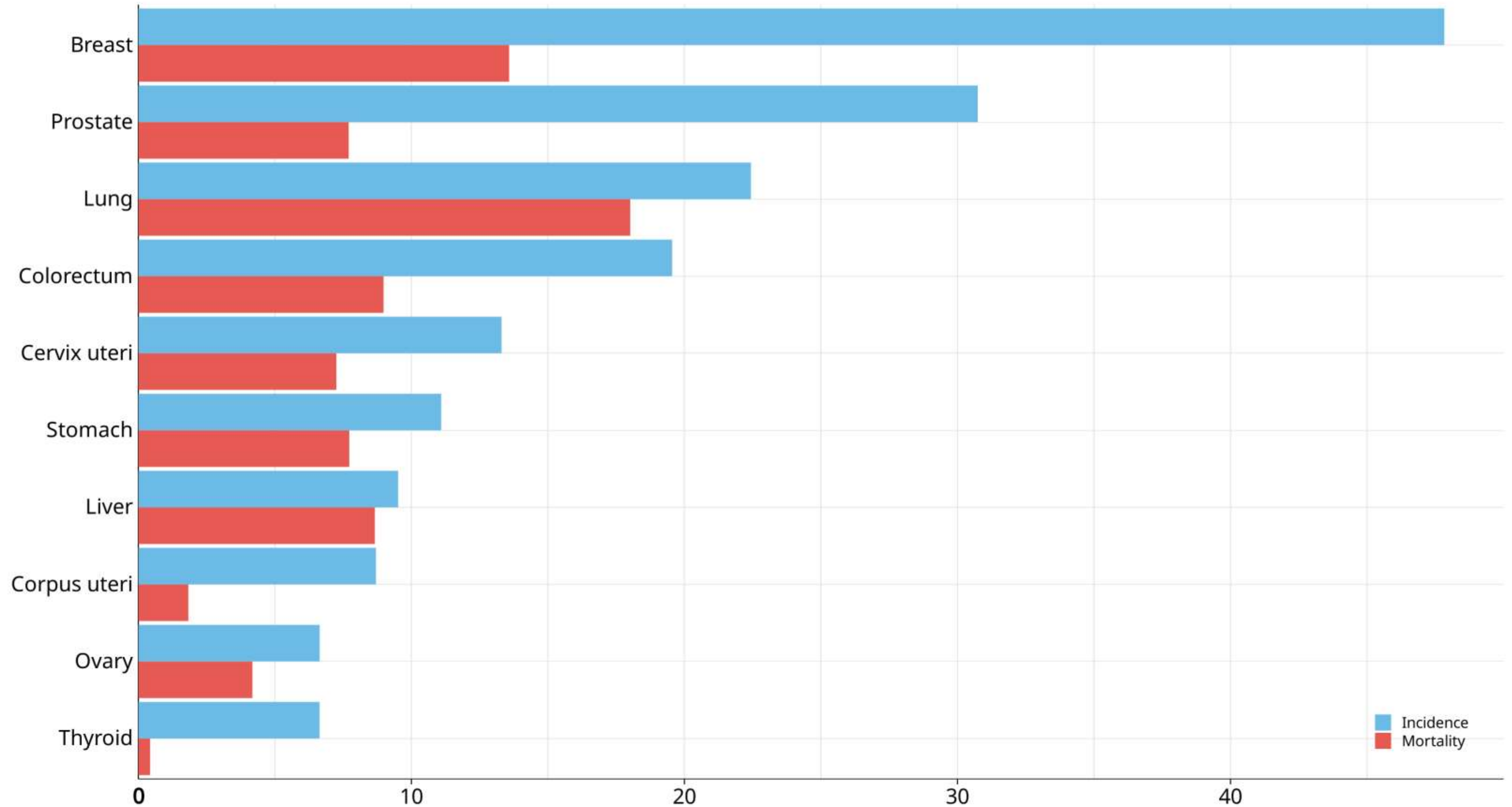


INTRODUCTION – Understanding the Problem

- **Cancer?:** Uncontrolled growth and spread of abnormal cells.
- **Global Impact:** Positioned as a major health challenge worldwide.

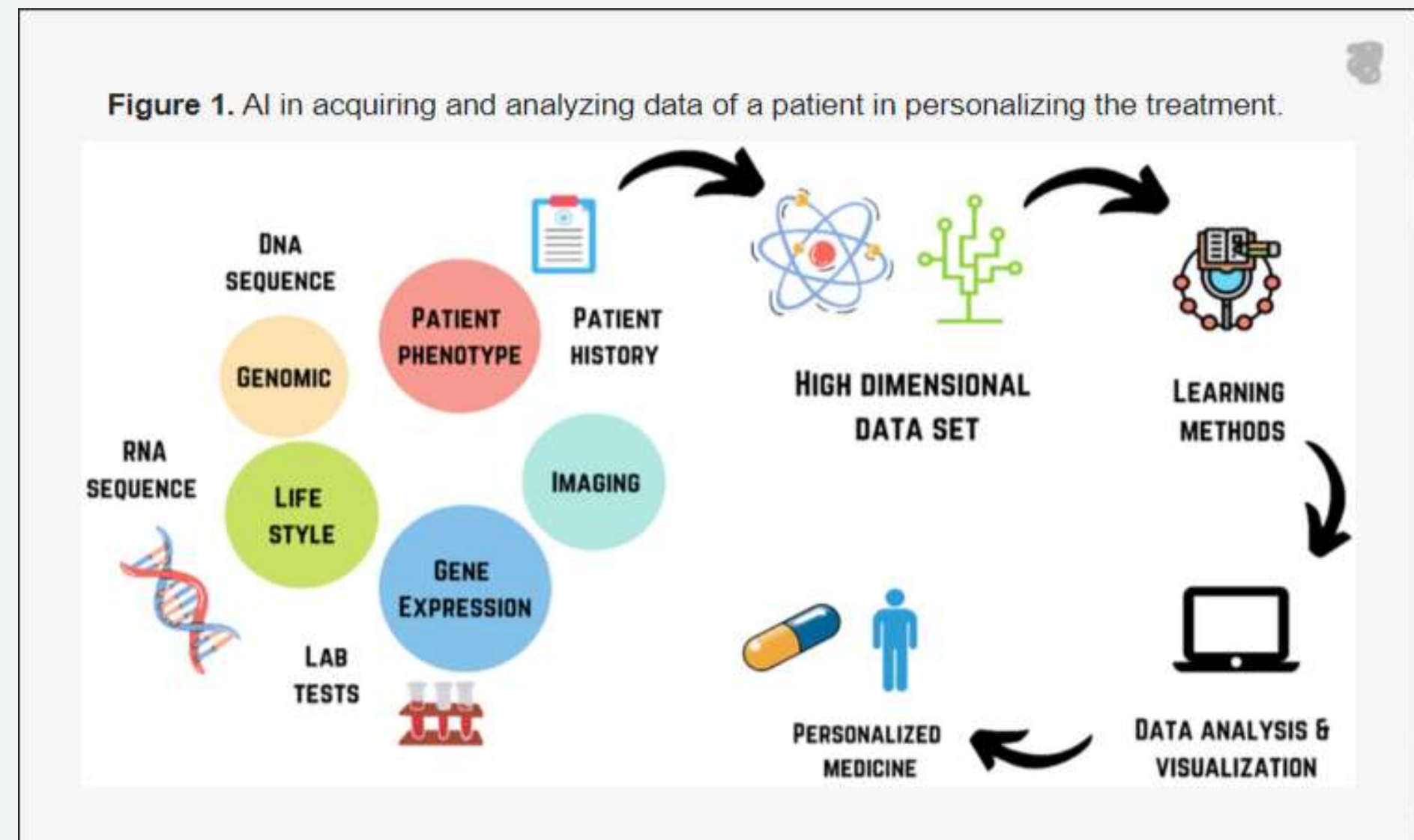


Estimated age-standardized incidence and mortality rates (World) in 2020, World, both sexes, all ages (excl. NMSC)



INTRODUCTION – Significance of Cancer Survival Prediction

- **Purpose of Cancer Patients Survival Prediction:** Tailoring individualized treatment plans and improving patient outcomes.
- **Resource Management:** Efficiently allocation of medical resources and early interventions identification.



INTRODUCTION – Challenges in Predictive Analysis

- **Patient Diversity:** Vast heterogeneity in demographic, genetic, and lifestyle factors affecting predictions.
- **Cancer's Complexity:** Various subtypes, progression rates, and treatment responses.

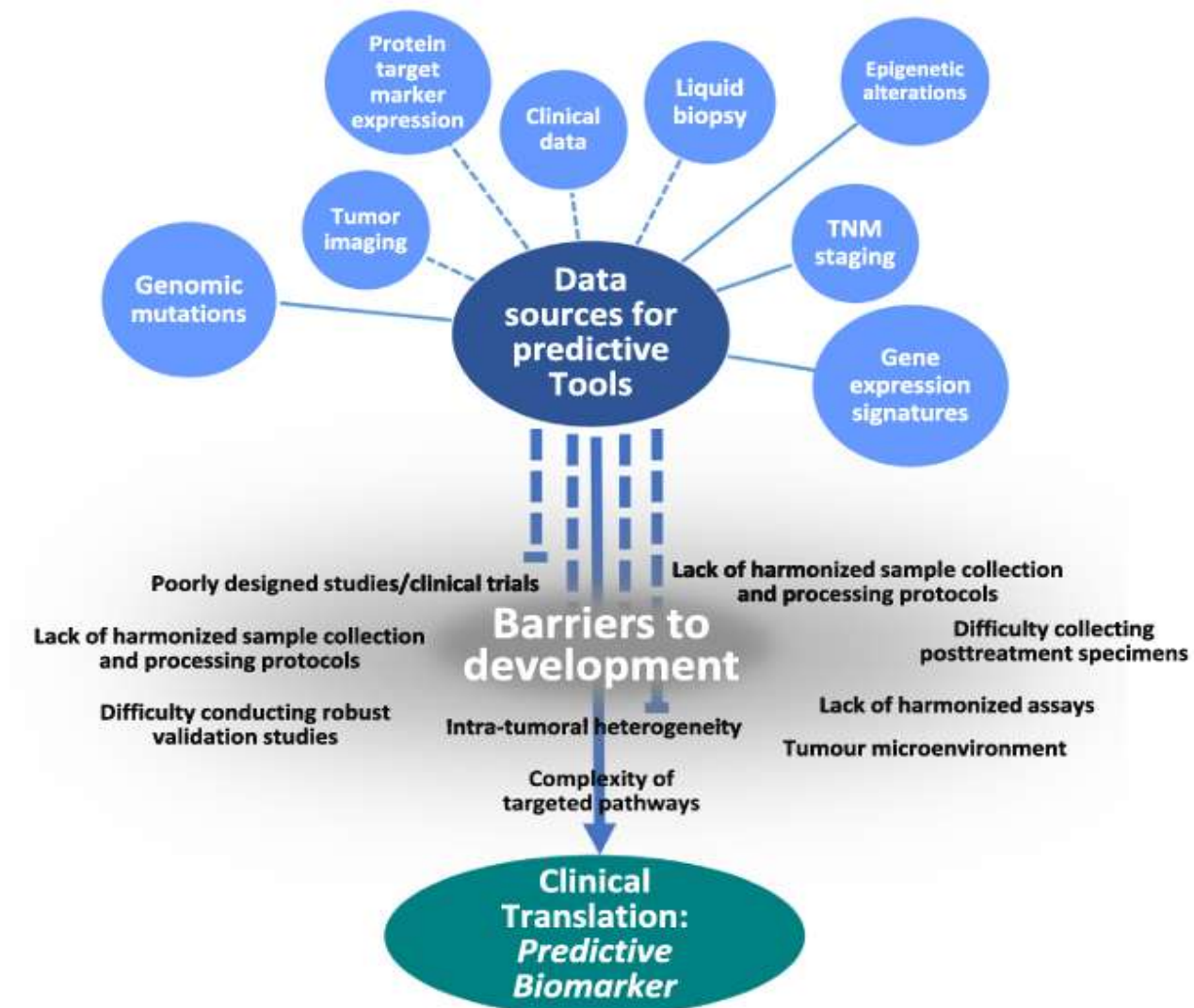


Fig. 1. Pictorial representation of data sources for predictive biomarker development and barriers that prevent successful clinical translation. Individual data variables (blue) may be predictive but some may be prognostic (such as TNM) but in combination form a predictive tool.

Figure 2 Source: Batis, N., Brooks, J. M., Payne, K., & Mehanna, H. (2021). Lack of predictive tools for conventional and targeted cancer therapy: Barriers to biomarker development and clinical translation. *Advanced drug delivery reviews*, 176, 113854. <https://doi.org/10.1016/j.addr.2021.113854>

MULTIMODAL ANALYSIS

- **What it is:** An approach that merges multiple data sources.
- **Why it Matters:** Offers a holistic view of diseases; essential for cancer prognosis and role in Oncology.
- Use a combination of clinical data, gene expression data and copy number alteration (CNA) data.

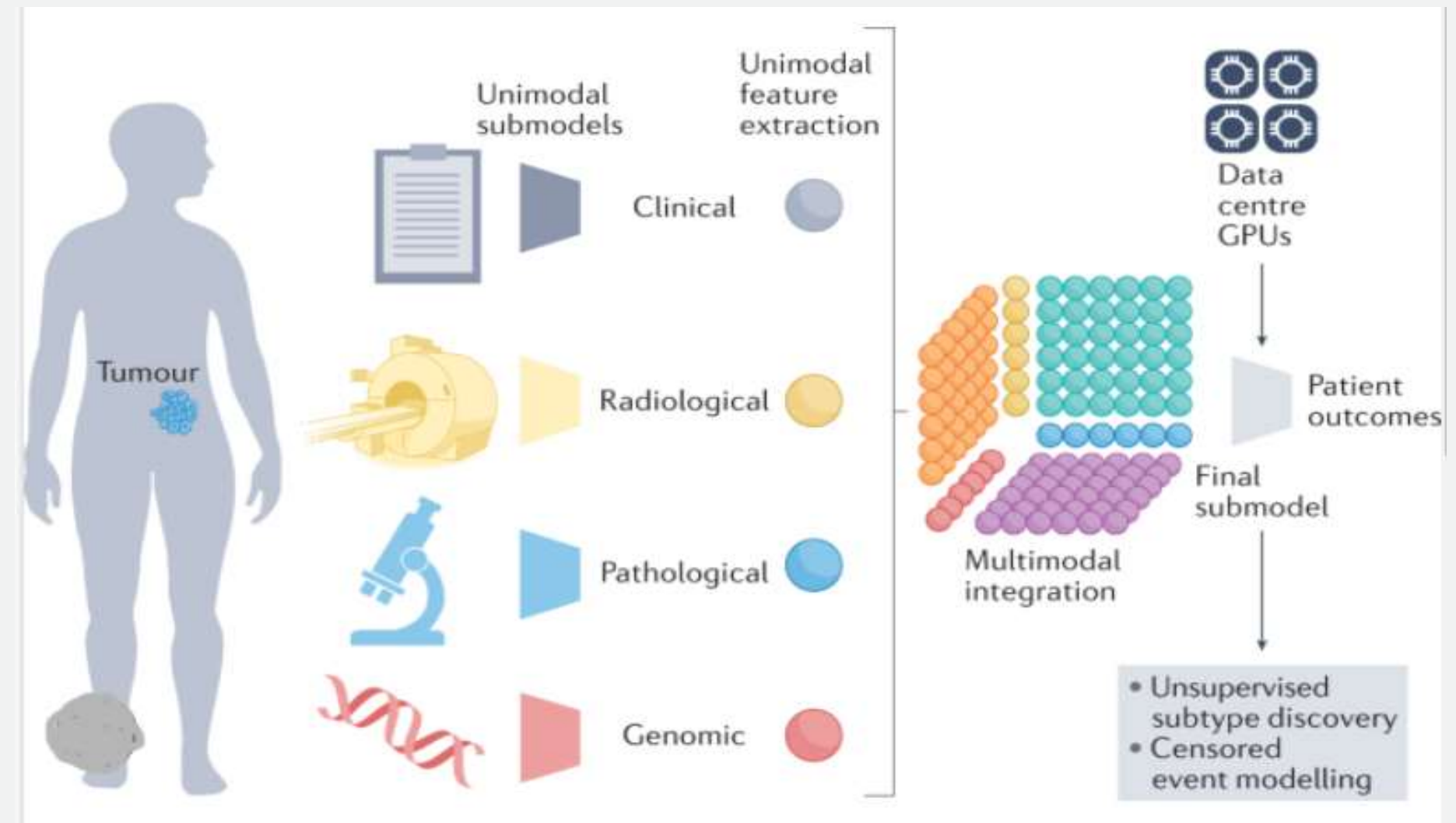


Figure 3: Multimodal models integrate features across modalities

Figure 3 Source: Boehm, Kevin & Khosravi, Pegah & Vanguri, Rami & Gao, Jianjiong & Shah, Sohrab. (2021). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*. 22. 1-13. 10.1038/s41568-021-00408-3.

Multimodal Data Analysis- Clinical Features

1. Clinical Features:

- Collection of medical records including
- age,
- gender,
- diagnosis,
- treatment received, and
- other personal health data of a patient.
- Essential for understanding patient health, predicting outcomes, and planning treatments.



Multimodal Data Analysis- Clinical Features

1. age at diagnosis:

- This parameter refers to the age of the patient at the time they were diagnosed with breast cancer.

2. type of breast surgery:

- This details the surgical procedure undertaken to treat the breast cancer, which might be a lumpectomy (removal of the tumor and a small portion of surrounding tissue) or mastectomy (removal of all breast tissue).

3. cancer type:

- This parameter indicates the general type of cancer that has been diagnosed, which could be invasive or non-invasive, referring to whether the cancer has spread beyond the milk ducts or lobules of the breast.

Multimodal Data Analysis- Clinical Features (Cont...)

4. **cancer type detailed:**

- This gives more detailed information about the specific subtype of breast cancer.

5. **cellularity:**

- This refers to the abundance of cells in the tumor sample. Low cellularity indicates that the tumor sample contains fewer cancer cells and more non-cancerous cells, while high cellularity indicates a higher proportion of cancer cells.

6. **chemotherapy:**

- This parameter indicates whether the patient underwent chemotherapy as a part of their treatment. It can be a binary variable (Yes/No) indicating the application or non-application of chemotherapy.

Multimodal Data Analysis- Clinical Features (Cont...)

7. **pam50 + claudin-low subtype:**

- PAM50 is a gene expression profile test that categorizes breast cancers into one of the four intrinsic or molecular subtypes (Luminal A, Luminal B, HER2-enriched, and Basal-like). Claudin-low is another molecular subtype, characterized by low expression of cell-cell adhesion molecules.

8. **cohort:**

- This could refer to a specific group or subgroup of patients being studied. Cohorts might be defined by specific characteristics, like disease stage .

9. **er status measured by ihc:**

- ER (Estrogen Receptor) status, as measured by Immunohistochemistry (IHC), indicates whether the breast cancer cells are positive or negative for estrogen receptors. ER-positive cancers may respond to hormone therapy.

Multimodal Data Analysis- Clinical Features (Cont...)

10. nottingham prognostic index:

- This is a prognostic tool that helps to predict the survival of patients diagnosed with breast cancer. It takes into account the size of the tumor, the degree to which it has spread to nearby lymph nodes, and the grade of the tumor (how abnormal the cells look under a microscope).

11. overall survival months:

- This parameter indicates the total number of months from diagnosis that the patient has survived.

12. overall survival:

- This might be a binary parameter indicating whether the patient is alive (1) or has passed away (0) at the time of the last follow-up.

Multimodal Data Analysis- Gene Expression Features

2. Gene Expression Features:

- Refers to the process where specific genes are activated to produce a required protein.
- Levels of expression can indicate how actively a gene is producing the protein for which it codes.
- Features:
 - **Under-expression (-1):** Gene activity is lower than the normal baseline; this might indicate the absence or malfunction of certain proteins.
 - **Over-expression (1):** Gene activity is higher than the normal baseline; could signify overactive protein production.
 - **Baseline (0):** Normal gene activity; indicates standard protein production.



Multimodal Data Analysis- **Copy Number Alteration Features**

3. **Copy Number Alteration (CNA) Features:**

- Refers to the variation in the number of copies of a particular gene in the genome of a cell.
- Changes in copy number can affect the gene's activity and the amount of protein it produces.
- Features:
 - **Homozygous deletion (-2):** Both copies of the gene (from each parent) are deleted. This likely results in a complete loss of gene function.
 - **Hemizygous deletion (-1):** One of the two copies of the gene is deleted. This can reduce the gene's activity by half.
 - **Neutral/no change (0):** The gene has the expected two copies, suggesting no alterations.
 - **Gain (1):** There's an extra copy (or copies) of the gene, potentially leading to increased gene activity.
 - **High-level amplification (2):** Many extra copies of the gene exist, which can significantly boost its activity.



LITERATURE REVIEW

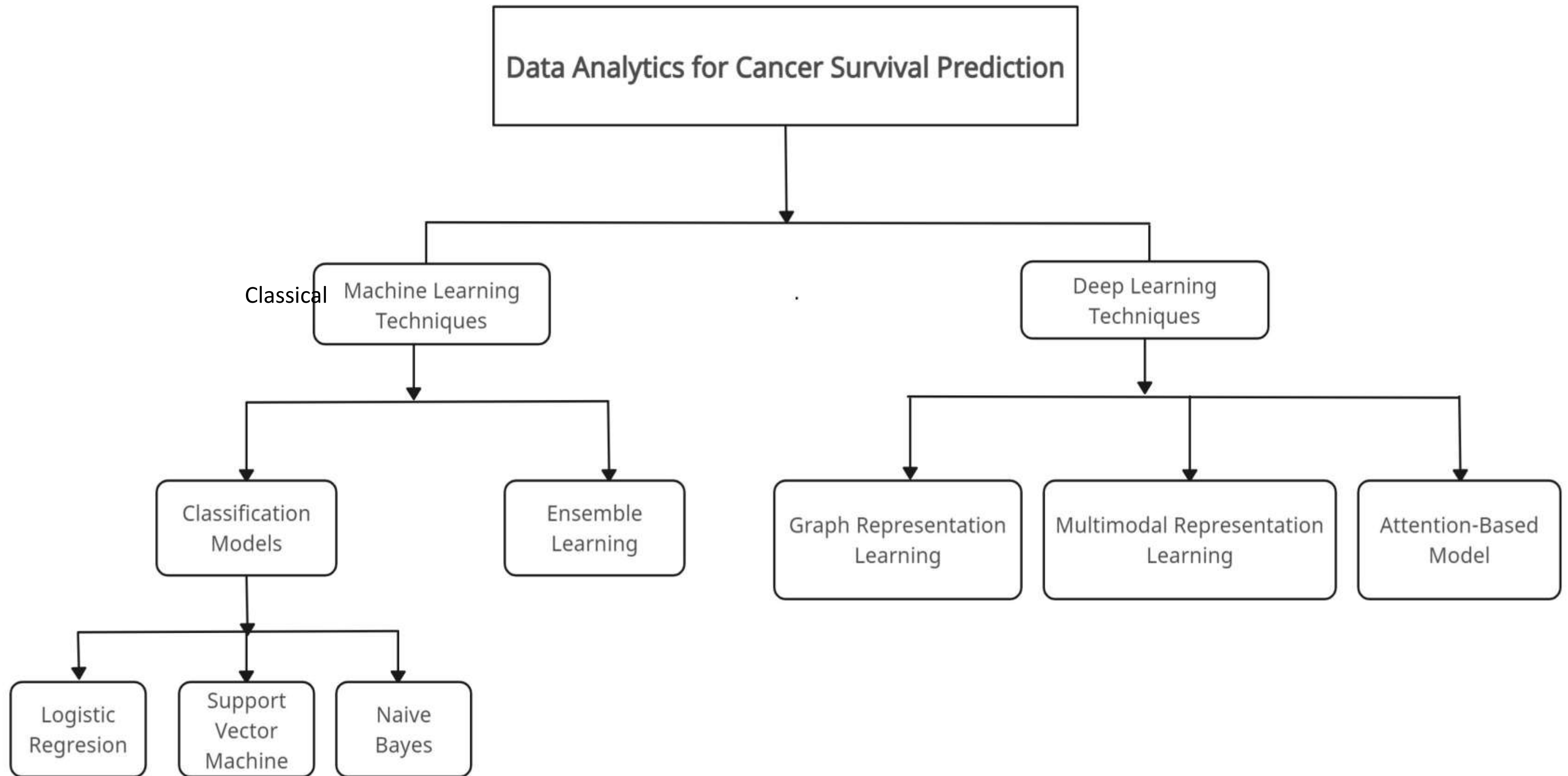


Figure 4: Various ML Techniques used for cancer survival prediction

TABLE 1: Overview of Classical ML Techniques in Cancer Survival Prediction

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2023	Bozkurt et al. [1]	Classification-based approach using multiple classifier, Logistic Regression for feature selection	Breast, Lung, Prostate, Stomach	Medical Information Mart in Intensive Care IV (MIMIC-IV)	Using fewer features is efficient
2023	Arya et al. [2]	Principal Component Analysis, Variational Autoencoders, Support Vector Machine, Random Forest	Breast	The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA)	More modalities = more robustness
2023	Zolfaghari et al. [3]	Ensemble Classifiers incorporating deep learning	Multiple Cancer Types	Not Specified	Review on ensemble methods used in cancer prognosis and diagnosis

TABLE 1: Overview of Classical ML Techniques in Cancer Survival Prediction (Cont...)

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2022	Azar et al. [4]	K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, XGBoost, Shapley Additive Explanations (SHAP)	Ovarian	Surveillance, Epidemiology, and End Results database (SEER)	Random Forest and XGBoost achieved the best performance for classification and regression respectively
2022	Yan et al. [5]	Priori knowledge- and stability-based feature selection (PKSFS), two-stage heterogeneous stacked ensemble learning model (BQAXR)	Gastric, Skin	Surveillance, Epidemiology, and End Results database (SEER)	PKSFS performed well in processing high-dimensional datasets. BQAXR outperformed mainstream ML
2020	S. P. et al. [6]	Logistic Regression-based models within a Quantitative Radiomic Framework	Lung	CT images	Use of advanced imaging techniques for predicting lung cancer patient survival

TABLE 1: Overview of Classical ML Techniques in Cancer Survival Prediction (Cont...)

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2018	Bartholomai et al. [7]	ANOVA for factor selection, RF (for classification and regression), Linear Regression, Gradient Boosted Machines (GBM)	Lung	Surveillance, Epidemiology, and End Results database (SEER)	Random Forest performed best for survival times ≤ 6 and > 24 months, while Gradient Boosted Machines performed best for 7–24 months
2017	Lynch et al. [8]	Linear regression, Decision Tree, Gradient Boosted Machines, Support Vector Machine, Ensemble	Lung	Surveillance, Epidemiology, and End Results database (SEER)	Ensemble had best performance
2016	Montazeri et al. [9]	ML techniques like Naïve Bayes, Tree Random Forest, 1NN, AD, Support Vector Machine, RBFN, MLP with 10-cross fold	Breast	Specific dataset with records of 900 patients	Tree Random Forest was the best model with the highest accuracy

TABLE 2: Overview of **DL** Techniques in Cancer Survival Prediction

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2023	Wu et al. [10]	Cross-Aligned Multimodal Representation Learning (CAMR)	Multiple (including Breast, Lung and Brain)	Three cancer datasets obtained from TCGA including LGG, Breast Invasive Carcinoma (BRCA) and LUSC	CAMR effectively reduces modality gaps, generating both modality-invariant and -specific representations for enhanced cancer survival prediction.
2023	Fan et al. [11]	Multimodal integrative DL with unsupervised representation learning	Pancancer	TGCA, University of California Santa Cruz Xena (UCSC Xena) (including clinical data, gene expression (mRNA) data, microRNA expression (miRNA) data and gene copy number variation (CNV) data)	Model performs best using clinical and mRNA modalities, Adding more data modalities risks overfitting, Including the CNV modality reduced prediction performance, potentially due to introducing noise.
2022	Arya et al. [12]	Stacked-based ensemble model architecture	Breast	Multi-modal datasets (gene expression profile, copy number alteration, clinical data)	CNN for feature extraction then stacked-based approach utilizing three modalities of data improves predictive performance especially for imbalanced datasets

TABLE 2: Overview of **DL** Techniques in Cancer Survival Prediction (Cont...)

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2022	Kanwal et al. [13]	Artificial Algae Algorithm (AAA) for feature extraction combined with Double DEEP Q-NETWORK (DDQN), Convolution eXtreme Gradient Boosting (CNN-XGBOOST), and Convolution Support Vector Machine (CNN-SVM)	Multiple (including Brain, Prostate, Bladder, Colorectal and Breast)	Lower Grade Glioma in the Brain (BRAIN-TCGA), Prostate Cancer Dataset, Bladder Cancer Dataset (BLADDER-TCGA), Metastatic Colorectal Cancer Dataset (MSKCC), and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	A novel framework was introduced that integrates DL/ML and RL with AAA for improved cancer prognosis prediction using multimodal data, incorporating early and late fusion techniques
2022	Li et al. [14]	Distribution-based Multiple-Instance Survival Learning algorithm (DeepDisMISL)	Colorectal	Two international Colorectal Cancer (CRC) datasets: MCO CRC and TCGA COAD-READ	Combining percentile-scored patches with highest and lowest scored ones , Including neighborhood instances around percentiles further boost prediction accuracy

TABLE 2: Overview of **DL** Techniques in Cancer Survival Prediction (Cont...)

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2022	Wu et al. [15]	Stacked Autoencoder-based Survival Prediction Neural Network (SAEsurv-net)	Glioblastoma multiforme (GBM), Ovarian serous cystadenocarcinoma (OV), Breast invasive carcinoma (BRCA)	Gene expression, Copy number variations (CNV), Clinical information for GBM, OV, and BRCA datasets from The Cancer Genome Atlas (TCGA) project via UCSC Xena	SAEsurv-net outperforms single-data-type models and other state-of-the-art methods, Effective handling of multi-omics heterogeneity and dimensionality reduction
2019	Cheerla et al. [16]	Developed a DL survival model with multimodal representation	Obtained from TCGA (including 20 different cancer types)	Multimodal dataset (including clinical, genomic and Whole Slide Image (WSIs))	Demonstrated efficient use of multimodal data, even with missing modalities, Proposed efficient While Slide Image analysis by sampling key Region of Interests.

TABLE 2: Overview of **DL** Techniques in Cancer Survival Prediction (Cont...)

Year of the Study	Authors	Techniques	Cancer Type	Datasets	Important Finding
2019	Huang et al. [17]	Survival Analysis Learning with Multi-Omics Neural Networks (SALMON)	Breast	583 female breast invasive carcinoma patients with multi-omics data types (gene expression, miRNA data, copy number burden, tumor mutation burden, clinical information)	Eigengene matrices reduce dimensionality and enhance model robustness, Improved performance with more omics data, Age-based stratification of prognosis.
2018	Sun et al. [18]	Multimodal Deep Neural Network by integrating Multi-dimensional Data (MDNNMD)	Breast	Multi-modal datasets (gene expression profile, copy number alteration, clinical data)	MDNNMD integrates multi-dimensional data for better prognosis prediction; outperforms single-dimensional methods

RESEARCH GAP

Limitations of the existing Cancer Survival Prediction Models:

- Most of the existing prediction models *exclude essential data modalities*, hindering performance.
- Model performance is not up to the mark due to *lack of data integration*.
- Very little work has been done on multimodal data analysis based on cancer survival prediction.
- *Overfitting and noise* introduction are potential risks when indiscriminately adding data modalities.



OBJECTIVE OF THE WORK

Objective 1: Exploratory Analysis of Multimodal Data used for cancer survival prediction:

- In this work we try to understand the effect of multi-modality on the performance of cancer survival prediction models.
- Conducted experiments on both unimodal and multimodal datasets.
- Investigated the limitations of traditional ML techniques in capturing intricate relationships between modalities.



OBJECTIVE OF THE WORK (Cont...)

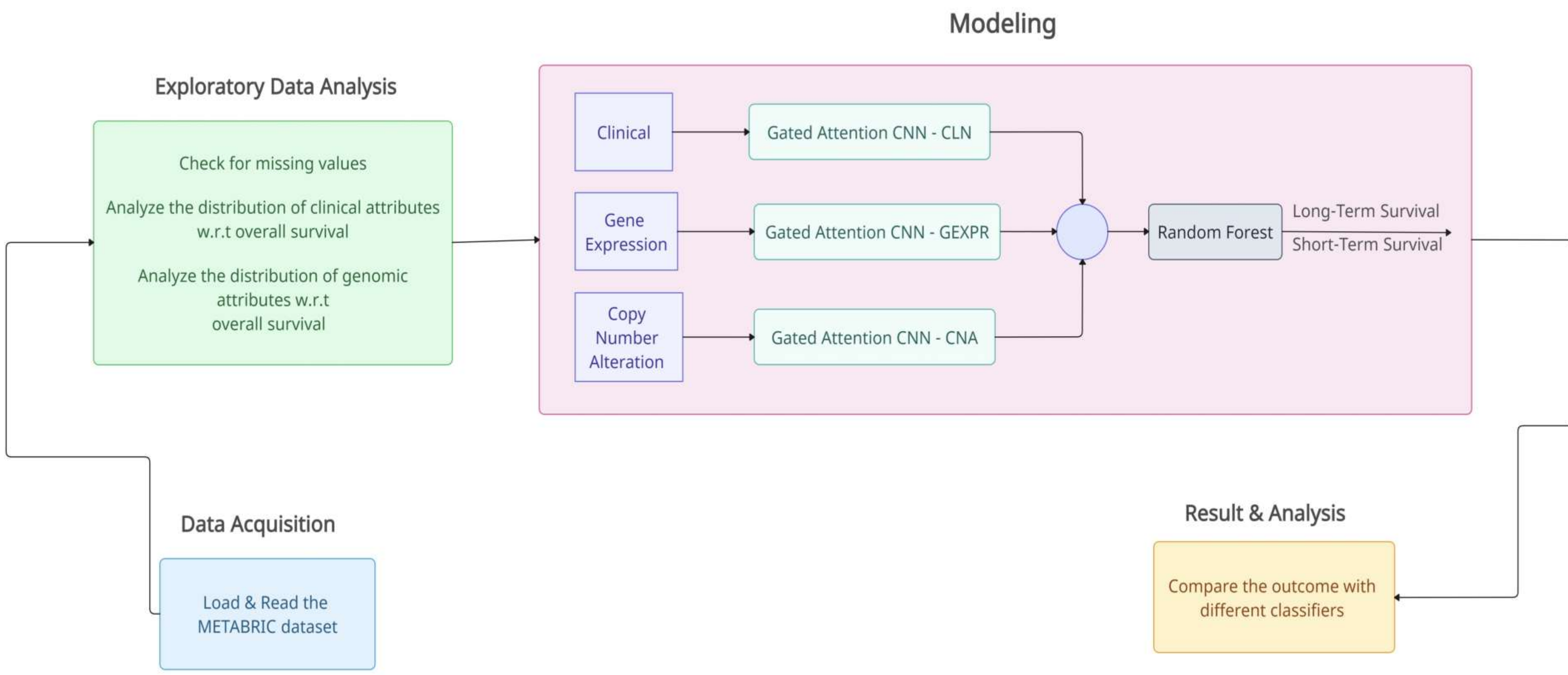
Objective 2: Use of Deep learning for multimodal data analysis based cancer survival prediction:

- In this work we try to leverage the power of deep learning for improved predictions using multimodal data.
- Specifically, we propose to use a popular deep learning architecture called "Gated Attention CNN" for multimodal data analysis based cancer survival prediction.
- We also built separate models for each modality, extracting unique features.
- Then, integrated features across all modalities to create a comprehensive model.
- We aimed to understand how combining features from all modalities enhances the model's ability to capture complex inter-modality relationships.





DESIGN OF CANCER SURVIVAL PREDICTION SYSTEM



Proposed Architecture for Cancer Survival Prediction using Multimodal Dataset

EXPLORATORY DATA ANALYSIS

Age at diagnosis:

- Survivors: Bimodal distribution — two age peaks observed.
- Non-survivors: Left-skewed — majority diagnosed at older ages, suggesting younger patients are less affected by cancer mortality.

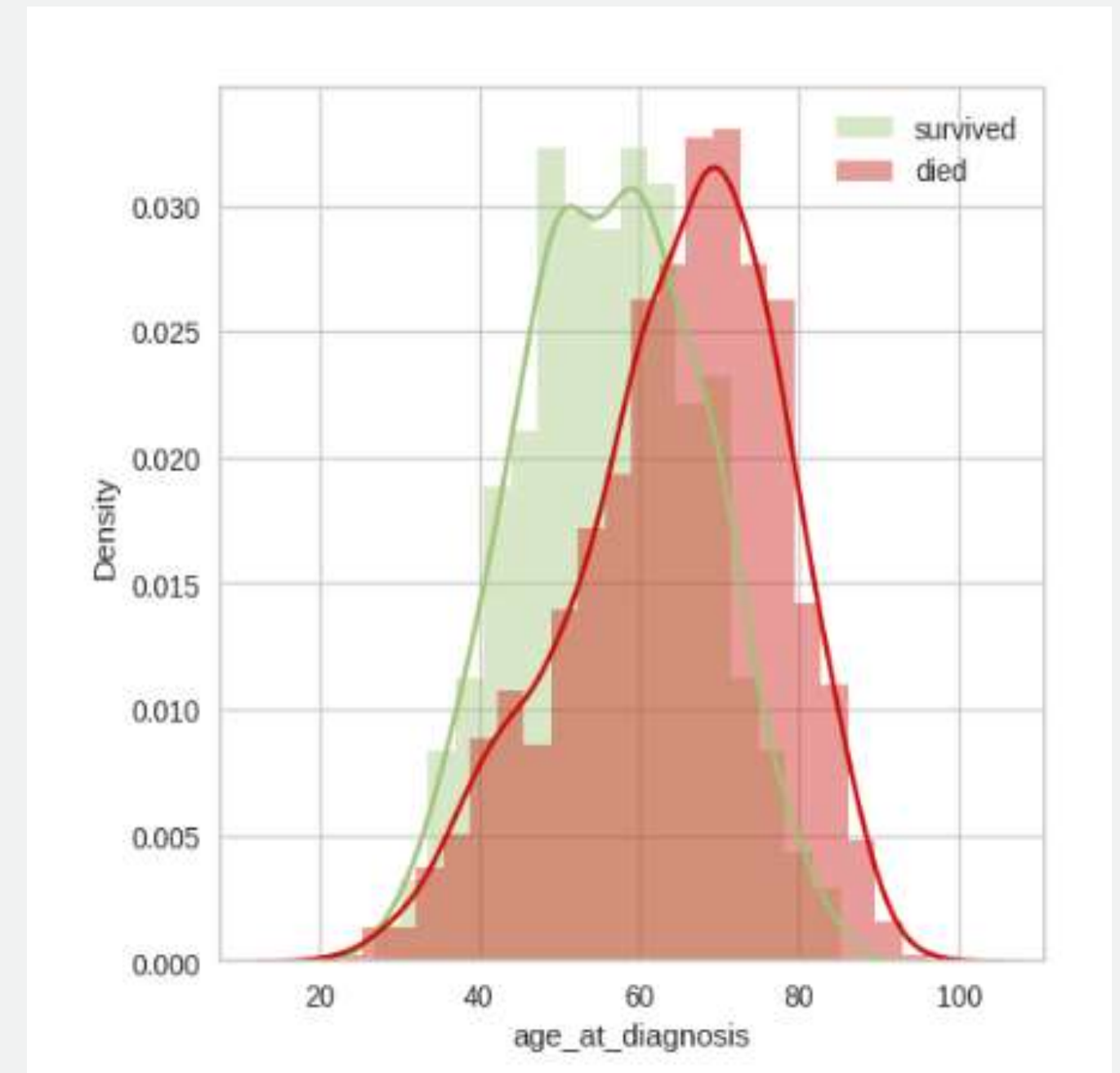


Figure 5: EDA performed on “age at diagnosis” attribute

EXPLORATORY DATA ANALYSIS (Cont...)

Lymph nodes examined positive:

- Both groups: Right-skewed distribution.
- Indicates majority with a lower number of positive lymph nodes; few patients with high counts, leading to skewness.

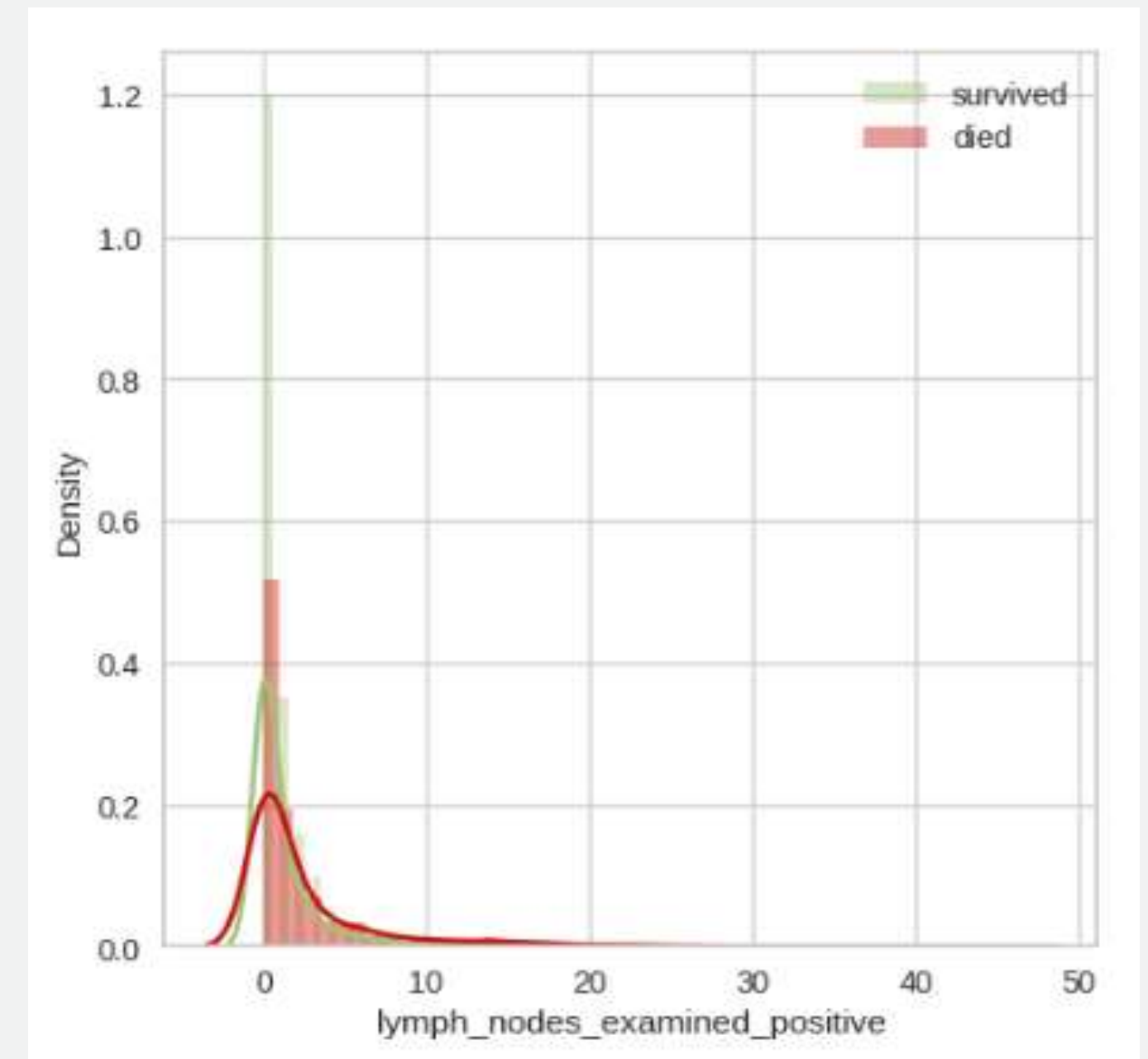


Figure 6: EDA performed on “lymph nodes examined positive” attribute

EXPLORATORY DATA ANALYSIS (Cont...)

Mutation count:

- Both groups: Right-skewed distribution.
- Most patients have fewer mutations; outliers with high mutation counts.

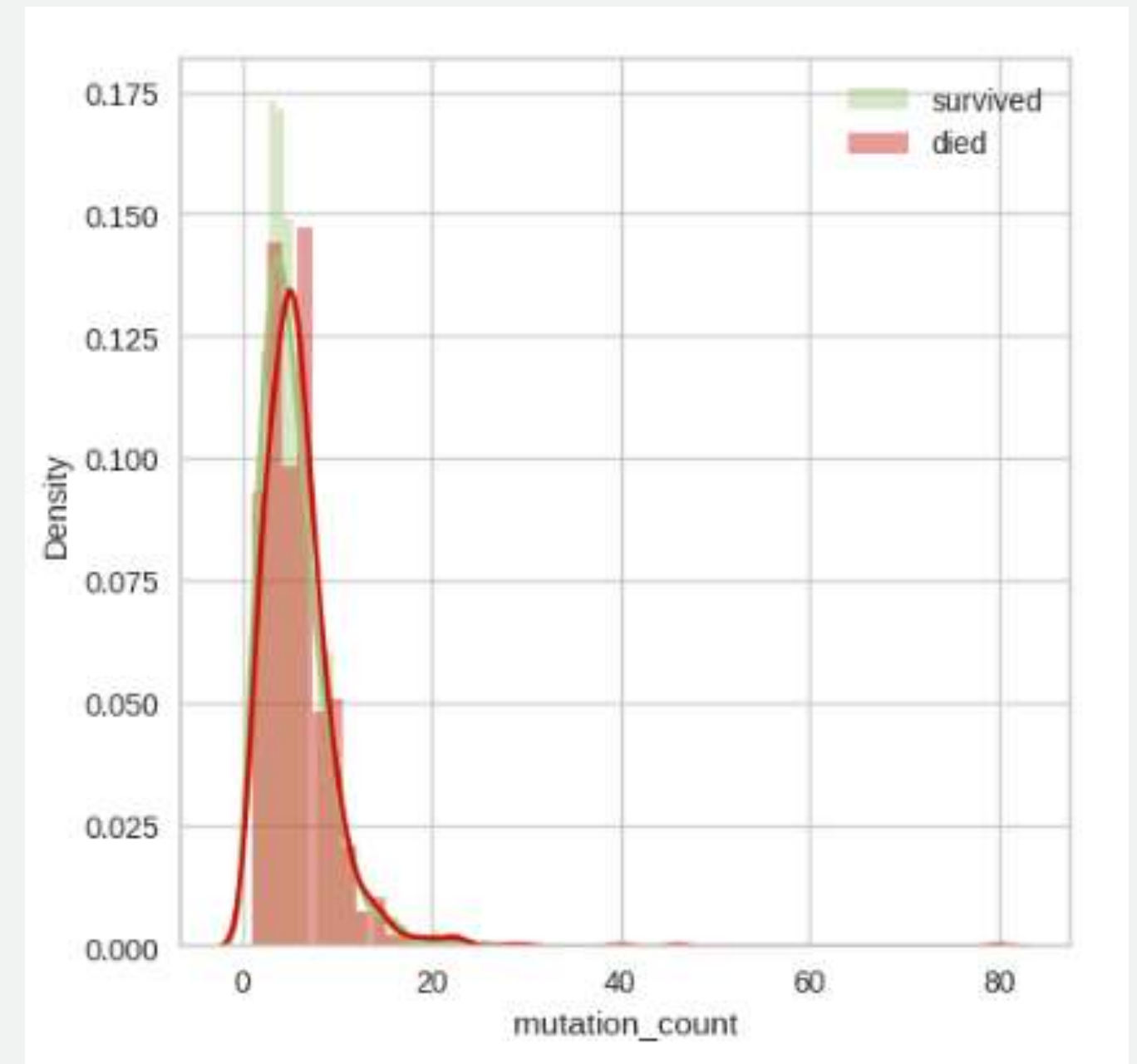


Figure 7: EDA performed on “mutation count” attribute

EXPLORATORY DATA ANALYSIS (Cont...)

Nottingham Prognostic Index (NPI):

- Both groups: Multimodal distribution — multiple peaks/clusters.
- Suggests different risk categories in NPI, aiding clinicians in decision-making.

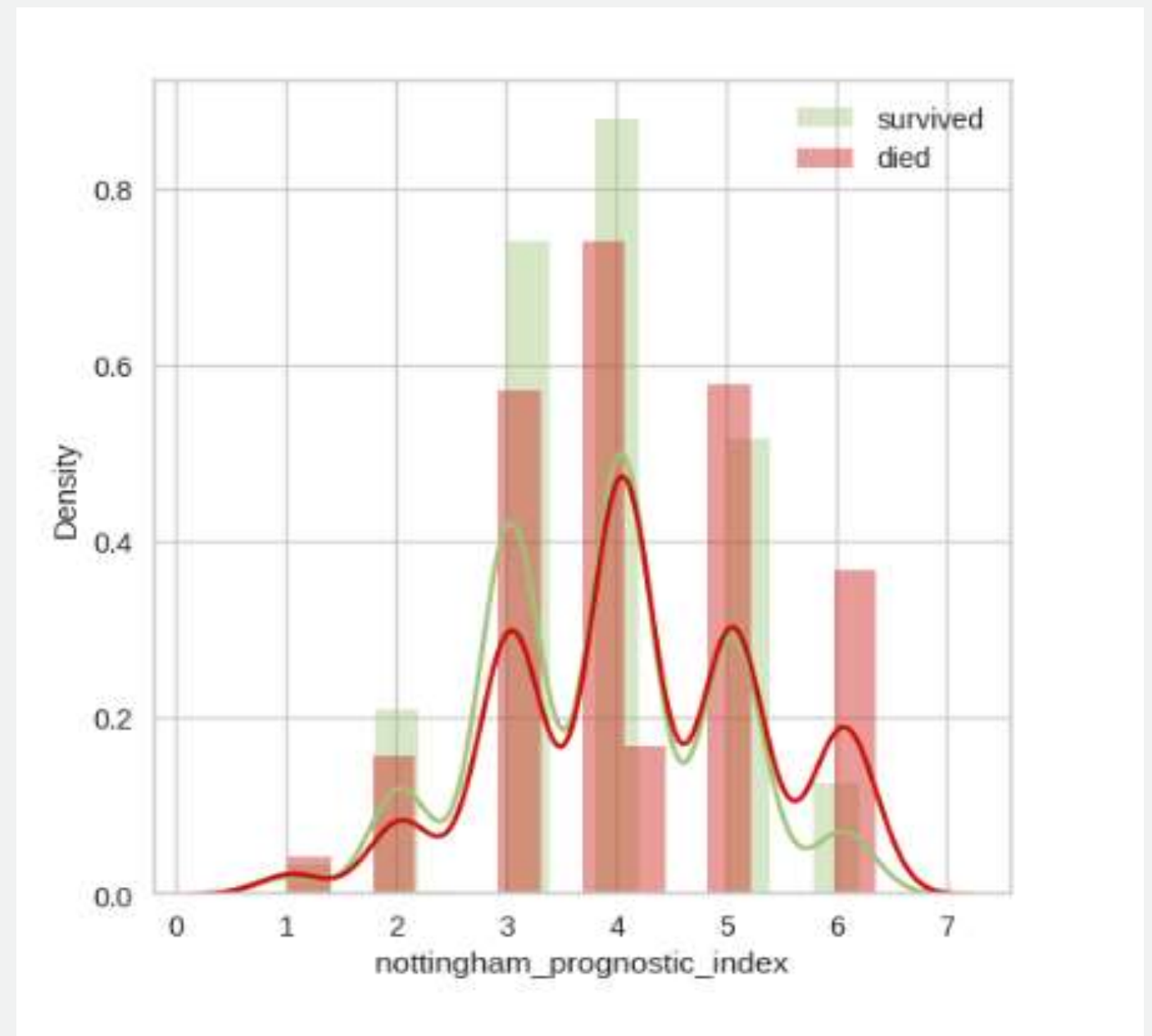


Figure 8: EDA performed on “Nottingham prognostic index” attribute

EXPLORATORY DATA ANALYSIS (Cont...)

Overall survival months:

- Survivors: Bimodal distribution — two distinct survival duration peaks.
- Non-survivors: Right-skewed — majority had shorter survival times.

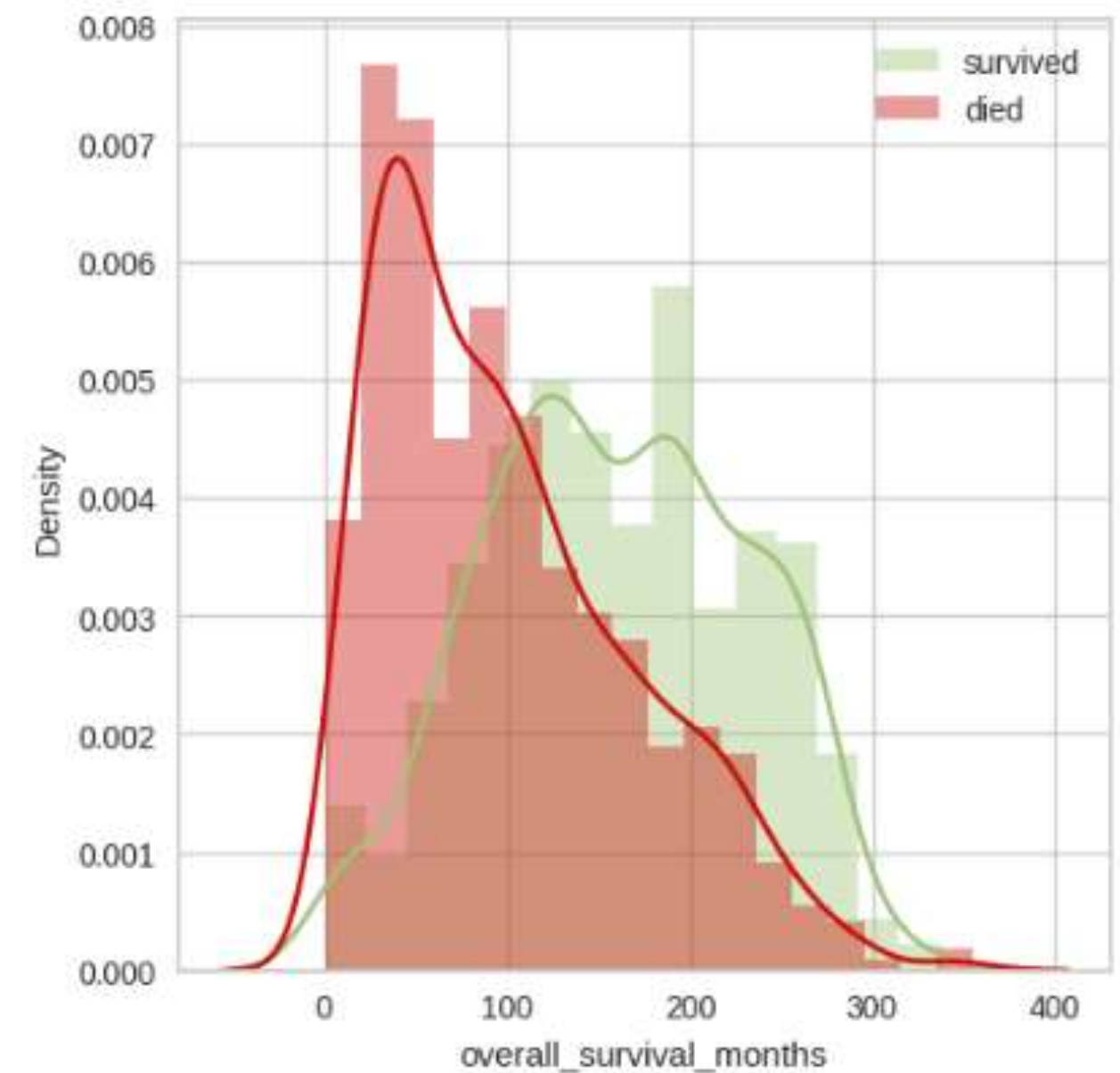
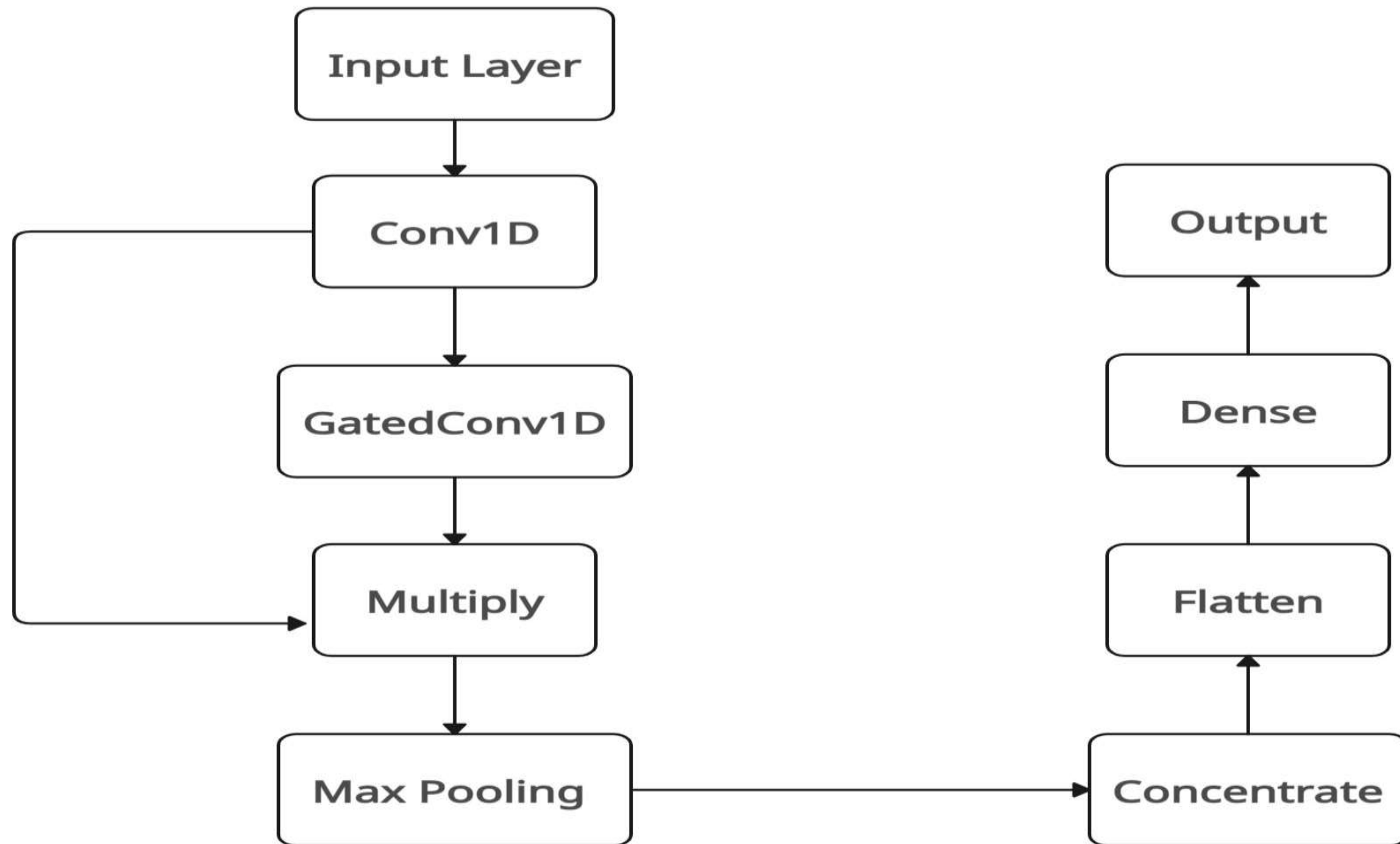


Figure 9: EDA performed on “overall survival month” attribute



Architecture of Gated Attention CNN Model

PARAMETER DETAILS of PROPOSED MODEL

Table 3: Specification of the Convolution Layer

Number of layers	2
Dimensions of the Filter	2, 3
Total filters used	30
Step size	2
Padding	Identical
Function for Activation	Rectified Linear Unit (ReLU)



PARAMETER DETAILS of PROPOSED MODEL

Table 4: Specification of the Gated Attention Layer

Number of layers	2
Dimensions of the Filter	1, 3
Total filters used	30
Step size	2
Padding	Identical
Function for Activation	Sigmoid



PARAMETER DETAILS of PROPOSED MODEL

Table 5: Specification of the Max Pooling Layer

Pool size	2
Stride	1
Padding	Identical



PARAMETER DETAILS of PROPOSED MODEL

Table 6: Specification of the Fully Connected Layer

Total count of hidden layers	3
No. of hidden neurons in hidden layers	200, 150, 100
Dropout	50%
Function for Activation	Hyperbolic Tangent

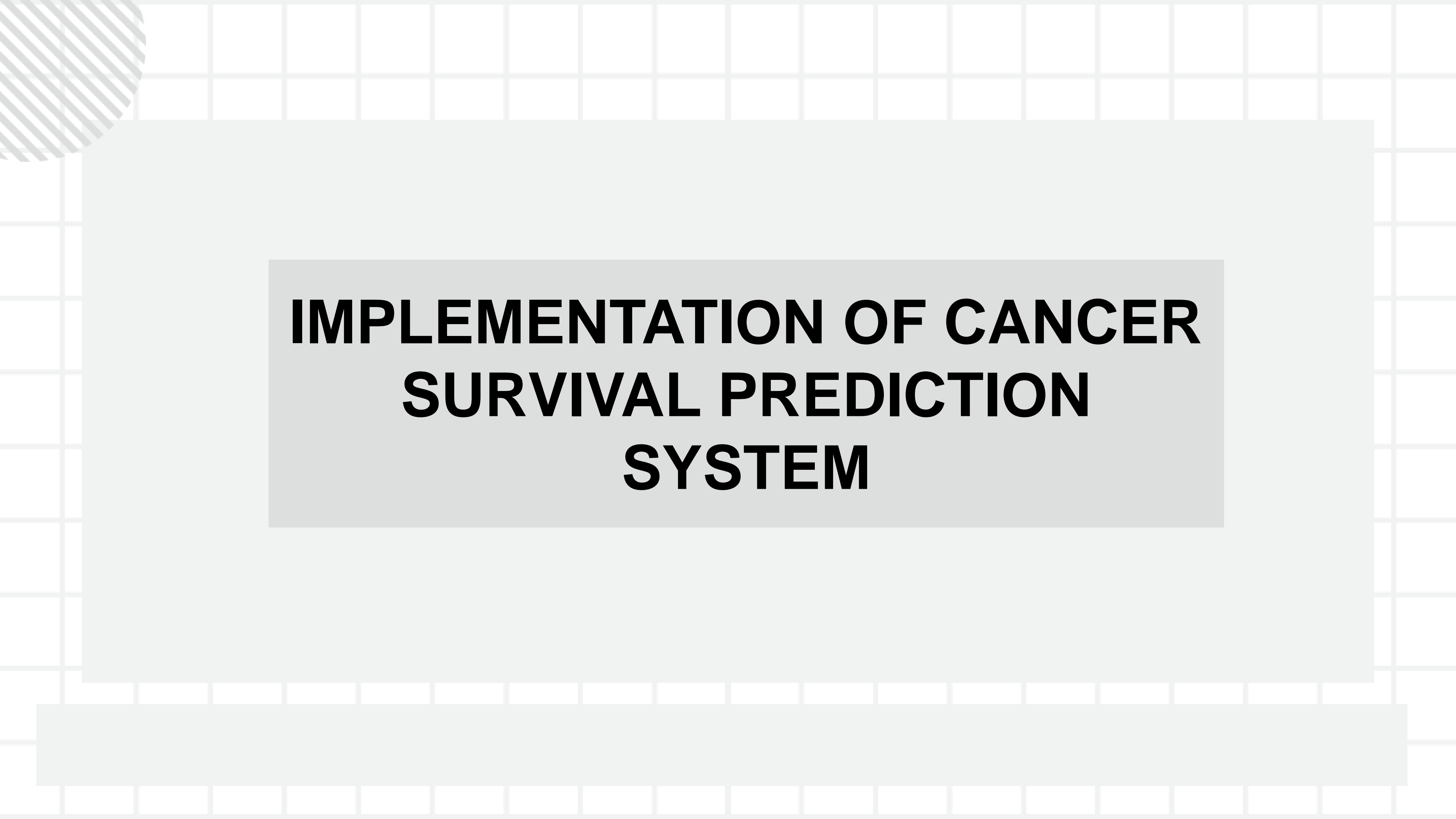


PARAMETER DETAILS of PROPOSED MODEL

Table 7: Parameter Details of Other Parameters

Activation function in the output layers	Sigmoid activation
Batch size during training	8
Number of Training cycles	50
Function to calculate model loss	Binary cross-entropy + L2 regularization





IMPLEMENTATION OF CANCER SURVIVAL PREDICTION SYSTEM

DATASET DESCRIPTION

1. Source & METABRIC Data Used:

- Publicly accessible from cbioportal.org [22].
- Named: Molecular Taxonomy of Breast Cancer International Consortium (METABRIC).

2. Patient Data:

- Comprises 1,980 valid breast cancer patient data.
- Contains multi-dimensional data including gene expression profile, copy number alteration profile, and clinical information.
- 491 patients are short-term survivors (<5 years) labeled as 0.
- 1,489 patients are long-term survivors (>5 years) labeled as 1.

SUMMARY -- DATASET DESCRIPTION

Table 4: Detailed description about METABRIC [22] dataset

Disease	Breast
Number of Patients	1980
Survival threshold (years)	5
Survival > (5 years) Class : 1	1489
Survival < (5 years) Class : 0	491
Number of modalities	3
Modalities	Clinical, Gene expression and CNA

SUMMARY -- DATASET DESCRIPTION (Cont...)

Table 5: Selected Features for the Proposed Model

Data Category	Selected feature numbers for the prediction model as used in the base paper
Clinical	25
Gene Expression	400
Copy Number Alteration	200

#checking the data
df2

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	chemotherapy	pam50+_claudin-low_subtype	cohort	er_status_measured_by_ihc	...	nottingham_prognostic_index	oncotree_code	overall_sur
0	0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	0	claudin-low	1	Positive	...	6.044	IDC	
1	2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumA	1	Positive	...	4.020	IDC	
2	5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1	LumB	1	Positive	...	4.030	IDC	
3	6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	1	LumB	1	Positive	...	4.050	MDLC	
4	8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	1	LumB	1	Positive	...	6.080	MDLC	
...	
1899	7295	43.10	BREAST CONSERVING	Breast Cancer	Breast Invasive Lobular Carcinoma	High	0	LumA	4	Positive	...	5.050	ILC	
1900	7296	42.88	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumB	4	Positive	...	5.040	IDC	
1901	7297	62.90	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumB	4	Positive	...	6.050	IDC	
1902	7298	61.16	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Moderate	0	LumB	4	Positive	...	5.050	IDC	
1903	7299	60.02	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumB	4	Positive	...	5.040	IDC	

Figure 10: Various parameters of the clinical data.

34s



```
from google.colab import files
import io
data = files.upload()
```



Choose Files METABRIC...ression.csv

- **METABRIC_gene-expression.csv**(text/csv) - 1609865 bytes, last modified: 6/2/2023 - 100% done

Saving METABRIC_gene-expression.csv to METABRIC_gene-expression (1).csv

1s



```
data = pd.read_csv(io.StringIO(data['METABRIC_gene-expression (1).csv'].decode('utf-8')),low_memory=False)
data
```

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_10	...	g_393	g_394	g_395	g_396	g_397	g_398	g_399	g_400	Patient_id	label
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	Pid_1	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_2	0
2	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	Pid_3	0
3	0	0	1	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	Pid_4	1
4	0	0	0	0	0	0	0	-1	0	0	...	0	0	0	0	0	0	1	0	Pid_5	0
...
1975	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1976	1
1976	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1977	1
1977	0	0	0	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	Pid_1978	1
1978	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	Pid_1979	0
1979	0	0	0	0	0	0	0	0	0	0	...	0	0	-1	0	0	0	0	0	Pid_1980	1

1980 rows x 402 columns

Figure 11: The figure showcases the transformed gene expression data. Columns represent different genes, ranging from g_1 to g_{400} , while rows correspond to patients, from pid_1 to pid_{1980} . Each table entry indicates the expression level of a particular gene for a specific patient. The values within the table emphasize the strength of the association between individual patients and genes.

```
[10] from google.colab import files
import io
data_1 = files.upload()
```

Choose Files METABRIC_cna.csv

- **METABRIC_cna.csv**(text/csv) - 828505 bytes, last modified: 6/2/2023 - 100% done
Saving METABRIC_cna.csv to METABRIC_cna.csv

```
data_1 = pd.read_csv(io.StringIO(data_1['METABRIC_cna.csv'].decode('utf-8')),low_memory=False)
data_1
```

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_10	...	g_193	g_194	g_195	g_196	g_197	g_198	g_199	g_200	Patient_id	label
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1	0
1	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	1	0	0	0	Pid_2	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	-1	0	0	Pid_3	0
3	0	0	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0	Pid_4	1
4	0	-1	0	0	-1	0	-1	0	0	0	...	0	-1	0	0	0	-1	0	-1	Pid_5	0
...
1975	0	2	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1976	1
1976	0	0	0	0	0	-1	0	0	0	0	...	0	0	0	0	0	-1	0	0	Pid_1977	1
1977	1	0	0	0	0	0	2	0	0	0	...	0	0	1	0	0	0	0	0	Pid_1978	1
1978	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1979	0
1979	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1980	1

1980 rows × 202 columns

Figure 12: The figure showcases the transformed copy number alteration (CNA) profile. Columns represent different genes, ranging from g_1 to g_{200} , while rows correspond to patients, from pid_1 to pid_{1980} . Each table entry indicates the expression level of a particular CNA for a specific patient. The values within the table emphasize the strength of the association between individual patients and CNA.

SYSTEM IMPLEMENTATION

1. Programming Language and Version Used:

- **Language: Python**
 - Python is a versatile and widely-used programming language, known for its ease of use and a vast library that makes it applicable in various domains like data analysis, artificial intelligence, scientific computing, etc.
- **Version: 3.11.3**
 - Version 3.11.3 is one of the latest versions of Python, which includes new features and optimizations, ensuring efficient and effective coding with reduced bugs and improved functionality.

SYSTEM IMPLEMENTATION

```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC Clinical dataset
dataset_clinical = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/M

# split into input (X) and output (Y) variables
X_clinical = dataset_clinical[:,0:25]
Y_clinical = dataset_clinical[:,25]
```

"""

1. Fix Random Seed:

This line sets the random seed to 1 for reproducibility of random processes using NumPy.
| Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC Clinical Dataset:

This line loads the METABRIC Clinical dataset from the specified file path.
| The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 26 columns.
| The first 25 columns represent the input features (X_clinical), and the last column represents the output labels (Y_clinical).

3. Split Input (X) and Output (Y) Variables:

This code splits the loaded dataset into input (X_clinical) and output (Y_clinical) variables.
| `X_clinical` contains all the rows of the dataset and the first 25 columns, representing the features or independent variables.
| `Y_clinical` contains all the rows of the dataset and the last column, representing the labels or dependent variable.

"""

Figure 13: Code for Initialization and Preprocessing of the METABRIC Clinical Dataset

SYSTEM IMPLEMENTATION (Cont...)

```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC EXP dataset
dataset_exp = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/METABR

# split into input (X) and output (Y) variables
X_exp = dataset_exp[:,0:400]
Y_exp = dataset_exp[:,400]
```

"""

1. Fix Random Seed:

This line sets the random seed to 1 for reproducibility of random processes using NumPy.
| Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC CNV Dataset:

This line loads the METABRIC CNV dataset from the specified file path.
| The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 201 columns.
| The first 400 columns represent the input features (X_exp), and the last column represents the output labels (Y_exp).

3. Split Input (X) and Output (Y) Variables:

This code splits the loaded dataset into input (X_exp) and output (Y_exp) variables.
| `X_exp` contains all the rows of the dataset and the first 400 columns, representing the features or independent variables.
| `Y_exp` contains all the rows of the dataset and the last column, representing the labels or dependent variable.

"""

Figure 14: Code for Initialization and Preprocessing of the METABRIC Gene Expression Dataset

SYSTEM IMPLEMENTATION (Cont...)

```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC CNV dataset
dataset_cnv = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/METABR

# split into input (X) and output (Y) variables
X_cnv = dataset_cnv[:,0:200]
Y_cnv = dataset_cnv[:,200]
```

"""

1. Fix Random Seed:

This line sets the random seed to 1 for reproducibility of random processes using NumPy.
| Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC CNV Dataset:

This line loads the METABRIC CNV dataset from the specified file path.
| The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 201 columns.
| The first 200 columns represent the input features (X_cnv), and the last column represents the output labels (Y_cnv).

3. Split Input (X) and Output (Y) Variables:

This code splits the loaded dataset into input (X_cnv) and output (Y_cnv) variables.
| `X_cnv` contains all the rows of the dataset and the first 200 columns, representing the features or independent variables.
| `Y_cnv` contains all the rows of the dataset and the last column, representing the labels or dependent variable.

"""

Figure 15: Code for Initialization and Preprocessing of the METABRIC Copy Number Alteration Dataset

SYSTEM IMPLEMENTATION (Cont...)

```
conv_clinical1 = Conv1D(filters=num_filters,kernel_size=1,strides=2,padding='same',name='Conv1D_clinical1',kernel_initializer='glorot_uniform')
#activ = nlrelu(conv_clinical1,'nrelu')
gatedAtnConv_clinical1 = Conv1D(filters=num_filters,kernel_size=1,strides=1,padding='same',name='GatedConv1D1',activation='relu',kernel_initializer='glorot_uniform')
gatedAtnConv_clinical1_1 = Conv1D(filters=num_filters,kernel_size=3,strides=1,padding='same',name='GatedConv1D1_1',activation='relu',kernel_initializer='glorot_uniform')
mult_1_1 = multiply([gatedAtnConv_clinical1,conv_clinical1])
mult_1_1_1 = multiply([gatedAtnConv_clinical1_1,conv_clinical1])
pooled_clinical1 = MaxPooling1D(pool_size=2, strides=1, padding='same')(mult_1_1)
pooled_clinical1_1 = MaxPooling1D(pool_size=2, strides=1, padding='same')(mult_1_1_1)
```

■■■■ ■■■■ ■■■■

```
7. conv_clinical1 = Conv1D(filters=num_of_filters, kernel_size=1, strides=2, padding='same', name='Conv1D_clinical1', kernel_initializer='glorot_un
```

Here are the details:

```
# `Conv1D`: This function creates a 1D convolutional layer.
```

`filters=num_of_filters`: This specifies the number of filters (or output channels) in the convolutional layer.

```

`num_of_filters` is a variable that you defined earlier in the code and seems to be set to 25.

```

```
# `kernel_size=1`: This sets the size of the convolutional kernel to 1.
```

[illegible]

```
# `strides=2`: This sets the stride of the convolutional layer to 2.
```

The stride determines the step size at which the kernel slides over the input.

In this case, the kernel moves two steps at a time, leading to downsampling.

```
# ~padding='same': This sets the padding mode to 'same', meaning the input is padded with zeros so that the output size matches the input size.
```

```
# `name='Conv1D_clinical1`: This assigns a name to the layer for identification.
```

```
# `kernel_initializer='glorot_uniform': This sets the weight initialization method for the convolutional layer.
```

`'glorot_uniform'` is an initializer that draws weights from a uniform distribution based on the Glorot uniform

```
# `bias_initializer=bias_init`: This sets the bias initializer for the convolutional layer.
```

The constant bias value of 0.1 is used for initialization, as defined earlier.

Figure 16: Model Definition i.e. Gated Attention CNN model on Clinical Data

SYSTEM IMPLEMENTATION (Cont...)

```
roc_auc = auc(fpr, tpr)
plt.plot(fpr, tpr, 'r', label = 'Gated_Attention_CNN-Clinical = %0.3f' %roc_auc)
plt.xlabel('1-Sp (False Positive Rate)')
plt.ylabel('Sn (True Positive Rate)')
plt.title('Receiver Operating Characteristics')
plt.legend()
plt.show()
```

"""

The code you provided is used to create and display a Receiver Operating Characteristic (ROC) curve for evaluating the performance of a classification model.

`roc_auc = auc(fpr, tpr)` : This line calculates the Area Under the Curve (AUC) for the ROC curve.

| The `auc` function from the `sklearn.metrics` module is used to compute the AUC, which quantifies the overall performance of the model across different thresholds.

`plt.plot(fpr, tpr, 'r', label='SiGaAtCNN-CLN = %0.3f' % roc_auc)` : This line plots the ROC curve using the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. The `r` argument specifies that the line should be red.

The `label` parameter provides a label for the plot legend, including the calculated AUC value formatted with three decimal places.

`plt.xlabel('1-Sp (False Positive Rate)')` : This line sets the label for the x-axis to "1-Sp (False Positive Rate)", indicating the false positive rate (1 - Specificity).

`plt.ylabel('Sn (True Positive Rate)')` : This line sets the label for the y-axis to "Sn (True Positive Rate)", indicating the true positive rate (Sensitivity).

`plt.title('Receiver Operating Characteristics')` : This line sets the title of the plot to "Receiver Operating Characteristics", describing the type of plot.

`plt.legend()` : This line adds a legend to the plot, displaying the label provided in the plot function.

`plt.show()` : This line displays the plot.

"""

.

Figure 17: Area Under Curve Evaluation Metrics using for checking the model performance on clinical data



EXPERIMENTS AND RESULTS

EXPERIMENTAL SETUP

1. Platform and Specification:

- **Platform: Google Colab**

- Google Colab is a free, cloud-based version of Jupyter Notebooks. It's a platform that allows you to write and execute Python in your browser with zero configuration, free access to GPUs, and easy sharing of your work.

- **CPU: Intel Xeon CPU with 2 vCPUs and 13GB RAM**

- The available CPU and RAM in Google Colab allow for robust data processing and model training, especially beneficial for machine learning and data analysis tasks.

EXPERIMENTAL SETUP (Cont...)

2. Hardware Details:

- **Processor: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz**
 - The processor is crucial for the overall speed and efficiency of the computer. The Intel i5-8265U provides reliable performance and can handle multiple tasks without significant slowdown.
- **RAM: 12.0 GB (11.9 GB usable)**
 - Adequate RAM ensures that your machine can handle running multiple applications simultaneously without hampering performance.

EXPERIMENTAL SETUP (Cont...)

3. Software Details:

- **TensorFlow Version: 2.8.0**
 - TensorFlow is an open-source library developed by Google for numerical computation and machine learning. Version 2.8.0 includes various features and optimizations that enhance model training and deployment.

4. Operating System Details:

- **Edition: Windows 10 Home Single Language**
 - A version of Windows 10 designed for home use with a simple and user-friendly interface, ensuring ease of use and a wide range of consumer features.

EXPERIMENTAL SETUP (Cont...)

4. Operating System Details (Cont...):

- **Version: 22H2**
 - The version number indicates the specific update or feature set included in the OS.
- **OS build: 19045.3448**
 - The build number provides information about the specific iteration of the OS, often used for troubleshooting and support.

EVALUATION METRICS USED IN PREDICTION

1. **Accuracy:** Defined as the proportion of correctly classified samples to the total number of samples.

$$Acc = \frac{TP + TN}{TP + TN + FN + FP}$$

2. **Precision:** Defined as the proportion of samples that were actually positive, which was classified as positive.

$$Pre = \frac{TP}{TP + FP}$$

3. **Sensitivity (Sn):** Represents the proportion of all positive samples that are correctly classified, and measured the classifier's ability to recognize positive samples.

$$Sn = \frac{TP}{TP + FN}$$

4. **Receiver Operating Characteristics Curve (ROC):** An ROC curve is a graph showing the performance of a classification model at all classification thresholds.

RESULTS

Table 6: Comparative Analysis of Proposed Model Performance: Unimodal vs. Multimodal Data Using Accuracy as a Metric

Model	Accuracy
Unimodal Gated Attention CNN – Clinical	0.813
Unimodal Gated Attention CNN – CNA	0.893
Unimodal Gated Attention CNN – Gene expression	0.841
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.912

RESULTS (Cont...)

Table 7: Comparative Analysis of Proposed Model Performance: Unimodal vs. Multimodal Data Using Precision as a Metric

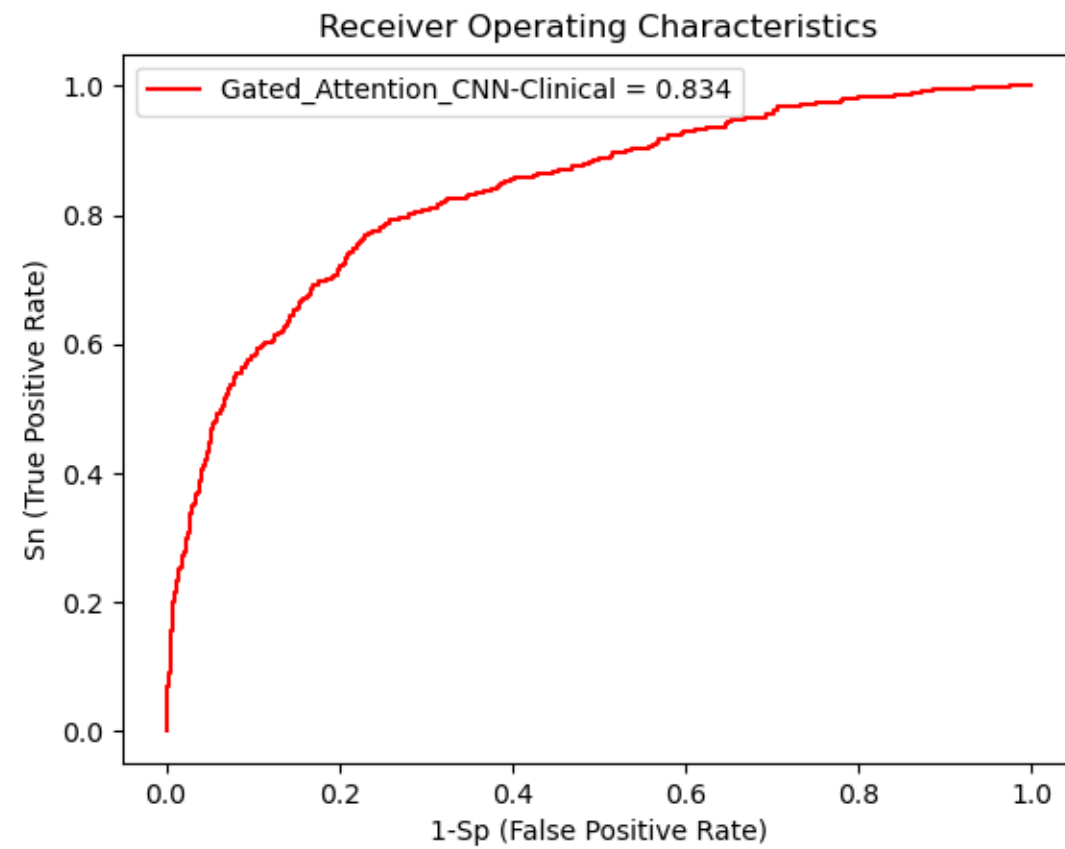
Model	Precision
Unimodal Gated Attention CNN – Clinical	0.712
Unimodal Gated Attention CNN – CNA	0.841
Unimodal Gated Attention CNN – Gene expression	0.779
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.841

RESULTS (Cont...)

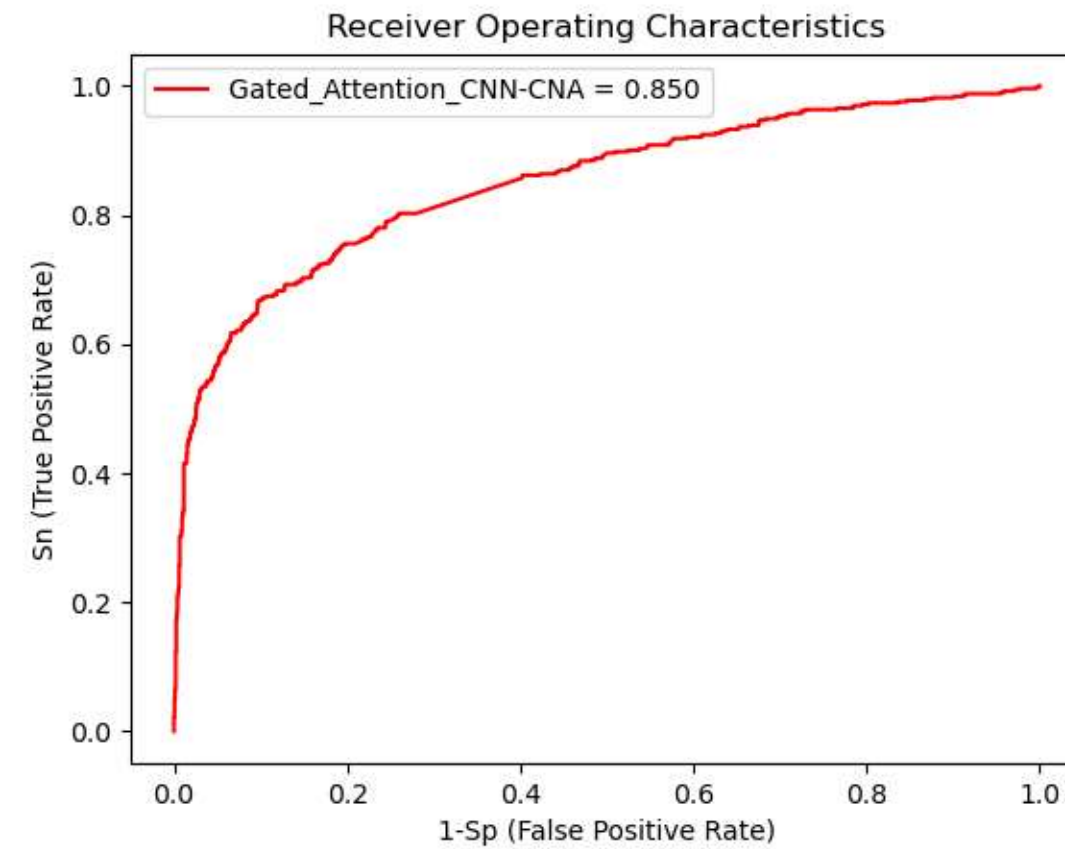
Table 8: Comparative Analysis of Proposed Model Performance: Unimodal vs. Multimodal Data Using Sensitivity as a Metric

Model	Sensitivity
Unimodal Gated Attention CNN – Clinical	0.413
Unimodal Gated Attention CNN – CNA	0.702
Unimodal Gated Attention CNN – Gene expression	0.505
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.798

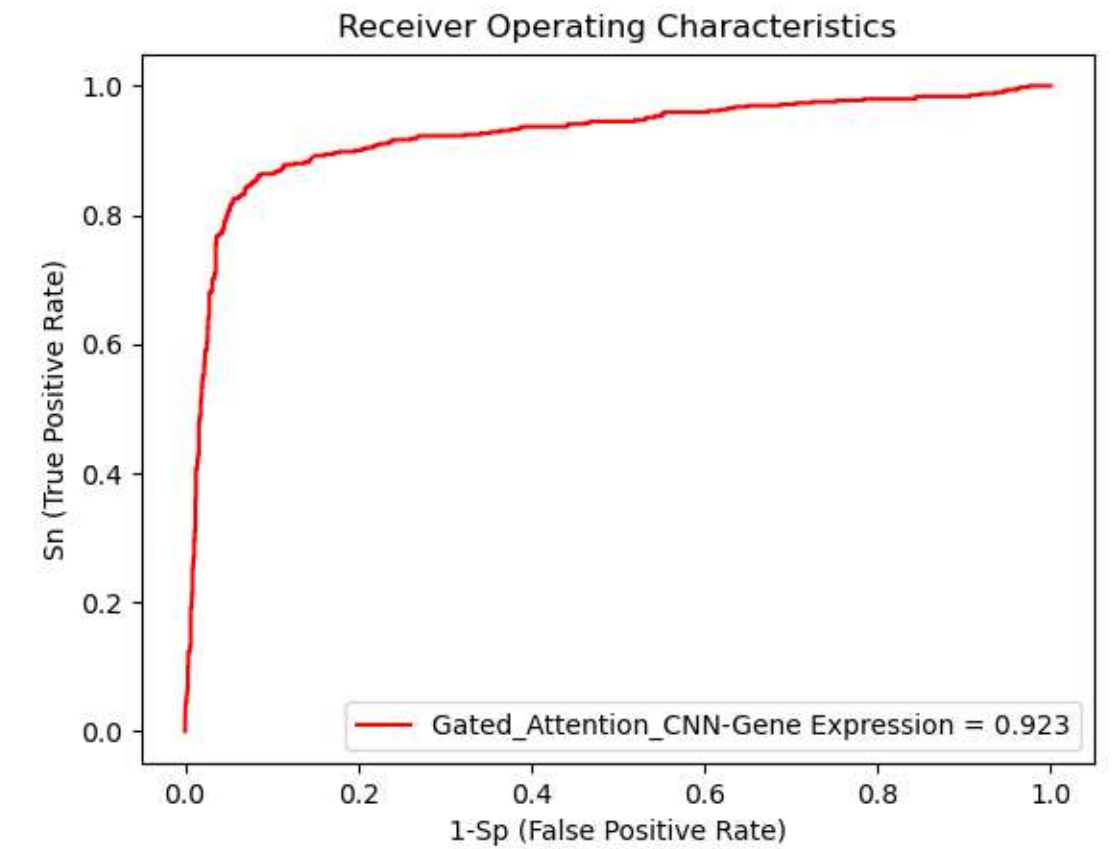
RESULTS (Cont...)



a) Gated Attention CNN – Clinical ROC curve



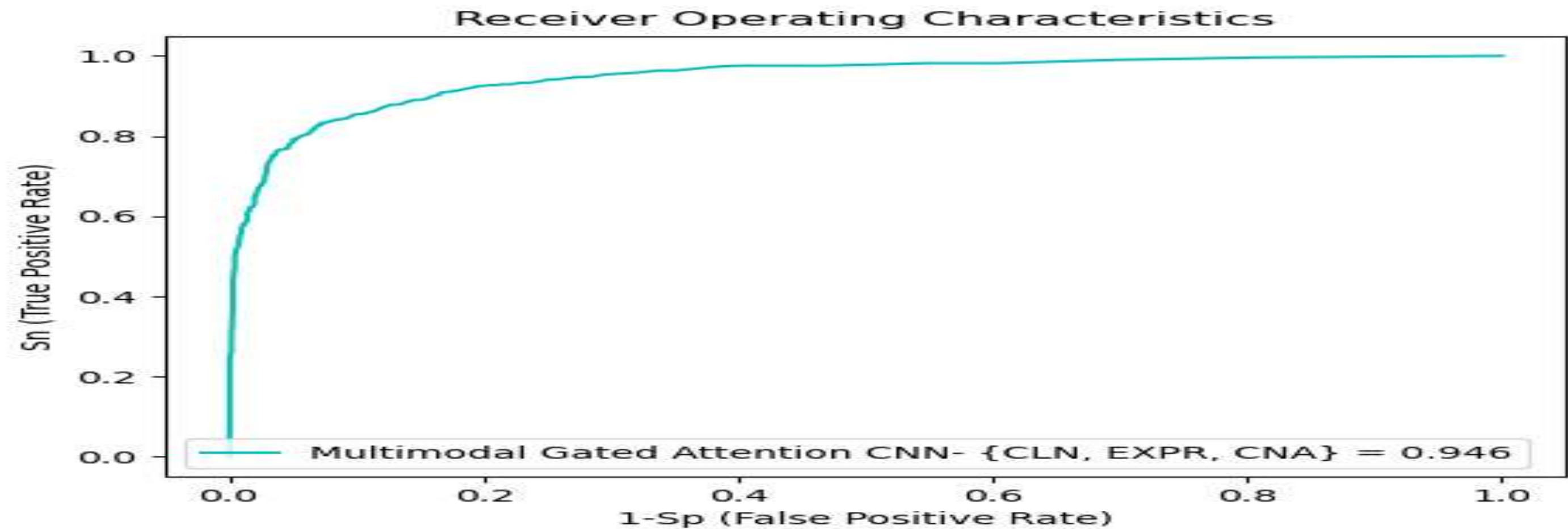
b) Gated Attention CNN – CNA ROC curve



c) Gated Attention CNN – Gene Expression ROC curve

Figure 18: ROC curve of Gated Attention CNN model trained on each data modality

RESULTS (Cont...)



d) Gated Attention CNN – {Clinical, Gene expression & CAN} ROC curve

Figure 19: ROC curve of Gated Attention CNN model trained on combination of each modality

RESULTS (Cont...)

Table 9: Comparative Analysis of Proposed Model Performance: Unimodal vs. Multimodal Data Using AUC as a Metric

Model	Area under curve
Unimodal Gated Attention CNN – Clinical	0.834
Unimodal Gated Attention CNN – CNA	0.850
Unimodal Gated Attention CNN – Gene expression	0.923
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.946

RESULTS (Cont...)

Table 10: Comparison of the proposed multimodal prediction model with other state of the art multimodal prediction model

Model	Accuracy
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.912
MDNNMD [18]	0.826
Stacked RF [12]	0.902
Logistic Regression [19]	0.760
Random Forest [20]	0.791
Support Vector Machine [21]	0.805

RESULTS (Cont...)

Table 10: Comparison of the proposed multimodal prediction model with other state of the art multimodal prediction model (Cont.)

Model	Precision
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.841
MDNNMD [18]	0.749
Stacked RF [12]	0.841
Logistic Regression [19]	0.549
Random Forest [20]	0.766
Support Vector Machine [21]	0.708

RESULTS (Cont...)

Table 10: Comparison of the proposed multimodal prediction model with other state of the art multimodal prediction model (Cont.)

Model	Sensitivity
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.798
MDNNMD [18]	0.450
Stacked RF [12]	0.747
Logistic Regression [19]	0.183
Random Forest [20]	0.226
Support Vector Machine [21]	0.365

RESULTS (Cont...)

Table 10: Comparison of the proposed multimodal prediction model with other state of the art multimodal prediction model (Cont.)

Model	Area under curve
Multimodal Gated Attention CNN – {Clinical, CNA, Gene expression}	0.950
MDNNMD [18]	0.845
Stacked RF [12]	0.930
Logistic Regression [19]	0.663
Random Forest [20]	0.801
Support Vector Machine [21]	0.810

CONCLUSION

1. **Multimodal Data:** Enhances predictive models; traditional ML often misses intricate inter-modal relationships.
2. **Gated Attention CNN:**
 - Custom architecture excels in integrating features across modalities.
 - Highlights the strength of combined data insights.
3. **METABRIC Dataset Analysis:**
 - Reveals insightful patterns, like bimodal age distribution for survivors.
 - Emphasizes the depth and richness of multimodal breast cancer data.



CONCLUSION (Cont...)

4. Performance Metrics:

- Highest accuracy (91.2%) and ROC-curve (0.950) achieved with all three modalities combined.
- Confirms superior predictive power of combined datasets.

5. **Overall:** Combining multimodal data with advanced deep learning offers a potential revolution in predictive oncology analytics.



FUTURE WORK

1. Enhancing Model Capabilities:

- Exploration of other deep learning architectures to increase prediction accuracy.
- Utilize transfer learning to leverage pre-trained models and expedite the training process.
- Investigate techniques to counter overfitting especially when integrating multiple data modalities.
- Collaborate with patient advocacy groups to understand patient needs better and refine prediction criteria.



FUTURE WORK (Cont...)

2. Expanding Dataset Diversity:

- Collaborate with additional hospitals and institutions to gather a larger and more diverse set of patient data.
- Integration of other relevant medical data types, such as MRI or PET scans, into the model.
- Consider other forms of cancer to test the model's adaptability and relevance across diseases.



FUTURE WORK (Cont...)

3. Model Interpretability & Explainability:

- Develop techniques to visualize how the model makes decisions, aiding clinicians in understanding its predictions.
- Implement feature importance algorithms to identify the most critical factors in survival prediction.
- Explore methods to counteract biases in predictions, ensuring equitable treatment recommendations across patient demographics.



FUTURE WORK (Cont...)

4. Concluding Thoughts:

- The future is promising with endless opportunities for refining and deploying our cancer survival prediction model.
- Continuous feedback, research, and collaboration will be key to our model's success and adaptability.
- We're committed to pushing the boundaries of what's possible in this vital field.



REFERENCES

1. Bozkurt, Caner & Aşuroğlu, Tunç. (2023). Mortality Prediction of Various Cancer Patients via Relevant Feature Analysis and Machine Learning. SN Computer Science. 4. [10.1007/s42979-023-01720-5](https://doi.org/10.1007/s42979-023-01720-5)
2. Arya, Nikhilanand & Saha, Sriparna & Mathur, Archana & Saha, Snehanstu. (2023). Improving the robustness and stability of a machine learning model for breast cancer prognosis through the use of multi-modal classifiers. Scientific Reports. 13. [10.1038/s41598-023-30143-8](https://doi.org/10.1038/s41598-023-30143-8)
3. Zolfaghari, Behrouz & Mirsadeghi, Leila & Bibak, Khodakhast & Kavousi, Kaveh. (2023). Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning. ACM Computing Surveys. 55. [10.1145/3580218](https://doi.org/10.1145/3580218)
4. Sorayaie Azar, Amir & Babaei Rikan, Samin & Naemi, Amin & Bagherzadeh, J. & Pirnejad, Habibollah & Mohasefi, Matin & Wiil, Uffe. (2022). Application of machine learning techniques for predicting survival in ovarian cancer. BMC Medical Informatics and Decision Making. 22. [10.1186/s12911-022-02087-y](https://doi.org/10.1186/s12911-022-02087-y)
5. Yan, F., Feng, Y. (2022). A two-stage stacked-based heterogeneous ensemble learning for cancer survival prediction. *Complex Intell. Syst.* 8, 4619–4639. <https://doi.org/10.1007/s40747-022-00791-w>

REFERENCES

6. S P, S., I, S., A H, K., R, H., S, K., H, G., & A, S. Z. (2020). Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework. *Journal of biomedical physics & engineering*, 10(4), 479–492. <https://doi.org/10.31661/JBPE.V0I0.1027>
7. Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *Proceedings of the ... IEEE International Symposium on Signal Processing and Information Technology*. *IEEE International Symposium on Signal Processing and Information Technology*, 2018, 632–637. <https://doi.org/10.1109/ISSPIT.2018.8642753>
8. Lynch, Chip & Abdollahi, Behnaz & Fuqua, Joshua & deCarlo, Alexandra & Bartholomai, James & Balgemann, Rayeanne & Berkel, Victor & Frieboes, Hermann. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*. 108. [10.1016/j.ijmedinf.2017.09.013](https://doi.org/10.1016/j.ijmedinf.2017.09.013)



REFERENCES (Cont...)

9. Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 24(1), 31–42. <https://doi.org/10.3233/THC-151071>
10. Wu, X., Shi, Y., Wang, M., & Li, A. (2023). CAMR: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics (Oxford, England)*, 39(1), btad025. <https://doi.org/10.1093/bioinformatics/btad025>
11. Fan, Z., Jiang, Z., Liang, H., & Han, C. (2023). Pancancer survival prediction using a deep learning architecture with multimodal representation and integration. *Bioinformatics advances*, 3(1), vbad006. <https://doi.org/10.1093/bioadv/vbad006>
12. Arya, N., & Saha, S. (2022). Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2), 1032–1041. <https://doi.org/10.1109/TCBB.2020.3018467>

REFERENCES (Cont...)

13. Summrina Kanwal, Faiza Khan, Sultan Alamri. (2022). A multimodal deep learning infused with artificial algae algorithm – An architecture of advanced E-health system for cancer prognosis prediction, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part A, Pages 2707-2719, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2022.03.011>.
14. Li, X., Jonnagaddala, J., Cen, M., Zhang, H., & Xu, S. (2022). Colorectal Cancer Survival Prediction Using Deep Distribution Based Multiple-Instance Learning. *Entropy (Basel, Switzerland)*, 24(11), 1669. <https://doi.org/10.3390/e24111669>
15. Wu, Xing & Fang, Qiulian. (2022). Stacked Autoencoder Based Multi-Omics Data Integration for Cancer Survival Prediction. [10.48550/arXiv.2207.04878](https://arxiv.org/abs/10.48550/arXiv.2207.04878)
16. Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics (Oxford, England)*, 35(14), i446–i454. <https://doi.org/10.1093/bioinformatics/btz342>
17. Huang, Zhi & Zhan, Xiaohui & Xiang & Huang, Kun. (2019). SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in Genetics*. 10. [10.3389/fgene.2019.00166](https://doi.org/10.3389/fgene.2019.00166)

REFERENCES (Cont...)

18. Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, <https://doi.org/10.1109/TCBB.2018.2806438>
19. Jefferson, & Horan, M. A. (1997). Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer*, 79(7). <https://doi.org/10.1002>
20. Nguyen, Cuong & Wang, Yong & Nguyen, Ha-Nam. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*. 06. 551-560. 10.4236/jbise.2013.65070
21. Xu, Xiaoyi & Zhang, Ya & Zou, Liang & Wang, Minghui & Li, Ao. (2012). A gene signature for breast cancer prognosis using support vector machine. 5th International Conference on Biomedical Engineering and Informatics. 928-931. 10.1109/BMEI.2012.6513032
22. Curtis, C., Shah, S., Chin, SF. et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature* 486, 346–352. <https://doi.org/10.1038/nature10983>

THANK YOU

