UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA

FINAL YEAR PROJECT PRESENTATION

RECURSIVE FEATURE ELIMINATION AND CONVOLUTIONAL NEURAL NETWORK FOR MULTIMODAL BREAST CANCER SURVIVAL CLASSIFICATION

Prepared by: Chang Min Xuan
Matric No: A20EC0024
Supervisor: Dr. Nies Hui Wen

Presentation Video Link: https://youtu.be/aXM-k-v6EeU
Demo Video Link: https://youtu.be/HIzgZ7FtQRs

OUTLINE

01 **INTRODUCTION**

02 **LITERATURE REVIEW**

03 **RESEARCH METHODOLOGY**

04 **RESEARCH DESIGN AND IMPLEMENTATION**

05 **RESULTS, ANALYSIS AND DISCUSSION**

06 **CONCLUSION**

Chapter 01

INTRODUCTION

INTRODUCTION

Breast cancer occurs due to the **uncontrollable growth of cells in the breast**, which most likely to appear in women. It has become so common in recent years with the increasing **incidence rates** by 0.5%. **Deep learning** is preferred over traditional machine learning methods due to its ability in handling **complex and huge breast cancer datasets**. It may also provide better **feature representation** since it is able to select **optimal and relevant features** from raw data without human interference.

PROBLEM BACKGROUND

- If breast cancer can be **discovered in the early stage**, it may provide **better cancer outcomes** and there might be more options for **cancer treatment** which secure **high chances of breast cancer survivability**
- An **accurate predictive model** is essential for the classification of survival of breast cancer
- **Heterogeneity** of multimodal datasets and difficulties in determination of **driver genes and functional genes** are the problems that would be found in this research

PROBLEM STATEMENT

While **multimodal datasets** provide **comprehensive information** for predicting **breast cancer survival**, their **complexity**, due to **diverse data types** and sources, **complicates analysis** and integration, leading to potential **data inconsistency and interpretative challenges**

RESEARCH AIM

Enhance the classification of breast cancer survival outcomes using recursive feature elimination (RFE) and convolutional neural networks (CNN) architecture

RESEARCH OBJECTIVES

- To investigate the survival of breast cancer using convolutional neural network (CNN)
- To experiment recursive feature elimination (RFE) and existing convolutional neural network (CNN) architectures to improve for multimodal breast cancer survival classification
- To evaluate and compare the model performance using accuracy, precision, sensitivity, specificity and Matthew's correlation coefficient

RESEARCH SCOPES

01

DATASET

The METABRIC dataset is available at a web-based platform, cBioPortal, with the following link: https://www.cbioportal.org/study/summary?id=brca_metabric.

02

MODEL DEVELOPMENT

The algorithm to be applied in this research paper is deep learning method, convolutional neural network (CNN) as feature extraction and random forest (RF) as final classification.

03

PERFORMANCE MEASUREMENT

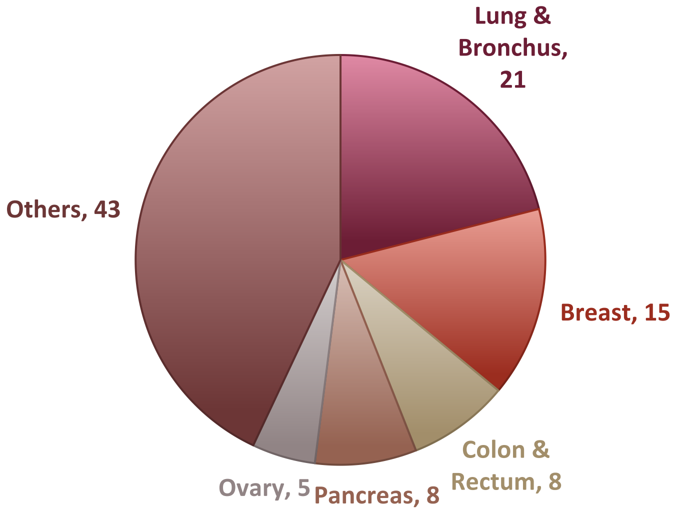
The performance evaluation to be used in this research are sensitivity, specificity, accuracy, precision and Matthew’s correlation coefficient while 10-fold cross validation will also be used for performance measurement.

Chapter 02

LITERATURE REVIEW

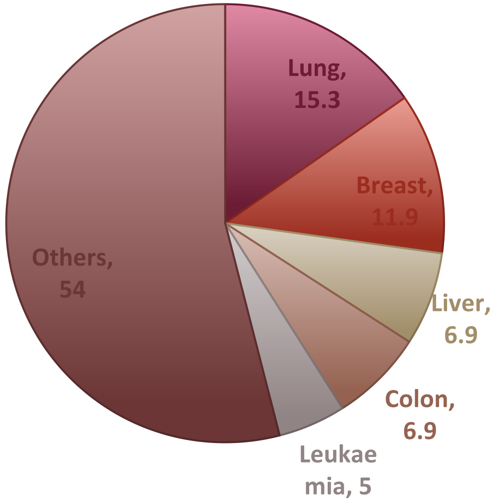
SURVIVABILITY OF BREAST CANCER

Top 5 Mortality Rate Cancer Cases in US



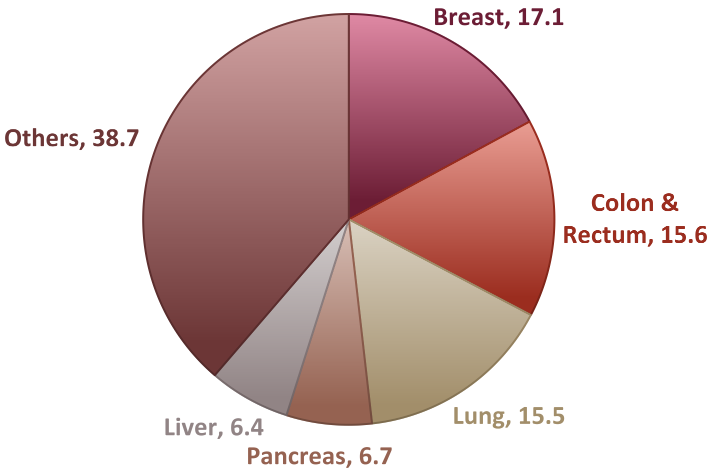
(Siegel *et al.*, 2023)

Top 5 Mortality Rate Cancer Cases in Malaysia



(*Malaysia - International Agency for Research on Cancer*, 2020)

Top 5 Mortality Rate Cancer Cases in Singapore



(Lee *et al.*, 2019)

BREAST CANCER

Causes

1. Dietary Habits
2. Reproductive Factors
3. Lifestyle Factors
4. Genetic Mutations

Signs and Symptoms

1. Early Diagnosis and Self-Observation
2. Diagnosis Procedures
   - Diagnostic & Screening Mammograms
3. Advanced Diagnostic Tests
   - Biopsies & Breast MRIs

Treatments

1. Surgical Options
   - Mastectomy
   - Lumpectomy

2. Radiotherapy
   - Reduces cancer recurrence and mortality rates

3. Chemotherapy
   - Neoadjuvant (before primary treatment)
   - Adjuvant (after primary treatment)

MULTIMODAL DATA

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Benefits

1. Accurate Classifications
   - Enhances classification accuracy and insights.
2. Robust Models
   - Compensates for missing/noisy data, leading to more reliable predictions.
3. Tailored Treatments
   - Combines diverse data for effective patient-specific treatments.
4. Personalized Plans
   - Guides personalized treatment plans, improving outcomes.

Data Sources

- Medical Records: Contain clinical data including healthcare status, clinical observations, diagnostic results, medications, patient demographics, and financial information.
- Radiology Scans: Medical imaging tests for capturing images of body parts.
- Histology Slides: Microscopic studies of biological specimens.
- Genomics: Study of genes and their functions.
- Proteomics: Study of proteins expressed by genes.
- Spatial Transcriptomics: Measures gene activity within specific tissue samples.
- Treatment Regimen: Refers to the patient's treatment plan.
- Familial History: Record of the biological family's medical history.

DEEP LEARNING

UTM UNIVERSITI TEKNOLOGI MALAYSIA

1 ARTIFICIAL NEURAL NETWORK (ANN)

- **Definition**: ANNs are computational algorithms used for modeling complex patterns and forecasting issues.
- **History**: Invented by psychologist Frank Rosenblatt in 1958, aimed at modeling human visual information processing and object recognition (perceptron).
- **Usage**: Used by researchers for studying human cognition.

2 DEEP NEURAL NETWORK (DNN)

- **Definition**: DNNs are a type of ANN with multiple hidden layers between the input and output layers.
- **Extension**: DNNs are an advanced extension of the shallow ANN structure.
- **Complexity**: Referred to as deep nets, indicating a neural network with significant complexity due to additional hidden layers.

3 RECURRENT NEURAL NETWORK (RNN)

- **Definition**: RNNs are a type of artificial neural network (ANN) designed to process sequential data and perform time series prediction.
- **Internal Memory**: RNNs have internal memory that allows them to memorize previous inputs.
- **Selective Memory**: LSTM uses gates (forget, input, and output gates) to selectively remember or forget information.

4 CONVOLUTIONAL NEURAL NETWORK (CNN)

- **Definition**: CNNs are a type of deep learning model primarily used in image and speech recognition.
- **Feature Extraction**: CNNs automatically extract data features using convolutional structures, eliminating the need for manual feature extraction.
- **Adoption**: Gained widespread use in the 2010s for large-scale classification tasks requiring advanced performance.

ADVANTAGES & DISADVANTAGES OF DEEP LEARNING MODELS

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Models	Advantages	Disadvantages
ANN	• Fault Tolerance • Fast and Parallel Processing Capabilities • Flexible and Non-linear	• Manual Feature Engineering • Limited Contextual Understanding • Less Efficient in Handling Large-Scale Data
DNN	• Automatic Feature Learning • Scalability • Able to Handle Structured and Unstructured Data	• Complexity and Longer Training Time • Prone to Overfitting • Computationally Intensive • Lack of Interpretability
RNN	• Able to Handle Sequential Data • Memory	• Incapable of Long-Term Dependencies • Slow Training Procedures • Prone to Gradient Vanishing and Exploding
CNN	• Efficient Image Processing • Automated Feature Extraction • Transfer Learning	• Prone to Overfitting • Transfer Learning Challenges

PREVIOUS RESEARCH ON DEEP LEARNING & BREAST CANCER SURVIVAL

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

No	Datasets	Method	Case Study	Performance Measurements	References
1	• Gene Expression Profile • Copy Number Alteration data • Clinical Data	• Multimodal Deep Neural Network by integrating Multi-dimensional Data (MDNNMD) • Feature Selection: Maximum Relevance-Minimum Redundancy (mRMR)	Breast Cancer Prognosis Prediction	Accuracy: 82.6% Precision: 74.9% Sensitivity: 45% Matthew correlation coefficient: 48.6%	(Sun *et al.*, 2020)
2	• Gene Expression Data • Copy Number Alteration Data • Clinical Data	• Multimodal data adversarial representation framework (MDAR) • Feature extraction: Multi-scale bilinear convolutional neural network (MS-B-CNN) • Feature Selection: Maximum Relevance-Minimum Redundancy (mRMR)	Predict the final breast cancer prognosis	Accuracy: 92.5% Precision: 91.1% Sensitivity: 81.2% Matthew correlation coefficient: 79.4% F1 score: 86%	(Du & Zhao, 2023)
3	• Gene Expression Profile • Copy Number Alteration Data • Clinical data	• Attention-based multimodal deep learning model • First phase: Sigmoid gated attention convolutional neural network • Second phase: Bimodal attention layer • Feature Selection: Maximum Relevance-Minimum Redundancy (mRMR)	Prediction of prognosis of breast cancer	Accuracy: 91.2% Precision: 84.1% Sensitivity: 79.8%	(Kayikci & Khoshgoftaar, 2023)
4	• Gene Expression Profile • Copy Number Alteration Data • Clinical data	• Multimodal Graph Neural Network framework (MGNN) • First part: Bipartite graph to process graph data • Second part: Graph Neural Network (GNN) layer to aggregate information from bipartite graph • Third part: Multimodal fusion neural layer to fuse medical features	Predicting survival of cancer patients	Accuracy: 95.4% Precision: 96.4% Sensitivity: 97.6% Matthew correlation coefficient: 87.5%	(Gao *et al.*, 2021)

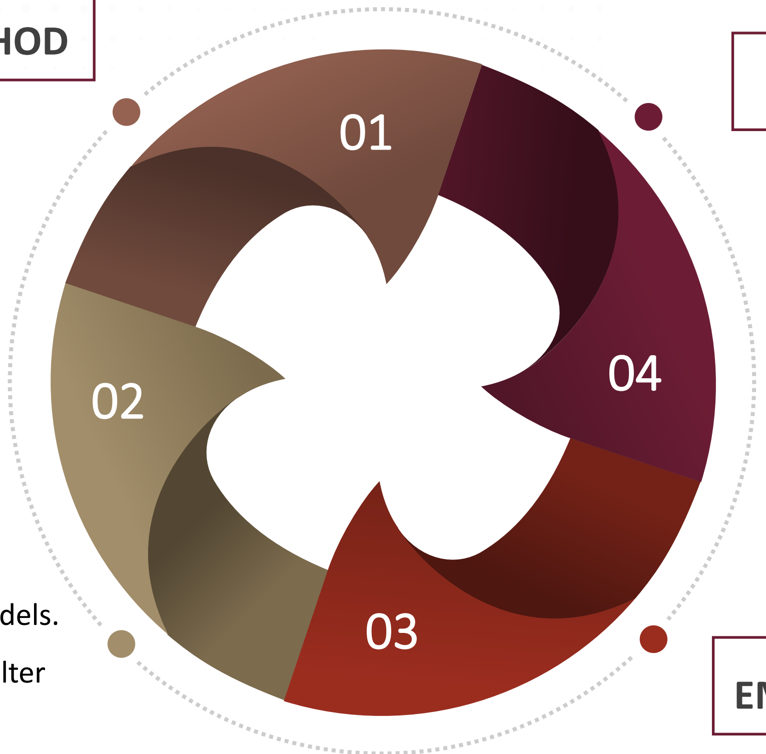
FEATURE SELECTION

FILTER METHOD

- **Selection Basis:** Intrinsic and statistical properties of data.
- **Advantages**: Low computation time and cost-effective.
- **Disadvantages**: Limited interaction with classifiers, potentially reducing predictive performance.
- **Examples**: ANOVA, correlation coefficient, variance threshold, mRMR.

WRAPPER METHOD

- **Selection Basis**: Performance of specific machine learning models.
- **Advantages**: Often achieves better prediction accuracy than filter methods.
- **Disadvantages**: Higher computation time, prone to overfitting with small datasets.
- **Examples**: Forward feature selection, backward feature elimination, stepwise feature selection, recursive feature elimination, exhaustive feature elimination.



HYBRID METHOD

- **Approach**: Combine strengths of filter and wrapper methods.
- **Advantages**: Aims for optimal feature subsets by leveraging both types of methods.
- **Challenges**: Handling the complexity of high-dimensional data effectively.
- **Examples**: Methods that integrate filter-based and wrapper-based approaches sequentially.

EMBEDDED METHOD

- **Integration**: Feature selection integrated within the model creation process.
- **Advantages**: Less computationally expensive than wrapper methods, efficient use of resources.
- **Disadvantages**: More complex to implement than filter methods.
- **Examples**: Lasso regularization, Ridge regression, decision trees, random forests.

RECURSIVE FEATURE ELIMINATION

- RFE iteratively removes the least important features until a desired number is reached, enhancing model simplicity and interpretability.
- RFE ranks features based on their importance and integrates with machine learning models to improve prediction accuracy.

ADVANTAGES

- Flexible, Easy to Use, and Effective in Selecting Relevant Features
- Enhancing Interpretability and Robustness
- Improving Robustness and Generalization
- Prevent Overfitting through Cross-Validation

Chapter 03

RESEARCH METHODOLOGY

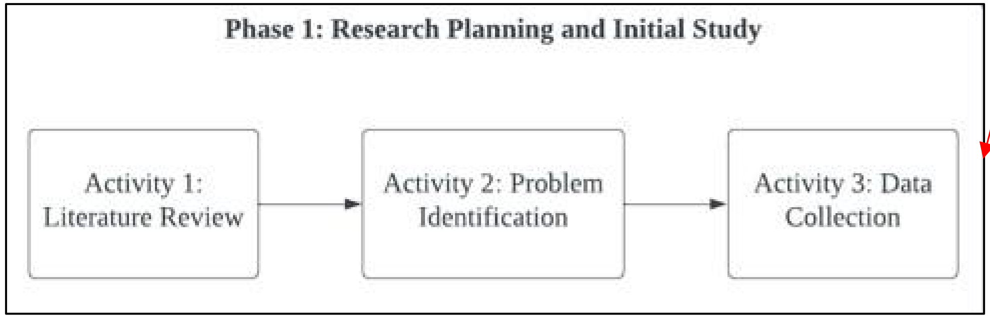
RESEARCH FRAMEWORK

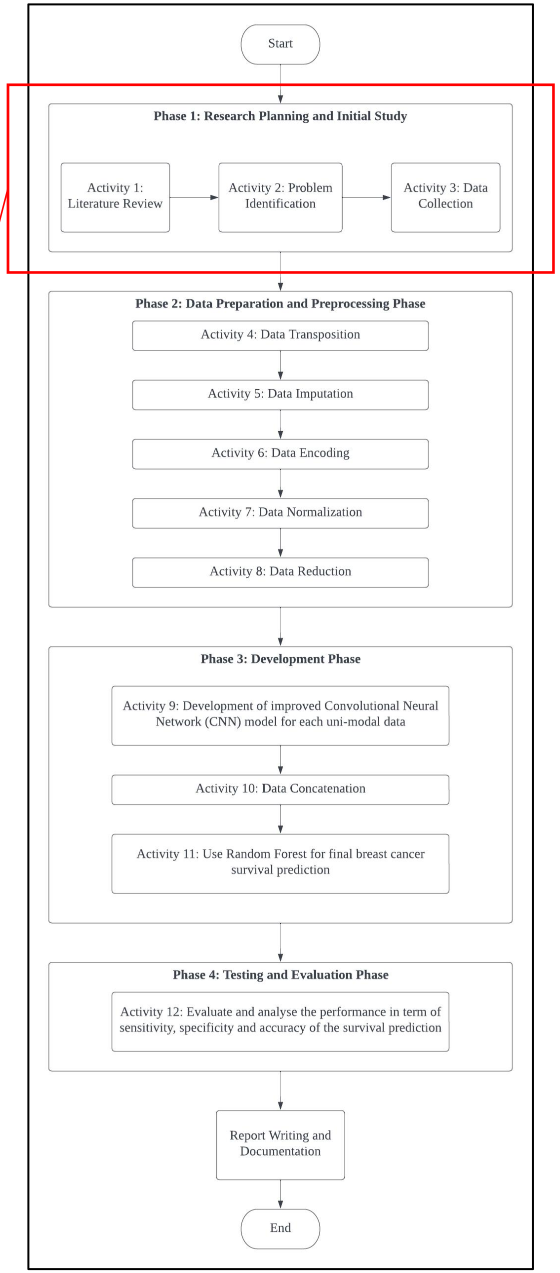
- **Phase 1**

- Literature review is vital as it is an extensive summary and outline of the research done previously by the researchers on a certain study topic.

- Clarify and define the problem faced so that we can search for the solutions effectively by conducting a suitable research on it

- Gather related information for further analysis and evaluation

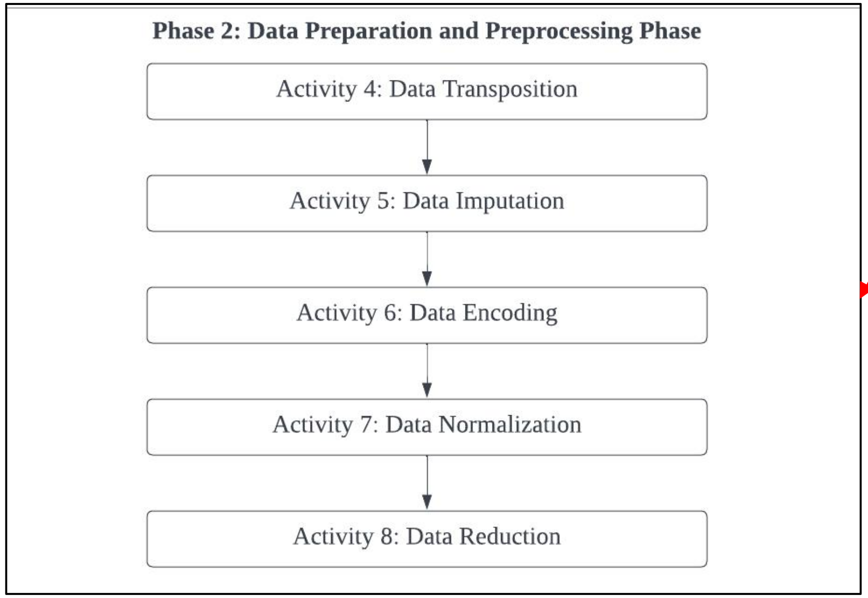


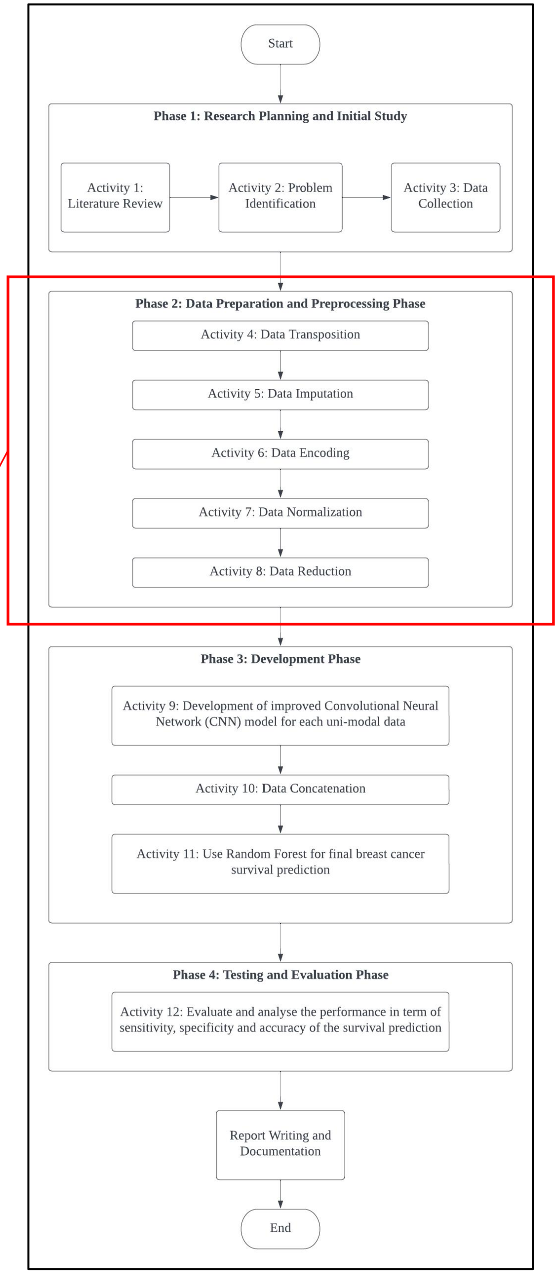


RESEARCH FRAMEWORK

- **Phase 2**

- Data transposition is to rearrange the structure of data, converting rows into columns and vice versa.
- Data imputation requires to fill up the predicted or estimated missing values
- Data encoding turns categorical data into numerical representation for further data analysis
- Data normalization works with numerical data to ensure all the features are on the same scale
- Data reduction reduce the complexity or dimensionality of datasets





RESEARCH FRAMEWORK

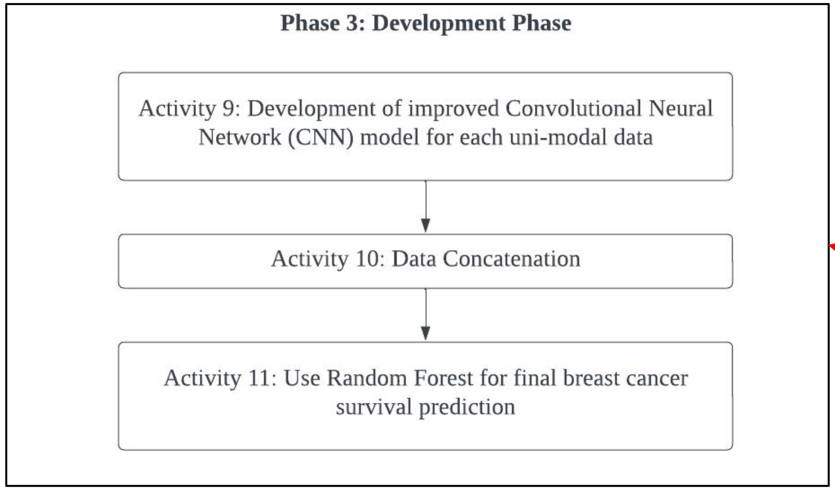
- **Phase 3**

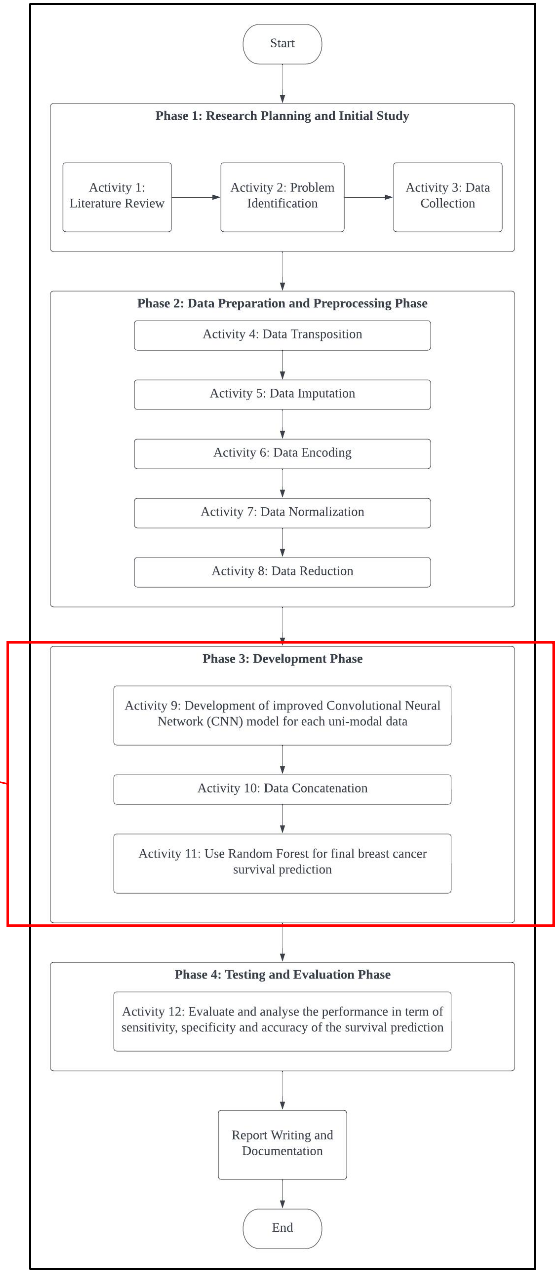
- Uses separate CNN for each unimodal data, that is clinical data, gene expression data as well as copy number alteration (CNA)

- Concatenate all the hidden features extracted from the CNN model so that it can form the stacked features

---

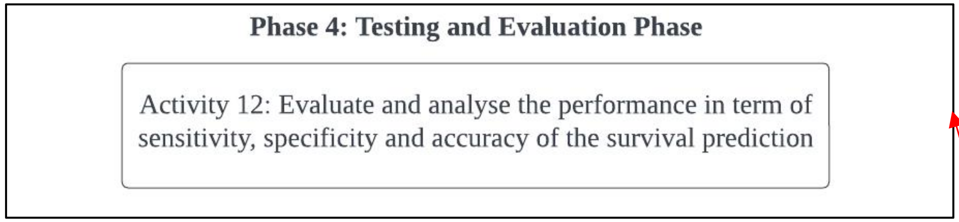
- Stacked features extracted from phase one are passed as the input for the model in phase two
- Random forest is used for final classification of output

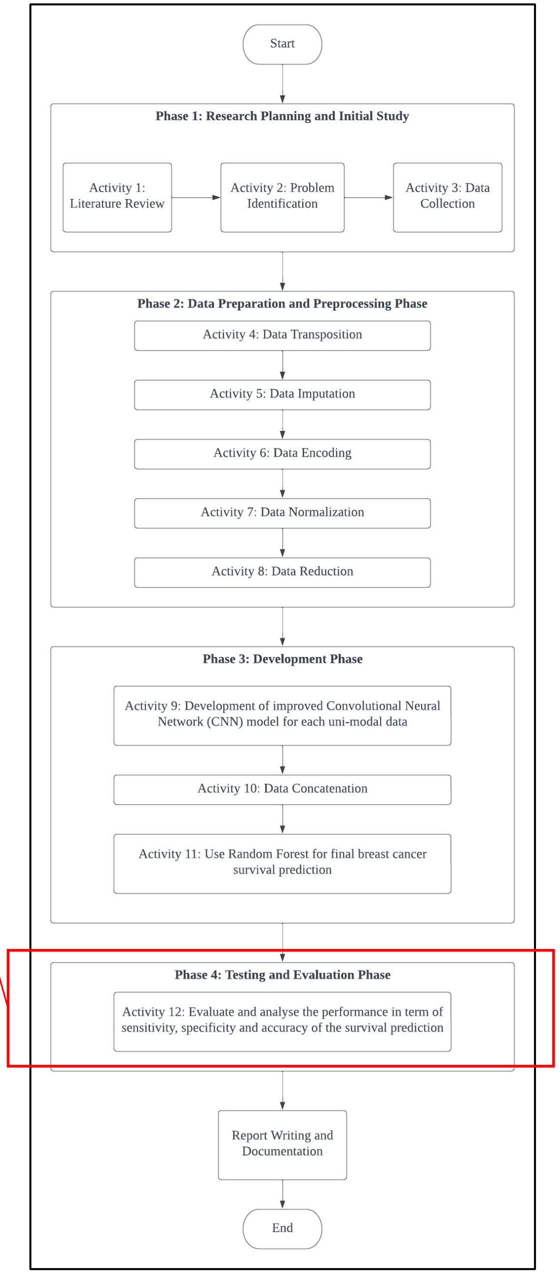




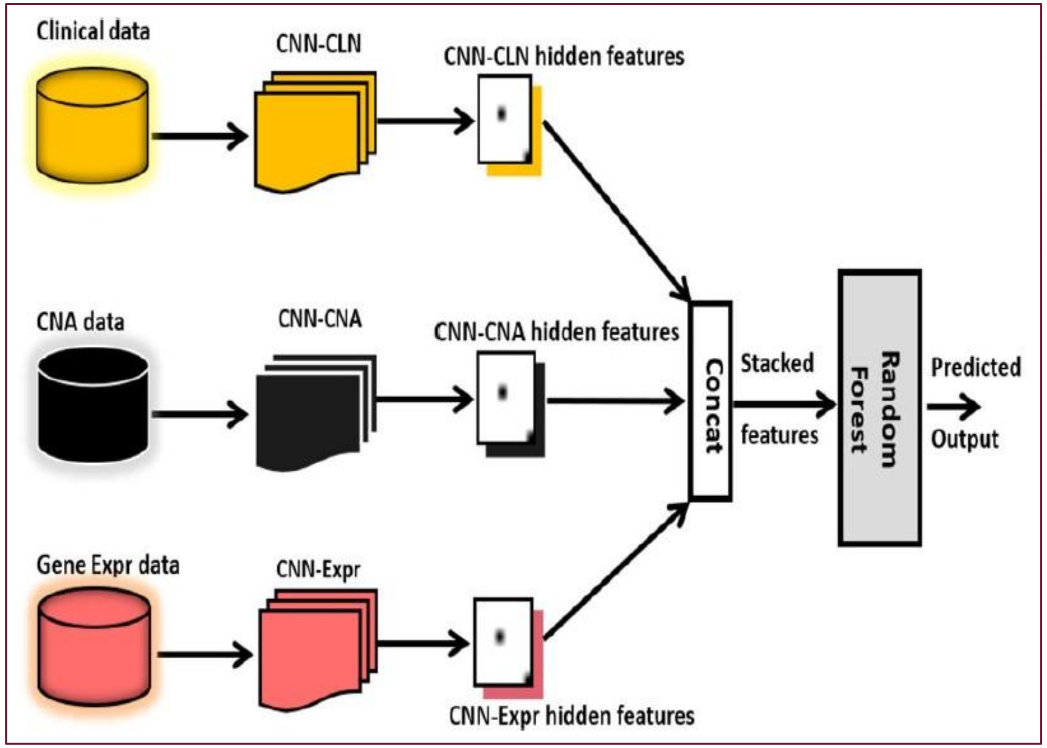
RESEARCH FRAMEWORK

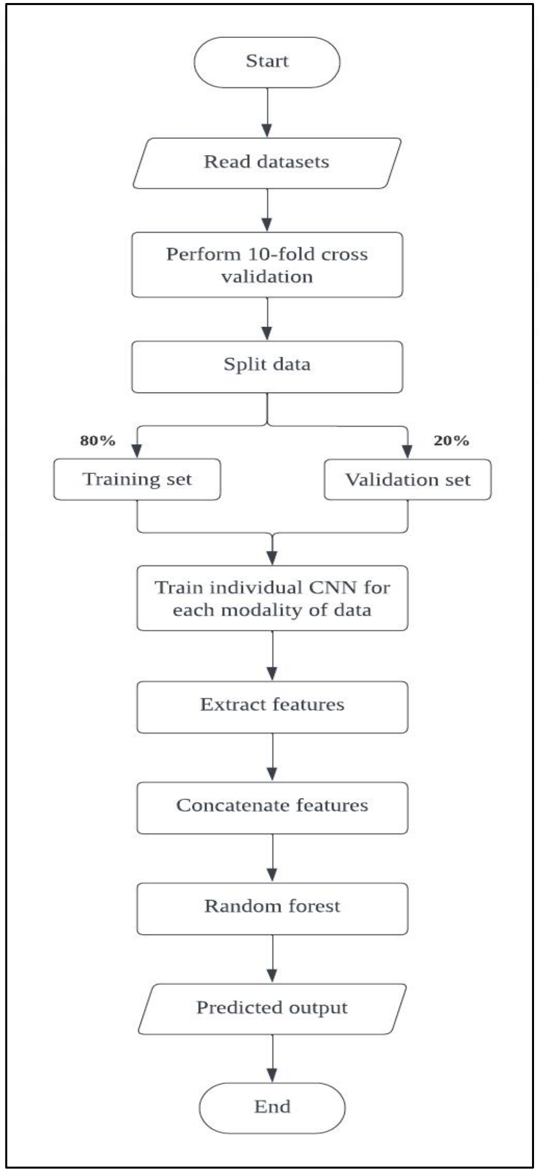
- **Phase 4**
- Calculate the performance metrics
- Compare the results of performance measurements





Breast Cancer Survival Model Architecture (Arya & Saha, 2022)

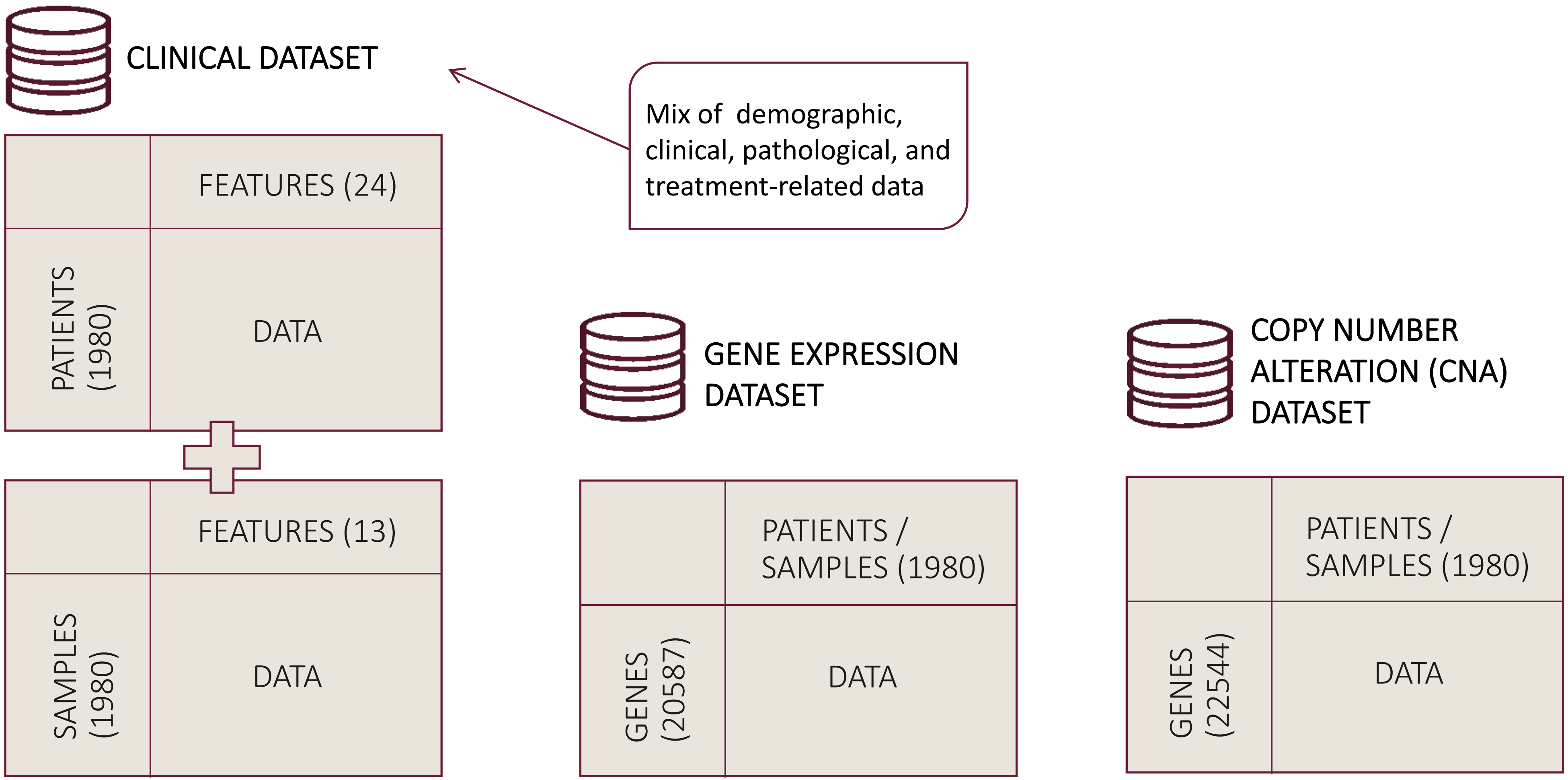




Flowchart for Model Architecture

DATASET

- The dataset used in this research is Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset



PERFORMANCE MEASUREMENTS

Confusion Matrix Table

		Actual Class	
		Positive (P)	Negative (N)
Predicted class	Positive (P)	True Positive (TP)	False Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

Equations of Performance Evaluations
Sensitivity $(Sn) = \frac{TP}{TP+FN}$
Specificity $(Sp) = \frac{TN}{TN+FP}$
Precision $(Pre) = \frac{TP}{TP+FP}$
Accuracy $(Acc) = \frac{TP+TN}{TP+TN+FP+FN}$
Matthew's correlation coefficient $(Mcc) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}}$

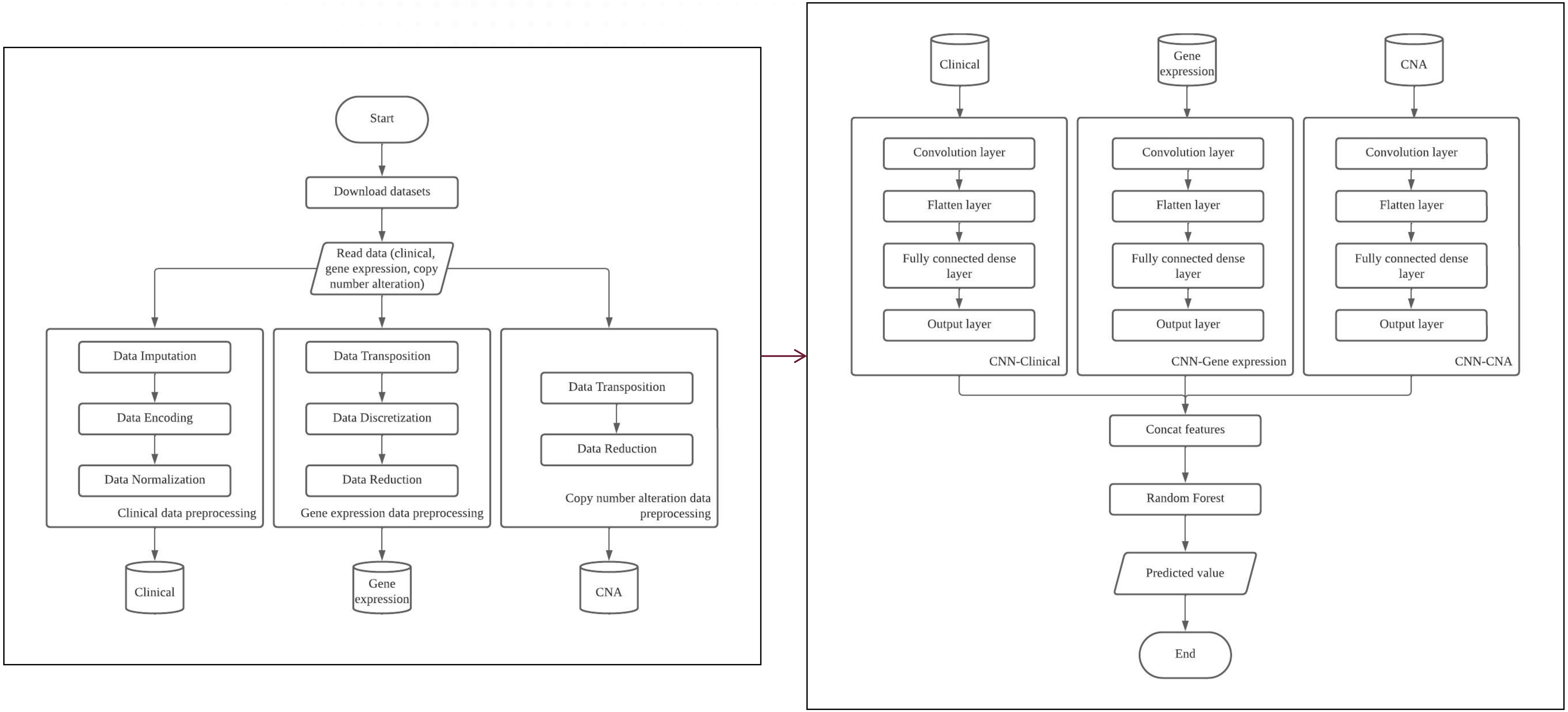
Where:
TP = True Positive, correct positive predictions
FP = False Positive, incorrect positive predictions
TN = True Negative, correct negative predictions
FN = False Negative, incorrect negative predictions

UTM UNIVERSITI TEKNOLOGI MALAYSIA
www.utm.my

Chapter 04

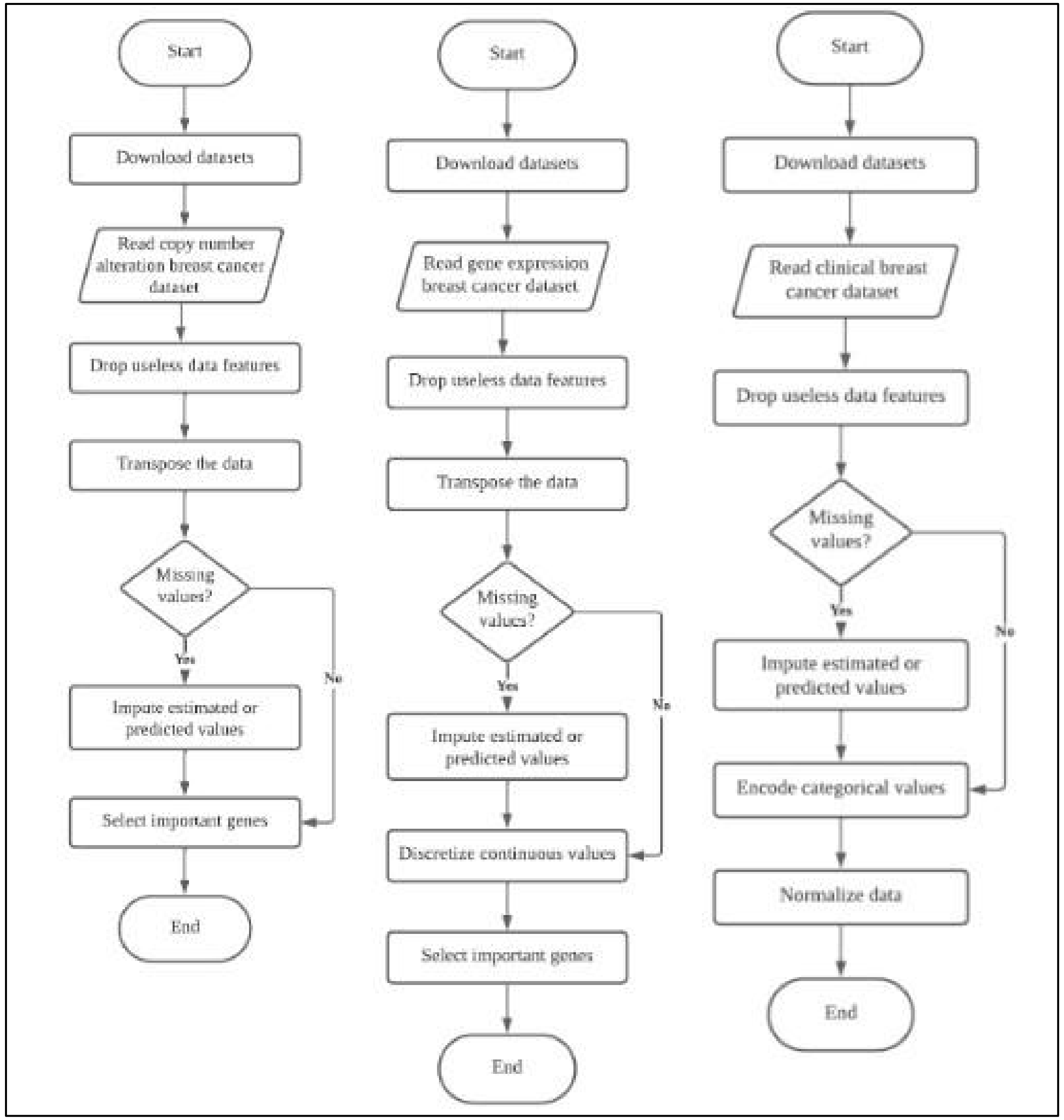
RESEARCH DESIGN AND IMPLEMENTATION

EXPERIMENT DESIGN



DATA PREPROCESSING

UTM UNIVERSITI TEKNOLOGI MALAYSIA



MAIN STEPS IN DATA PREPROCESSING:

- DATA TRANSPOSITION
- DATA IMPUTATION
- DATA ENCODING
- DATA NORMALIZATION AND DISCRETIZATION
- DATA REDUCTION

Flowcharts for Data Preprocessing

DATA PREPROCESSING

- Several useless data features are dropped to ease the progress and ensure more accurate classification. Those features are removed as they are not meaningful and had no significance analysis purposes (Zhang et al., 2022).
- Basic Transposition is applied to make sure dataset consistency for further analysis and concatenation.
- Feature selection is undergone as the results for deep learning methods will be influenced and bad performance of results are obtained if high dimensionality of data is used with low sample size.

Summary Table for Datasets after Data Preprocessing

Datasets	Data Shape (rows, columns)		
	Data Preparation	Data Transposition	Data Reduction
Clinical	(1980, 31)	(1980,31)	(1980, 27)
Gene expression	(20597, 1981)	(1980, 20597)	(1980, 400)
Copy number alteration	(22544, 1981)	(1980, 22544)	(1980, 200)

Overall Information for Clinical Dataset

Information	Values
Cut-off (years)	5
Total number of patients	1980
Long time survivors	1489
Short time survivors	491
Median age in diagnosis	61
Average survival (months)	110

CONVOLUTIONAL NEURAL NETWORK

Input Layer:

- The model takes three types of input data:
  - ✓ Clinical data: 27 features
  - ✓ Gene expression data: 400 features
  - ✓ CNA data: 200 features

Convolution Layer:

- The input data is processed through 4 filters of size 15x1, which extract features from the data.
- The resulting feature maps have an input shape of 4 (indicating that four feature maps are produced).

Flatten Layer:

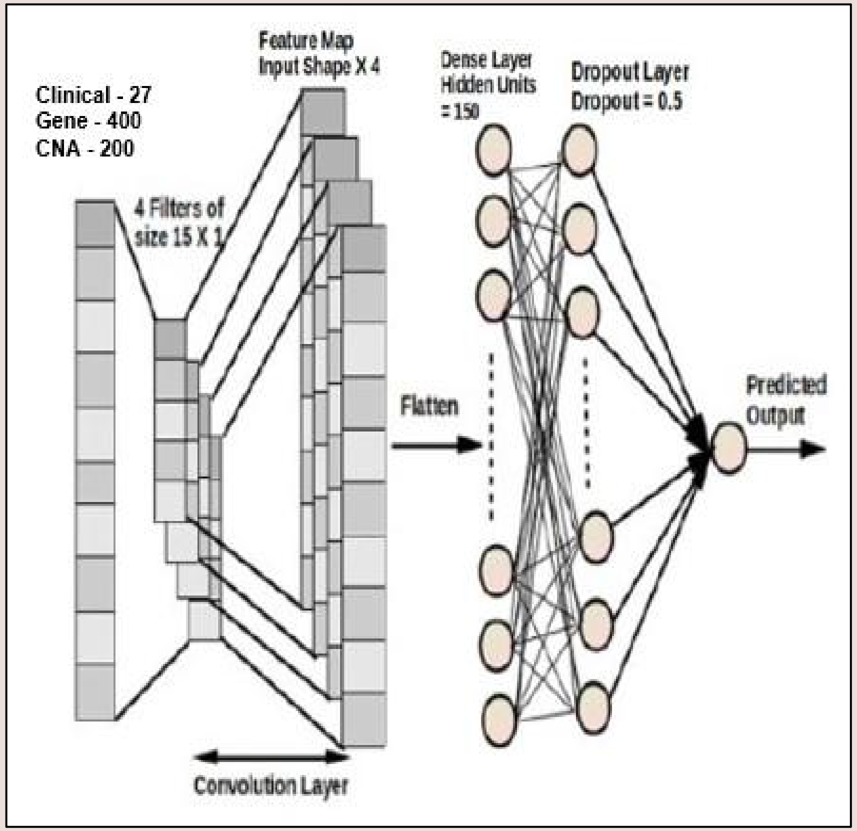
- The feature maps are then flattened into a one-dimensional vector to be fed into the dense layers.

Dense Layer:

- The flattened vector is processed by a dense layer consisting of 150 hidden units.
- A dropout layer with a dropout rate of 0.5 is applied to prevent overfitting.

Output Layer:

- The final dense layer outputs the predicted result, which is likely a classification for breast cancer survival.



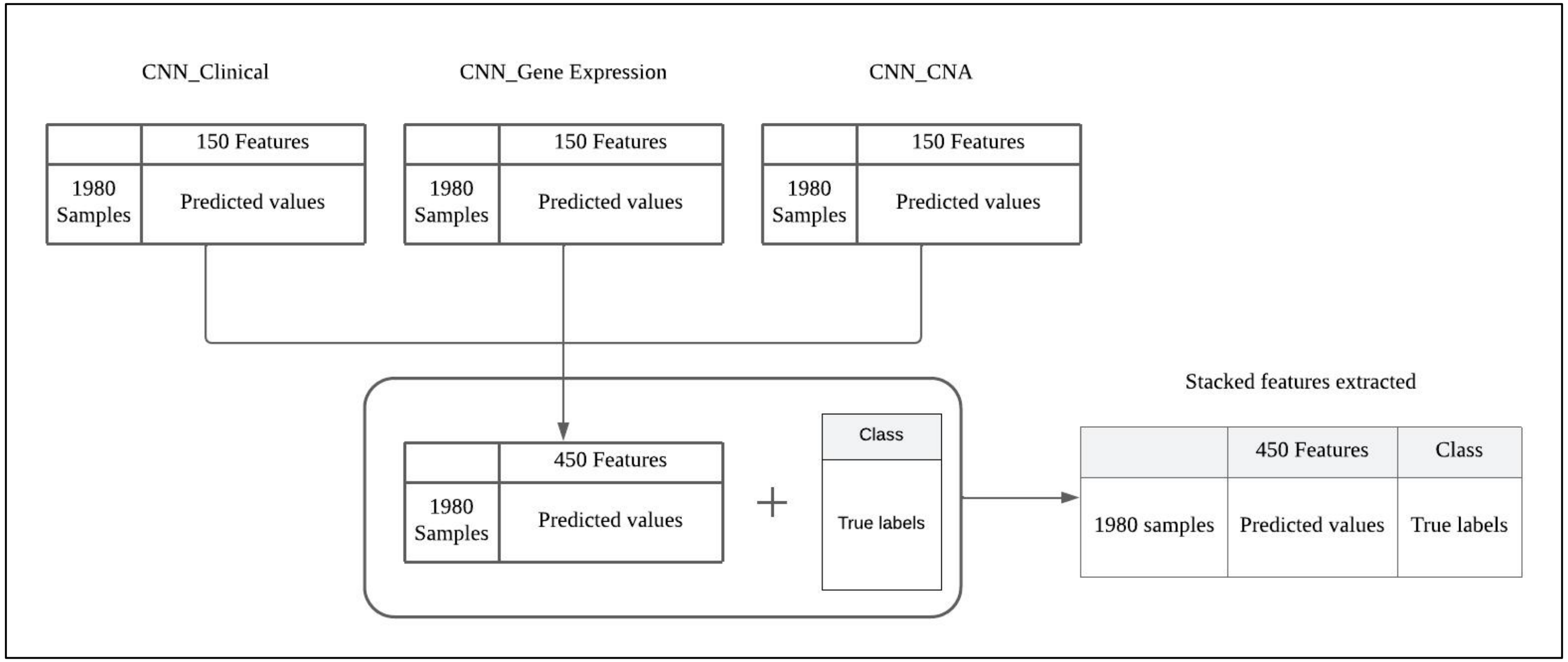
Hyperparameter Configurations of CNN Model

Parameters	Values
Number of convolutional layers	1
Number of filters in convolutional layer	4
Filter size of convolutional layer	15
Stride size in convolutional layer	2
Activation function (Convolution Layer)	TANH
Padding in convolutional layer	Same
Number of hidden layers	1
Number of hidden units in hidden layer	150
Activation function (Dense Layer)	TANH
Optimizer	Adam
Mini-batch size	8
Training epoch	20
Loss function	Binary cross-entropy + L2 Regularization

UTM UNIVERSITI TEKNOLOGI MALAYSIA
www.utm.my

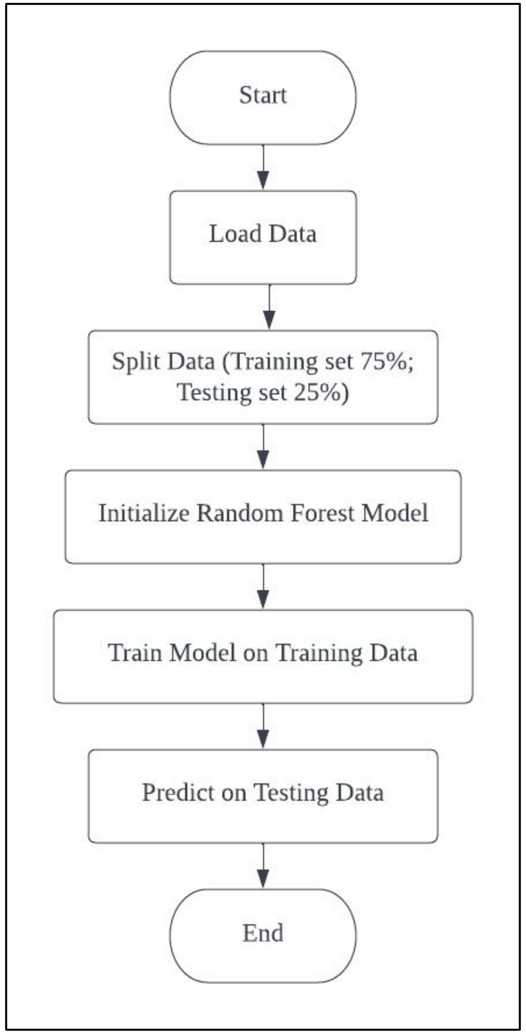
CONCATENATION OF STACKED FEATURES FOR FINAL CLASSIFICATION

- Features are extracted from the hidden layer of a CNN model. Each of the three CNN models produces 150 features from their hidden layers, making a total of 450 features when combined.
- These stacked features serve as input for a Random Forest classifier for the final classification task.



Concatenation Process

Random Forest



www.utm.my

Chapter 05

RESULTS, ANALYSIS AND DISCUSSION

DATA PARTITIONING

- First subset of data is used in allowing model to learn the underlying patterns, relationships, and features from the data
- Second subset of data is kept separate from the training process and acts as an indicator of the model's performance on new, unobserved data

CNN Model

Dataset	Data Division		Total
	Training (80%)	Testing (20%)	
Clinical	1584	396	1980
Gene Expression	1584	396	1980
Copy Number Alteration	1584	396	1980

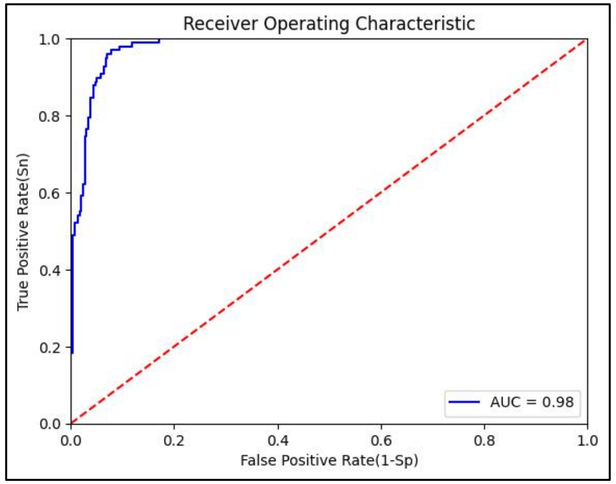
RF Model

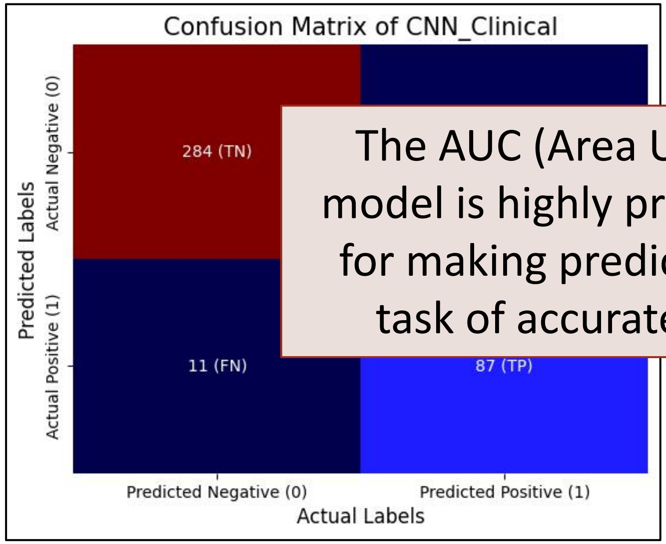
Dataset	Data Division		Total
	Training (75%)	Testing (25%)	
Stacked Metadata (Combination of extracted features from clinical data, gene expression data and copy number alteration data)	1485	495	1980

www.utm.my

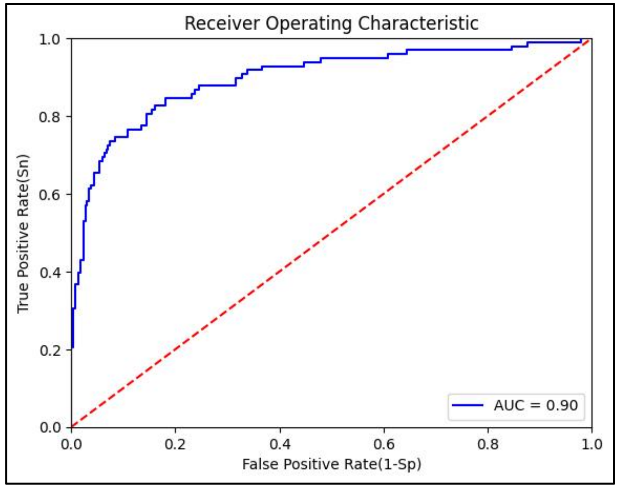
CNN MODEL PERFORMANCE

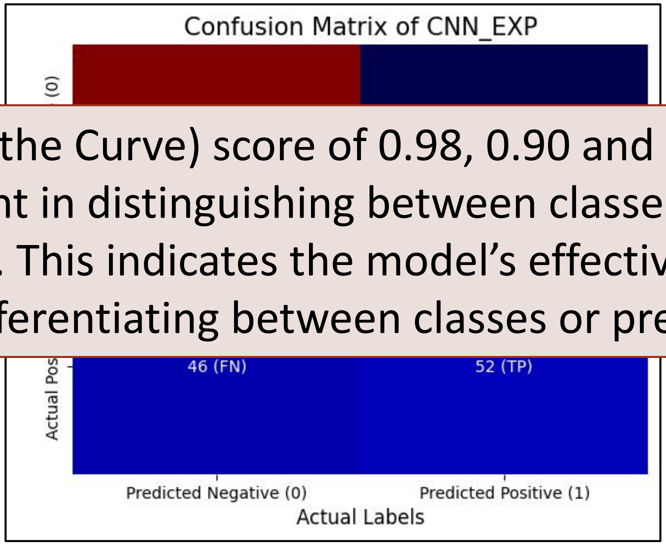
Clinical



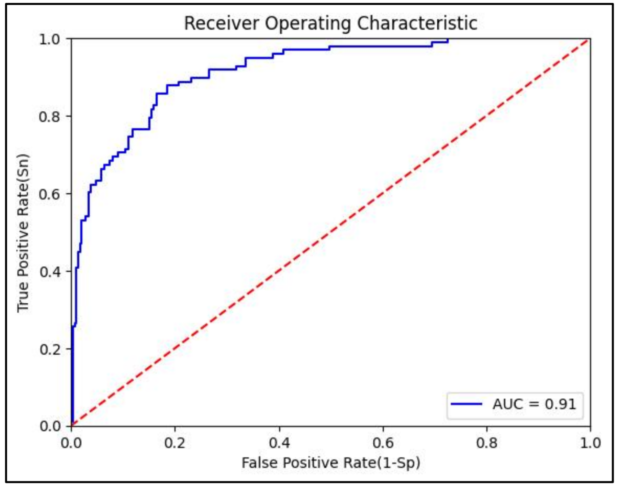


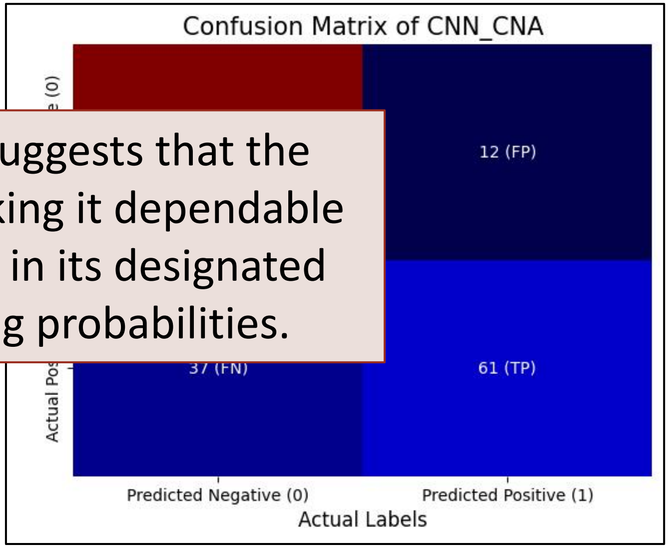
Gene Expression





Copy Number Alteration





The AUC (Area Under the Curve) score of 0.98, 0.90 and 0.91 suggests that the model is highly proficient in distinguishing between classes, making it dependable for making predictions. This indicates the model’s effectiveness in its designated task of accurately differentiating between classes or predicting probabilities.

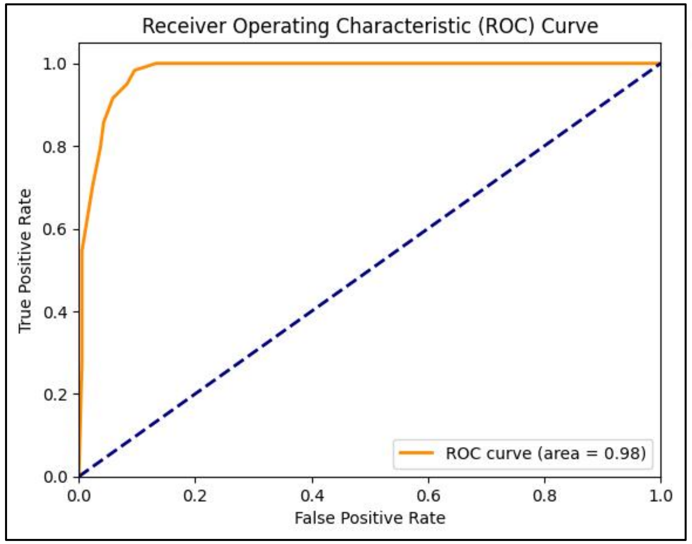
CNN MODEL PERFORMANCE

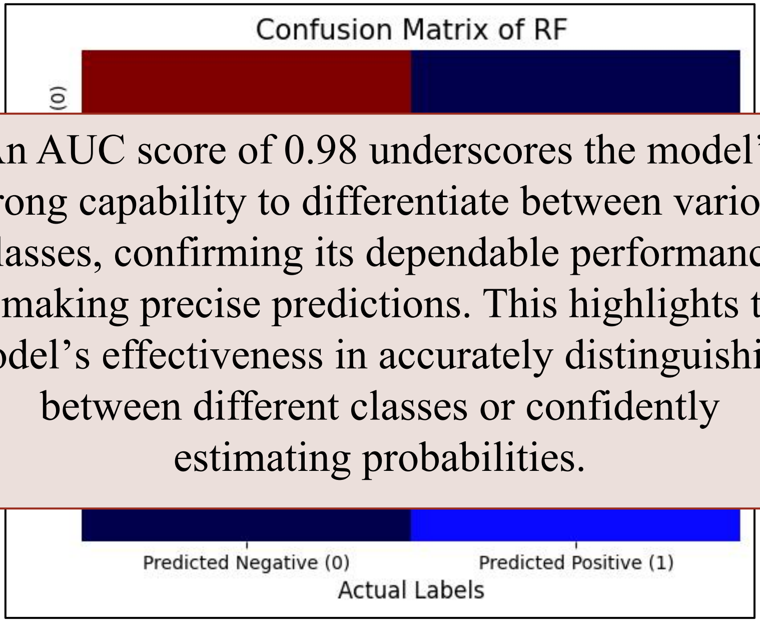
UTM UNIVERSITI TEKNOLOGI MALAYSIA

Comparison on Performance Measurement Results of CNN Model

Performance Metrics	Clinical		Gene Expression		Copy Number Alteration	
	Proposed Model Results	Existing Model Results	Proposed Model Results	Existing Model Results	Proposed Model Results	Existing Model Results
AUC	0.98	0.85	0.90	0.90	0.91	0.87
Sensitivity (%)	88.78	42.86	53.06	46.94	62.24	50.00
Specificity (%)	95.30	94.97	97.32	96.98	95.57	95.30
Precision (%)	86.14	73.68	86.67	83.64	83.56	77.78
Accuracy (%)	93.69	82.07	86.36	84.60	87.63	84.09
Matthew's correlation coefficient	0.8324	0.4650	0.6063	0.5480	0.6479	0.5345

RF MODEL PERFORMANCE





An AUC score of 0.98 underscores the model's strong capability to differentiate between various classes, confirming its dependable performance in making precise predictions. This highlights the model's effectiveness in accurately distinguishing between different classes or confidently estimating probabilities.

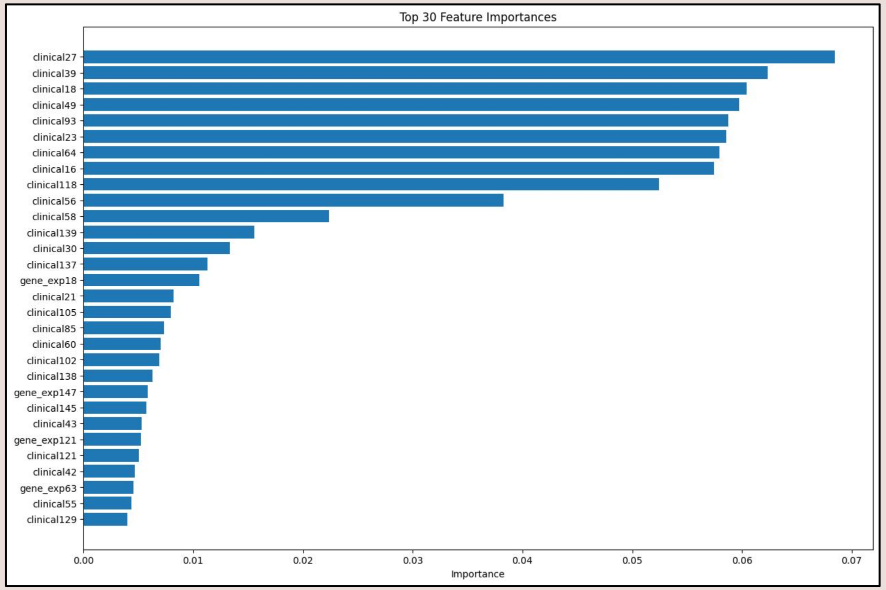
Comparison on Performance Measurement Results of RF Model

Performance Metrics	Proposed Model Results	Existing Model Results
ROC	0.98	0.93
Sensitivity (%)	87.39	59.84
Specificity (%)	95.74	95.11
Precision (%)	86.67	80.85
Accuracy (%)	93.74	86.06
Matthew's correlation coefficient	0.8290	0.6119

The combination of stacked features demonstrates superior performance across all metrics compared to individual datasets, emphasizing the benefit of integrating multiple data modalities for more accurate and reliable classifications of breast cancer survival.

www.utm.my

FEATURE IMPORTANCE



Ranking: Clinical > Gene Expression > CNA

Clinical Data:

- Includes demographic information, medical history, and diagnostic details directly relevant to patient outcomes, hence its dominance in feature importance

Gene Expression Data:

- Certain genes or expression patterns significantly influence survival outcome predictions. These genes likely relate to tumor biology, treatment response, or other critical factors.

Copy Number Alteration Data:

- Low Impact Features: CNA features may not strongly impact survival classification in this dataset.
- Data Quality or Coverage: Issues with data quality or incomplete coverage of CNA features.
- Interplay with Other Data: CNA data might require interaction with other features to contribute meaningfully to predictions rather than independently predicting outcomes.

STATISTICAL SIGNIFICANCE TEST

- A t-test was conducted to compare the cross-validation accuracies between the existing model results and the proposed model results using the METABRIC dataset of 1980 patients.

- **T-test Results**: The t-value obtained was 8.6982, and the p-value was reported as 0.0000. This indicates a significant difference between the two groups, as the p-value is less than 0.05, which is the typical threshold for statistical significance.

- **Mean Difference**: The mean difference in accuracy between the two groups was 0.0625. The 95% confidence interval for this difference was calculated to be between 0.05 and 0.08, further supporting the statistical significance of the t-test result.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA
www.utm.my

BIOLOGICAL VALIDATION

- Identify the correlation of genes obtained after feature selection with breast cancer survival
- 127 out of 600 genes are related with breast cancer prognosis markers
- 456 out of 600 are related with survival cases
- 101 out of 600 genes are really contributed to the breast cancer survival analysis

High expression of **CD44** in breast cancer, particularly in triple-negative subtype, is associated with poorer survival outcomes and contributes to aggressive tumor behavior through mechanisms involving cell proliferation, migration, and epithelial-mesenchymal transition (EMT).

High expression of **PIK3CD** in breast cancer is associated with poor prognosis, aggressive tumor characteristics, reduced survival rates, and therapeutic resistance, making it a significant prognostic biomarker and potential therapeutic target.

High stromal STAT3 expression and enriched **CD27** in breast cancer, particularly TNBC, are associated with poorer survival outcomes due to their role in promoting an inflammatory tumor microenvironment.

Downregulation of the tumor suppressor gene **STARD13**, often due to high miR-21-3p levels, is associated with disease progression and shorter breast cancer-specific survival.

Chapter 06

CONCLUSION

ACHIEVEMENTS

- Reviewed relevant data and literature to understand the research focus, algorithms, and CNN and RF models in detail

- This preparation ensures efficient experiment execution and the acquisition of multimodal datasets, achieving RO1

- Implemented preprocessing steps, using RFE as feature selection

- The cleaned datasets serve as inputs for a CNN model for feature extraction, with outputs concatenated to form stacked features for final classification via random forest.

- The success in preprocessing and classification confirms the achievement of RO2.

- Analyzed, evaluated and compared the model performance with existing model

- Obtained performance measurement results in classifying short-term and long-term breast cancer survival rates, achieving RO3

RESEARCH CONSTRAINTS

1. Dataset Limitations
   - ✓Missing values, varying scales, and different formats across datasets
   - ✓Require careful preprocessing and normalization

2. Feature Selection Complexity
   - ✓Computational complexity
   - ✓Time demands

3. Computational Resources
   - ✓Require powerful computing resources

SUGGESTIONS FOR IMPROVEMENTS & FUTURE WORKS

1. Addressing Computational Resources Limitations
   - ✓Google Colab can speed up model training and evaluation

2. Increasing Dataset Size
   - ✓Improve the informativeness and results of the experiments

3. Expanding Research Scope
   - ✓Incorporating additional data types such as gene methylation and imaging data, along with image processing techniques using breast cancer tissue images

UTM UNIVERSITI TEKNOLOGI MALAYSIA
www.utm.my

THANK YOU