

LB/TH/43/2025  
TH6017

**AN OPTIMIZED SUPERVISED LEARNING MODEL FOR PREDICTING  
SURVIVAL RATE**

T.M.D. Saumya

239356X

MSc in Computer Science

Department of Computer Science and Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

June 2025

## DECLARATION

I declare that this is my own work, and this MSc Research report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

Signature:

Date: 2025/06/22

I certify that the declaration above by the candidate is true to the best of my knowledge and that this project report is acceptable for evaluation for the MSc Research.

Name of Supervisor: Prof. Indika Perera

Signature of the Supervisor:

Date: 2025/06/22

## **ACKNOWLEDGEMENT**

I wish to extend my sincere appreciation to my supervisor Prof. Indika Perera for the continuous supervision and encouragement provide during this research study. And I would like to extend my thanks to University Library and the staff for enlightening me with the research materials required throughout this study. I am also thankful for my parents, daughter and my colleagues for their encouragements and the valuable support during this study.

## ABSTRACT

The incorporation of machine-learning techniques in medical informatics has facilitated major improvements in cancer survivability prognosis. In this study, the concentration is given on improving the accuracy for the prediction of breast cancer survivability and to come up with a prediction model with the METABRIC dataset, which have both clinical and genomic features. Further improvement of the prediction model through feature engineering and missing data handling was another main objective of this study. The study was started with several ML algorithms such as Logistic Regression, Support Vector Classification, Random Forest, Categorical Boosting and Extreme Gradient Boosting with the objective of selecting the best algorithm for this dataset and to further improve the prediction accuracy with algorithm customizations. Of these, algorithms Extreme Gradient Boosting algorithm provided the best baseline accuracy (81.10%), and then further improvements were done to prediction model pipeline to improve prediction accuracy through kernel-based multiple imputation for missing values in the dataset and also with advanced feature engineering techniques like log transformation, interaction features and categorical feature encoding on the METABRIC dataset which resulted in increase of prediction accuracy up to 87.5%. With the objective of further improvement of prediction accuracy, the Extreme Gradient Boosting Algorithm was customized with a new objective function composite with Asymmetric Cost Sensitivity and Smoothed Focal Loss which resulted in further prediction accuracy improvement of 1.5%. The final proposed model pipeline with customized Extreme Gradient Boosting algorithm offers highly accurate and clinically aligned survivability prediction model which can be used as the base for disease prognosis using transfer learning

**Keywords:** Survivability Prediction, Machine Learning, Supervised learning, Extreme Gradient Boosting, METABRIC

# TABLE OF CONTENTS

DECLARATION .....	ii
ACKNOWLEDGEMENT .....	iii
ABSTRACT .....	iv
TABLE OF CONTENTS .....	v
LIST OF FIGURES .....	ix
LIST OF TABLES .....	x
LIST OF ABBREVIATIONS .....	xi
CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1 Background and Motivation.....	1
1.1.1 Introduction to Cancer.....	2
1.1.1 Breast Cancer .....	2
1.2 Breast Cancer and Its Challenges.....	3
1.3 Societal Impact of Cancer and Survivability .....	4
1.4 Understanding Cancer Survivability .....	4
1.5 Role of Machine Learning and Data Mining in Cancer Prediction .....	5
1.6 Background to Research Problem .....	6
1.7 Research Questions .....	7
1.8 Research Objectives .....	7
1.9 Structure of the Thesis .....	8
CHAPTER 2 .....	10
LITERATURE REVIEW.....	10
2.1 IT Applications in Healthcare Sector .....	10
2.2 Data Mining Applications in Healthcare Sector .....	10
2.2.1 Clinical Care.....	10
2.2.2 Administration of Health Services .....	11
2.2.3 Medical Research .....	11
2.2.4 Education and Training .....	11
2.3 Cancer Survivability Prediction Studies .....	11

2.4	Open-Source Software Tools Used for Data Mining Research .....	16
2.4.1	Functionality Aspect .....	17
2.4.2	Usability Aspect.....	17
2.5	Analysis of Gaps and Limitations in Existing Studies.....	18
2.6	State-of-the-Art Models in Cancer Survivability Prediction .....	19
CHAPTER 3	.....	20
METHODOLOGY	.....	20
3.1	Research Approach .....	21
3.1.1	Handling Missing Data in the METABRIC Dataset.....	21
3.1.2	Identification of Key Prognostic Factors .....	21
3.1.3	Comparative Evaluation of Machine Learning Algorithms.....	22
3.1.4	Customization and Enhancement of Selected Algorithm .....	22
3.1.5	Integration of Feature Engineering Techniques .....	22
3.2	Rationale Behind Selected Techniques .....	23
3.2.1	Handling Missing Data in the METABRIC Dataset.....	23
3.2.2	Identifying the Most Important Factors .....	24
3.2.3	Comparing Different Machine Learning Algorithms .....	24
3.2.4	Customizing and Improving the Selected Model.....	24
3.2.5	Feature Engineering for Better Model Insights.....	25
3.3	Research Design.....	25
3.4	Data Source .....	27
3.5	Approach and Technology for Missing Data Handling .....	28
3.5.1	Identifying the Nature of Missingness .....	28
3.5.2	Kernel Imputation for Missing Data .....	29
3.6	Exploratory Analysis for Key Prognosis Factor Identification.....	29
3.6.1	Pearson Correlation Coefficients .....	30
3.6.2	Spearman Correlation Coefficients .....	30
3.6.3	Model-Based Feature Importance and SHAP Analysis .....	30
3.6.4	Comparative Model Deployment.....	31
3.7	Machine Learning Algorithms .....	31
3.7.1	Logistic Regression.....	32
3.7.2	Random Forest Classification .....	32

3.7.3 Support Vector Classification .....	33
3.7.4 XGBoost Algorithm .....	33
3.7.5 CatBoost Algorithm .....	34
3.8 Feature Engineering Methods .....	34
3.8.1 Log Transformation .....	34
3.8.2 Categorical feature encoding .....	34
3.8.3 Feature Interaction Generation.....	35
3.9 Customization of the Selected Algorithm to Improve Accuracy .....	35
3.9.1 Asymmetric Cost Sensitivity.....	35
3.9.2 Smoothed Focal Loss .....	35
3.10 Model Validation .....	35
3.10.1 Cross Validation.....	36
3.10.2 Domain Expert Validation .....	36
3.11 Model Evaluation .....	36
3.11.1 Accuracy .....	36
3.11.2 Precision.....	36
3.11.3 Recall.....	37
3.12 Clinical Relevance of Evaluation Metrics.....	37
3.13 Scientific Contributions of the Research.....	38
CHAPTER 4 .....	39
IMPLEMENTATION .....	39
4.1 Data Selection .....	39
4.2 Data Analysis and Prognosis Factor Identification.....	41
4.2.1 Analysis with Pearson Correlations .....	42
4.2.2 Spearman Correlations Analysis .....	43
4.2.3 SHAP Analysis .....	44
4.3 Data Pre-Processing and Missing Value Handling .....	46
4.3.1 Data Pre-Processing Stages.....	46
4.3.2 Preprocessed Data Segregation .....	49
4.4 Machine Learning Algorithms Used for Model Creation .....	49
4.4.1 Logistic Regression.....	50
4.4.2 Random Forest Classification .....	51

4.4.3 Support Vector Classifier .....	51
4.4.4 Extreme Gradient Boost Algorithm .....	52
4.4.5 Categorical Boosting Algorithm .....	52
4.5 Feature Engineering to Improve Prediction Accuracy .....	53
4.5.1 Log Transformation .....	53
4.5.2 Feature Interaction Generation and Categorical Feature Encoding .....	53
4.6 Customization of XGBoost Algorithm to Improve Prediction Accuracy .....	55
4.6.1 Asymmetric Cost Sensitivity.....	55
4.6.2 Smoothed Focal Loss .....	56
Chapter 5 .....	57
Evaluation .....	57
5.1 Cross-Validation and Accuracy Scores.....	57
5.2 Confusion Matrix .....	57
5.3 Classification Report.....	58
5.4 Evaluation of Machine Learning Algorithms on Prediction .....	59
5.4.1 Predictions with Clinical Data .....	59
5.4.2 Predictions with Genomic Data .....	61
5.4.3 Predictions with Combination of Clinical Data and Genomic.....	63
5.5 Progress of Accuracy on Prediction with Imputation .....	65
5.6 Improvement Achieved through Feature Engineering.....	66
5.7 Accuracy Improvement with XGBoosting Algorithm Customization.....	67
Chapter 6 .....	69
CONCLUSION .....	69
6.1 Results and Insights .....	70
6.2 Novel Algorithm Proposed through this Study for Survivability Prediction...	72
6.3 Prediction Model for Survivability Prediction on METABRIC Dataset .....	72
6.4 Limitations of this study .....	74
6.5 Future works .....	74
References .....	75
APPENDIX A .....	80

## LIST OF FIGURES

<b>Figure</b>	<b>Description</b>	<b>Page</b>
Figure 2.1:	Main Areas in Health Informatics	10
Figure 3.1:	Research Approach	23
Figure 3.3:	Research Design	26
Figure 4.1:	Pearson Correlation	43
Figure 4.2:	Spearman Correlation	44
Figure 4.3:	SHAP Value Analysis	45
Figure 4.4:	Missing data matrix	47
Figure 4.5:	Data Preprocessing Steps in ML Model with Missing Data Handling	48
Figure 4.6:	Train and Evaluate model	50
Figure 4.7:	Pearson Coefficients of Interaction Features	54
Figure 5.1:	Summary on accuracy scores for all the algorithms on clinical data	60
Figure 5.2:	Summary on accuracy scores for all the algorithms on genomic data	62
Figure 5.3:	Summary on accuracy scores for all the algorithms on combination of clinical and genomic data	64
Figure 5.4:	Summary on accuracy scores for all the algorithms on combination of clinical and genomic data without multiple imputations	65
Figure 5.5:	Summary on accuracy scores After Customization of Algorithm	68
Figure 6.1:	Customization of Algorithm	72
Figure 6.2:	Graphical Representation of Prediction Model Pipeline	73

## LIST OF TABLES

<b>Table</b>	<b>Description</b>	<b>Page</b>
Table 2.1:	Evaluation of Open-Source Data Mining Software	17
Table 4.1:	Clinical Factors in Dataset	39
Table 5.1:	Summary on accuracy scores for all the algorithms on clinical data	59
Table 5.2:	Summary on accuracy scores for all the algorithms on genomic data	61
Table 5.3:	Summary on accuracy scores for all the algorithms on combination of clinical and genomic data	63
Table 5.4:	Summary on accuracy scores for all the algorithms on combination of clinical and genomic data without multiple imputations	65
Table 5.5:	Summary on Accuracy Scores After Feature Engineering	66
Table 5.6:	Summary on Accuracy Scores After Customization of Algorithm	67

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Description</b>
DNA	Deoxyribonucleic Acid
UV	Ultraviolet
SEER	Surveillance, Epidemiology, and End Results
KDD	Knowledge Discovery in Databases
CDC	Centers for Disease Control and Prevention's
USA	United States of America
EM	Expectation Maximization
ANN	Artificial Neural Network
METABIC	Molecular Taxonomy of Breast Cancer International Consortium
LR	Logistic Régression
SVC	Support Vector Classification
XGBOOST	Extreme Gradient Boosting
CatBoost	Categorical Boosting
MICE	Multiple Imputation by Chained Equations
ML	Machine Learning