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

LIST OF FIGURES

Figure **Description** **Page**

LIST OF TABLES

Table **Description** **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
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