

Department of Electrical and Computer Engineering
North South University



Senior Design Project

Predicting Characteristics Associated with Breast Cancer Survival using Multiple Machine Learning Approaches

Mohammad Nazmul Haque

ID # 1712859042

Faculty Advisor

Dr. Mohammad Monirujjaman Khan

Associate Professor

Department of Electrical and Computer Engineering, Fall 2021

North South University

Declaration of Authorship

We, declare that this research titled ‘Predicting Characteristics Associated with Breast Cancer Survival using Multiple Machine Learning Approaches’ and the work presented in it are our own. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. Any material reproduced in this project has been properly acknowledged.

Student’s name & Signature

1. **Mohammad Nazmul Haque**

APPROVAL

This research report entitled “Predicting Characteristics Associated with Breast Cancer Survival using Multiple Machine Learning Approaches” has been submitted by **Mohammad Nazmul Haque (ID#1712859042)** is approved in partial fulfillment of the requirement of the Degree of Bachelor of Science in Computer Science and Engineering on May and has been accepted as satisfactory.

Supervisor’s Signature



Dr. Mohammad Monirujjaman Khan

Associate Professor

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh

Department Chair’s Signature

Dr. Mohammad Rezaul Bari

Associate Professor and Chairman

Department of Electrical Engineering and Computer Science

North South University,

Dhaka, Bangladesh

ACKNOWLEDGMENT

First of all, we wish to express our gratitude to the Almighty for giving us the strength to perform our responsibilities and complete the report.

The capstone project program is very helpful to bridge the gap between the theoretical knowledge and real-life experience as part of Bachelor of Science (BSc) program. This report has been designed to have a practical experience through the theoretical understanding.

We also acknowledge our profound sense of gratitude to all the teachers who have been instrumental for providing us the technical knowledge and moral support to complete the project with full understanding.

It is imperative to show our appreciation for our honorable faculty member Dr. Mohammad Monirujjaman Khan for his undivided attention and help to achieve this milestone. Also, our gratefulness is divine to the North South University, ECE department for providing us a course such as CSE 499 in which we could really work on this project and materialize it the way we have dreamt of.

We thank our friends and family for their moral support to carve out this project and always offer their support.

ABSTRACT

Breast cancer is one of the most commonly diagnosed female disorders globally. Numerous studies have been conducted to predict survival markers, although the majority of these analyses were conducted using simple statistical techniques. In lieu of that, this research employed machine learning approaches to develop models for identifying and visualizing relevant prognostic indications of breast cancer survival rates. A comprehensive hospital-based breast cancer dataset was collected from the National Cancer Institute's SEER Program's November 2017 update, which offers population-based cancer statistics. The dataset included female patients diagnosed between 2006 and 2010 with infiltrating duct and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3). The dataset included nine predictor factors and one predictor variable that were linked to the patients' survival status (alive or dead). To identify important prognostic markers associated with breast cancer survival rates, prediction models were constructed using K-nearest neighbor, decision tree, gradient boosting, random forest, adaboost, logistic regression, voting classifier, and support vector machine. All methods yielded close results in terms of model accuracy and calibration measures, with the lowest achieved from Logistic Regression (accuracy = 80.57 percent) and the greatest acquired from random forest (accuracy = 94.64 percent). Notably, the multiple machine learning algorithms utilized in this research achieved high accuracy, suggesting that these approaches might be used as alternative prognostic tools in breast cancer survival studies, especially in the Asian area.