Equity in post-HCT Survival Predictions

Industrial Oriented Mini Project (CS653PC)

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

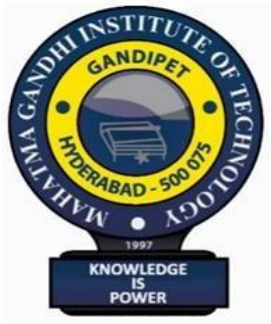
Computer Science and Engineering

by

Mohammed Abdul Kalam Khan (22261A05A4)

Under the guidance of

Dr.KOTOJU RAJITHA (Asst Professor)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MAHATMA GANDHI INSTITUTE OF TECHNOLOGY

(Affiliated to Jawaharlal Nehru Technological University

Hyderabad) Gandipet, Hyderabad-500 075, Telangana (India)

2024-2025

ABSTRACT

Accurate survival prediction for allogeneic Hematopoietic Cell Transplantation (HCT) patients is crucial for improving healthcare outcomes and ensuring equitable treatment across diverse populations. Existing predictive models often exhibit biases related to socioeconomic status, race, and geography, limiting their effectiveness in real-world applications. In this study, we leverage **SurvivalGAN**, a generative adversarial network designed to handle censored time-to-event data, to enhance survival analysis and address fairness disparities in HCT survival predictions.

Our approach involves generating high-quality synthetic survival data that mirrors real-world distributions while preserving the complexities of censoring and time-to-event relationships. By integrating **Kaplan-Meier estimation, Cox Proportional Hazards models, and machine learning techniques like XGBoost and deep learning**, we build predictive models that optimize for both **accuracy and fairness**. To evaluate performance, we employ the **Stratified Concordance Index**, ensuring equitable prediction quality across different racial groups.

Preliminary results indicate that our methodology improves model discrimination power while reducing disparities in survival predictions. By addressing key challenges in censored data modeling and fairness optimization, our work contributes to the development of more inclusive and effective predictive models for HCT patients. This research underscores the potential of AI-driven survival analysis in **personalized medicine, healthcare resource allocation, and equitable treatment planning**.

1.1 INTRODUCTION

The Allogeneic Hematopoietic Cell Transplantation (HCT) is a critical medical procedure for treating various blood disorders, but predicting patient survival outcomes remains a complex challenge. Current predictive models often suffer from biases linked to race, socioeconomic status, and geographical disparities, leading to **inequitable survival predictions**. Furthermore, handling **censored time-to-event data** presents significant difficulties, as traditional models struggle to accurately capture patient survival patterns while maintaining fairness.

To address these challenges, this study leverages **SurvivalGAN**, a generative adversarial network (GAN) designed for survival data generation. SurvivalGAN effectively models censored time-to-event relationships, mitigating common failure modes such as **over-optimism, over-pessimism, and short-sightedness** in synthetic data. By integrating this approach with machine learning techniques like **Kaplan-Meier estimation, Cox Proportional Hazards models, and XGBoost**, we aim to develop a **fair and accurate predictive model** that ensures equitable transplant survival predictions across diverse patient groups.

This work not only enhances prediction accuracy but also promotes fairness in healthcare decision-making, ultimately contributing to improved patient outcomes and trust in AI-driven medical models.

1.2 OBJECTIVE

This study aims to develop an advanced predictive model for allogeneic HCT patient survival that is both **accurate and fair across diverse racial groups**. The key objectives include:

1. **Improving Survival Prediction Accuracy** – Implementing **SurvivalGAN** to generate high-quality synthetic survival data that preserves real-world statistical properties, enabling robust predictive modeling.
2. **Ensuring Fairness Across Racial Groups** – Utilizing the **Stratified Concordance Index** to evaluate and mitigate racial biases in survival predictions.
3. **Handling Censored Time-to-Event Data** – Addressing the complexities of **censoring and time-horizon imbalance** using GAN-based synthetic data augmentation.

4. **Enhancing Clinical Decision-Making** – Providing equitable survival predictions that can aid medical professionals in making informed transplant decisions, optimizing patient care.

By achieving these objectives, this research contributes to the **advancement of AI-driven survival analysis in healthcare**, ensuring **equitable and reliable** transplant outcome predictions.

1.3 EXISTING SYSTEM

Current predictive models for allogeneic Hematopoietic Cell Transplantation (HCT) survival rely on traditional statistical methods such as **Cox Proportional Hazards models and Kaplan-Meier estimation**. While these models provide valuable insights, they struggle with **handling censored time-to-event data** and often exhibit biases related to race, socioeconomic status, and geography. Furthermore, machine learning-based survival models often suffer from **data imbalances and fairness issues**, making them less reliable for diverse patient populations.

Additionally, existing synthetic data generation techniques for survival analysis do not effectively **capture censoring imbalance** or **time-horizon imbalance**, leading to inaccuracies in survival predictions. These shortcomings limit the ability to make **equitable and precise** survival predictions for HCT patients.

DRAWBACKS:

- Inability to handle censored data efficiently, leading to biased survival estimates.
- Traditional models do not address fairness concerns in transplant survival predictions.
- Synthetic data generation techniques fail to capture real-world event distributions.
- Existing models struggle with generalization across different demographic groups.

1.4 PROPOSED SYSTEM

To overcome these limitations, we propose a **SurvivalGAN-powered predictive model** for HCT survival analysis, ensuring both **accuracy and fairness** in transplant survival predictions. SurvivalGAN, a **Generative Adversarial Network (GAN) designed for survival data**, effectively addresses challenges in censored time-to-event data by capturing intricate survival patterns while maintaining fairness across different racial groups.

Our approach integrates **Kaplan-Meier estimation, Cox models, XGBoost, and deep learning techniques** to build a robust prediction model. To evaluate model performance, we utilize the **Stratified Concordance Index**, ensuring fair and consistent predictions across different racial subgroups. This method enhances **clinical decision-making, healthcare resource allocation, and personalized treatment plans** for transplant patients.

ADVANTAGES:

- Effectively handles **censored survival data** with GAN-based data augmentation.
- Improves **fairness** by ensuring balanced predictive performance across racial groups.
- Enhances **accuracy and generalization** by learning from synthetic yet realistic patient data.
- Supports **personalized medicine**, improving clinical decision-making for HCT patients.

1.5 HARDWARE REQUIREMENTS

The hardware requirements ensure the system operates efficiently for data processing, survival modeling, and synthetic data generation.

- **PROCESSOR:** Intel Core i5/i7 or equivalent
- **RAM:** 8GB DDR4 (minimum)
- **HARD DISK:** 500GB SSD (recommended for faster computation)
- **GPU:** NVIDIA CUDA-enabled GPU (for deep learning models)

1.6 SOFTWARE REQUIREMENTS

The software requirements define the tools and technologies needed for model development and deployment.

- **FRONT END:** Python (Streamlit/Dash for visualization)
- **BACK END:** Python (Pandas, NumPy, SciPy, TensorFlow/PyTorch, Lifelines, XGBoost)
- **DATABASE:** PostgreSQL / MySQL (for storing patient data)
- **OPERATING SYSTEM:** Windows 10 / Linux (Ubuntu)
- **IDE:** Jupyter Notebook, VS Code

1.7 LITERATURE SURVEY

This literature survey examines various machine learning and survival analysis techniques for predicting survival in different medical contexts.

1. **Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction:**

   This study investigates breast cancer progression prediction using survival analysis models. It applies the Cox Proportional Hazards (PH) model, Random Survival Forest (RSF), and Conditional Inference Forest (Cforest) on two breast cancer datasets: GBSG2 and METABRIC. Model performances were evaluated using the Concordance Index (C-Index) and Prediction Error Curves (PEC). RSF and Cforest showed better predictive accuracy than the Cox PH model.

2. **Survival Prediction of Children After Bone Marrow Transplant Using Machine Learning Algorithms:**

   This study applies machine learning algorithms to predict survival after bone marrow transplant (BMT) using a dataset from the UCI ML repository. It evaluates models such as Random Forest (RF), Bagging Classifier, XGBoost, AdaBoost, Gradient Boost (GB), Decision Tree (DT), and K-Nearest Neighbors (KNN). The best models achieved 97.37% accuracy after feature selection and hyperparameter tuning with Grid Search Cross-Validation (GSCV).

3. **Machine Learning for the Prediction of Survival Post-Allogeneic Hematopoietic Cell Transplantation: A Single-Center Experience:**

   This study examines the effectiveness of machine learning models in predicting survival after allogeneic hematopoietic cell transplantation (HCT). Using a dataset of 2,697 patients, ML models were trained on 45 pre-transplant variables. Random Forest (RF) achieved the best performance (AUC 0.71) compared to logistic regression. Key survival factors included donor type, radiation dose, patient age, and lung function metrics.

4. **The Ensembles of Machine Learning Methods for Survival Prediction After Kidney Transplantation:**

   This study applies ensemble machine learning techniques for survival prediction after kidney transplantation. The Kaplan-Meier method is used for survival estimation, and multiple

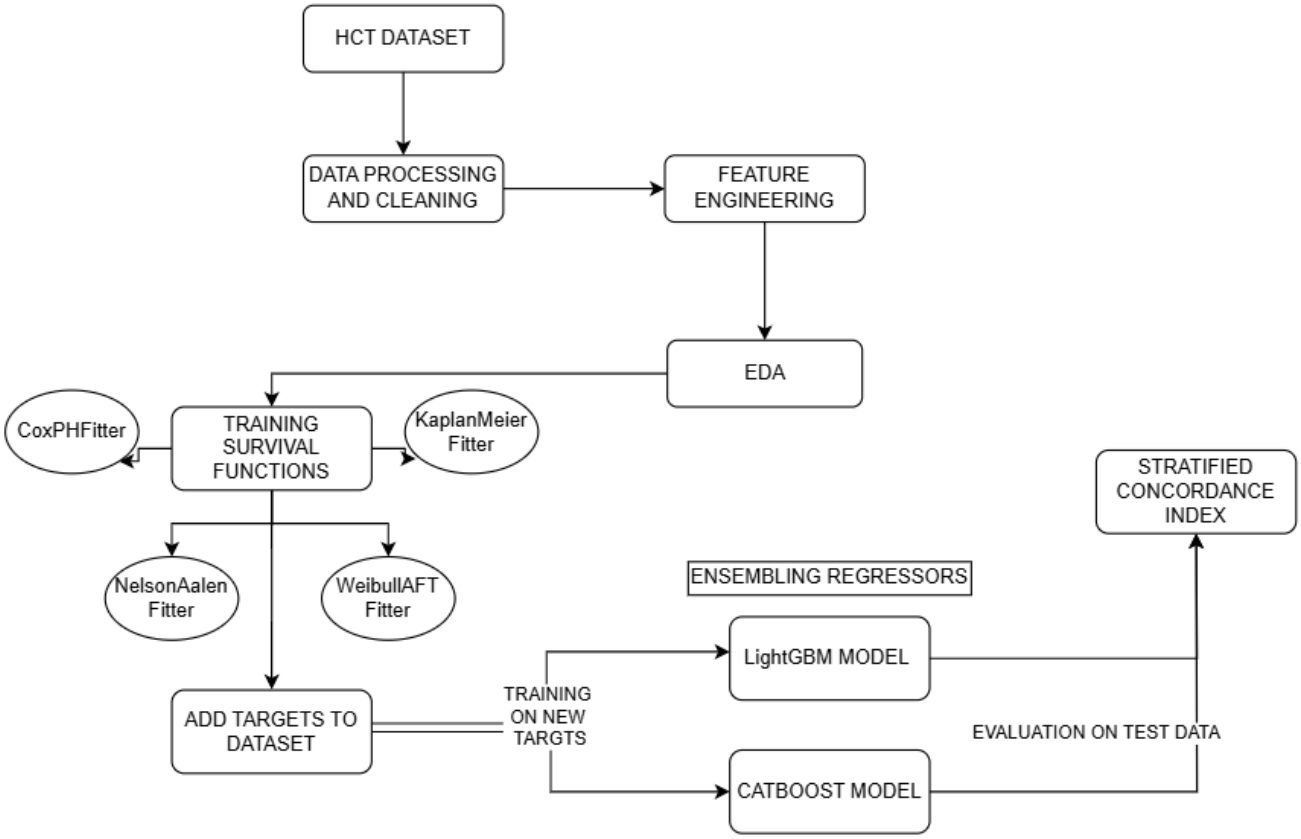
feature selection methods are implemented to refine predictive accuracy. Four ensemble ML models were developed, achieving over 90% classification accuracy using a stacking approach.

5. **Bayesian Weibull Tree Models for Survival Analysis of Clinico-Genomic Data:** This study presents a Bayesian Weibull tree model for survival prediction using clinico-genomic data. The model applies recursive partitioning to group patients into homogeneous subgroups, each following a Weibull survival distribution. Empirical Bayes methods dynamically update prior distributions, and predictions are made by averaging multiple models. The approach was validated on ovarian cancer survival data and identified genomic biomarkers affecting survival outcomes.

1.8 LITERATURE SURVEY TABLE

Author(s)	Title	Year	Concept	Merit	Demerit
Gonca Buyrukoğlu	Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction	2024	Evaluates Cox PH, Random Survival Forest, and Conditional Inference Forest for breast cancer survival prediction.	RSF and Cforest outperform Cox PH, improving predictive accuracy.	Limited to two specific datasets.
Hussam Alawneh, Ahmad Hasasneh	Survival Prediction of Children After Bone Marrow Transplant Using Machine Learning Algorithms	2024	Uses ML models (RF, XGBoost, AdaBoost, etc.) to predict pediatric bone marrow transplant survival.	Achieves 97.37% accuracy using feature selection and hyperparameter tuning.	Does not account for time-dependent covariates.
Hamed Shourabizadeh, Dionne M. Aleman, Louis-Martin Rousseau, Arjun D. Law, Auro Viswabandya, Fotios V. Michelis	Machine Learning for the Prediction of Survival Post-Allogeneic Hematopoietic Cell Transplantation: A Single-Center Experience	2023	Investigates ML models for HCT survival prediction using 2,697 patient records. RF achieved the best AUC of 0.71.	Identifies significant clinical predictors for survival stratification.	Limited to a single-center dataset, requiring external validation.
Yaroslav Tolstyak et al.	The Ensembles of Machine Learning Methods for Survival Predicting After Kidney Transplantation	2021	Applies ensemble ML models and Kaplan-Meier estimation to predict kidney transplant survival.	Uses multiple feature selection techniques for better predictive accuracy.	Limited dataset generalizability, requiring further validation.
Jennifer Clarke, Mike West	Bayesian Weibull Tree Models for Survival Analysis of Clinico-Genomic Data	2007	Uses Bayesian Weibull tree models for survival prediction via recursive partitioning.	Integrates genomic and clinical data effectively for personalized predictions.	Computationally intensive and requires domain expertise.

1.9 DESIGN METHODOLOGY



This diagram represents a **machine learning pipeline for survival analysis** using the **HCT (Hematopoietic Cell Transplantation) dataset**. The methodology follows a structured approach involving data preprocessing, survival function modeling, feature engineering, and ensemble regression models for improved survival prediction. Here's a breakdown of each step:

1. Data Acquisition and Preprocessing

- **HCT Dataset**: The pipeline begins with the dataset containing survival-related information.
- **Data Processing and Cleaning**: Includes handling missing values, outliers, and normalizing data to ensure consistency.

- **Feature Engineering**: Extracts relevant features to enhance predictive performance.
- **Exploratory Data Analysis (EDA)**: Analyzes feature distributions and relationships.

2. Survival Function Modeling

- **Training Survival Functions**: Different survival analysis models are trained to estimate survival probabilities:
    - **CoxPHFitter (Cox Proportional Hazards Model)**: Estimates hazard ratios for covariates.
    - **KaplanMeierFitter**: Computes non-parametric survival estimates.
    - **NelsonAalenFitter**: Estimates cumulative hazard functions.
    - **WeibullAFTFitter (Accelerated Failure Time Model)**: Fits parametric survival curves.
- **Adding Targets to Dataset**: The trained survival models generate survival probabilities and risk scores, which are added as new target variables.

3. Training Machine Learning Models

- **Ensembling Regressors**: Machine learning models use survival predictions as inputs:
    - **LightGBM Model**: A gradient boosting model that efficiently handles large datasets.
    - **CatBoost Model**: A high-performance gradient boosting algorithm optimized for categorical data.
- **Training on New Targets**: The ensemble models are trained using survival targets obtained from previous steps.

4. Model Evaluation

- **Evaluation on Test Data**: The trained models predict survival probabilities for unseen data.
- **Stratified Concordance Index**: Measures the agreement between predicted survival rankings and actual outcomes.

SUMMARY

- The methodology integrates **traditional survival analysis techniques** (e.g., CoxPH, Kaplan-Meier) with **modern ML models** (LightGBM, CatBoost).
- The **ensemble learning approach** improves prediction accuracy.
- The **Stratified Concordance Index** ensures robust model evaluation.