EQUITY IN HCT SURVIVAL PREDICTIONS

Mohammed Abdul Kalam Khan,
U.G. Student, Department of Computer and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

Dr. Rajitha Kotoju,
Asssistant Professor, Department of Computer and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

R.Mohan Krishna Ayyappa,
Asssistant Professor, Department of Computer and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

ABSTRACT:
Hematopoietic Cell Transplantation (HCT) remains a critical therapeutic option for various hematological conditions, yet predicting post-transplant survival outcomes remains a complex challenge—particularly across diverse racial and socioeconomic groups. This project proposes a machine learning–driven system that enhances survival prediction for allogeneic HCT patients by addressing existing disparities related to race, geography, and socioeconomic status. Leveraging ensemble models such as XGBoost, CatBoost, and LightGBM, the system integrates clinical and demographic data to estimate personalized survival risk scores. A user-friendly interface is developed using Next.js (frontend) and Flask (backend) to enable clinicians and researchers to interact with the model and obtain interpretable predictions. For fair and accurate model evaluation, we employ the Stratified Concordance Index (C-index), a metric specifically adapted to assess predictive performance across racial groups independently. This metric not only evaluates the model's ability to rank survival times reliably but also penalizes variability across racial subgroups, promoting equitable healthcare outcomes. By incorporating both algorithmic sophistication and a focus on fairness, this project aims to assist clinical decision-making, improve patient stratification, and foster greater trust in predictive healthcare systems.

Keywords: Hematopoietic Cell Transplantation, Survival Prediction, Machine Learning, LightGBM, Stratified Concordance Index, Fairness in AI, Flask, Next.js, Racial Equity in Healthcare

INTRODUCTION

Hematopoietic Cell Transplantation (HCT) is a critical therapeutic intervention used to treat a range of hematologic malignancies and disorders, including leukemia, lymphoma, and bone marrow failure. Despite its life-saving potential, predicting post-transplant survival outcomes remains a complex and high-stakes challenge, influenced by a multitude of clinical, biological, and treatment-related factors. Traditional scoring systems often fall short in capturing this complexity—particularly as treatment protocols evolve and new risk factors emerge.

To address this gap, our project develops an ensemble-based machine learning system designed to enhance the accuracy and fairness of survival predictions after allogeneic HCT. By integrating gradient boosting methods such as XGBoost, LightGBM, and CatBoost, the system is capable of learning from diverse clinical datasets that include both numerical and categorical variables. Advanced preprocessing techniques and HLA feature engineering are employed to optimize data quality and model performance.

Importantly, the system accounts for disparities across socioeconomic, racial, and geographic lines—factors that are frequently underrepresented in conventional models. Using a fairness-aware evaluation strategy based on the Stratified Concordance Index (Cindex), the system ensures equitable prediction performance across different racial subgroups. This makes the model not only more accurate but also more just and clinically relevant.

The final model outputs a personalized risk score (0–100) for each patient, estimating their post-transplant survival probability. This robust predictive pipeline is deployed through a user-friendly web interface built with Next.js (frontend) and Flask (backend), enabling clinicians to input patient data and receive real-time, interpretable risk assessments. Overall, this project demonstrates how machine learning can be harnessed to support equitable, datadriven decisions in complex clinical settings.

LITERATURE SURVEY

The [1] Gonca Buyrukoğlu(2024), “Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction” This study invesΘgates survival predicΘon in breast cancer paΘents using ensemble learning and tradiΘonal survival analysis models. It compares the Cox ProporΘonal Hazards model with Random Survival Forest (RSF) and CondiΘonal Inference Forest (Cforest) on two datasets—GBSG2 and METABRIC. EvaluaΘon was done using the Concordance Index (C-Index) and PredicΘon Error Curves (PEC), showing that RSF and Cforest outperformed the Cox PH model in predicΘve accuracy.

[2] Hussam Alawneh, Ahmad Hasasneh (2024), “Survival Prediction of Children After Bone Marrow Transplant Using Machine Learning Algorithms” This research applies machine learning algorithms to predict survival outcomes of children after bone marrow transplantation using data from the UCI Machine Learning Repository. Models including Random Forest, XGBoost, AdaBoost, Bagging Classifier, Gradient Boosting, Decision Tree, and K-Nearest Neighbors were evaluated. After feature selection and hyperparameter tuning with Grid Search Cross-Validation, the best-performing models achieved an accuracy of 97.37%.

[3] Hamed Shourabizadeh, Dionne M. Aleman (2023), “Machine Learning for the Prediction of Survival Post-Allogeneic Hematopoietic Cell Transplantation: A SingleCenter Experience” This study explores the use of machine learning for predicting survival after allogeneic hematopoietic cell transplantation (HCT). Using a dataset of 2,697 patients and 45 pretransplant clinical variables, various ML models were trained. Random Forest achieved the highest performance with an AUC of 0.71, outperforming traditional logistic regression. Key predictive factors included donor type, radiation dosage, patient age, and lung function parameters.

[4] Yaroslav Tolstyak et al. (2021), “The Ensembles of Machine Learning Methods for Survival Prediction After Kidney Transplantation” This research implements ensemble machine learning models for predicting survival after kidney transplantation. The study uses the Kaplan-Meier method for survival estimation and incorporates multiple feature selection techniques. Four ensemble models were developed and tested using a stacking approach, achieving classification accuracy of over 90%, demonstrating the robustness of ensemble learning in clinical survival prediction.

[5] Jennifer Clarke, Mike West (2007), “Bayesian Weibull Tree Models for Survival Analysis of Clinico-Genomic Data” This study proposes a Bayesian Weibull tree-based survival model to analyze clinicogenomic data. The model uses recursive partitioning to segment patients into subgroups that follow Weibull survival distributions. By applying empirical Bayes methods to update prior distributions, the approach achieves improved predictions

by averaging over multiple tree models. The model was validated using ovarian cancer data, revealing key genomic biomarkers affecting patient survival.

METHODOLOGY

To address the complex task of predicting post-transplant survival for Hematopoietic Cell Transplantation (HCT) patients, this project developed a scalable and interpretable ensemble machine learning system. The system integrates traditional survival analysis methods with modern machine learning algorithms, offering personalized and fair risk assessments via a real-time clinical web application.

Technologies Used:

- Python (Jupyter Notebook): Primary development environment due to its extensive ecosystem of scienΘfic compuΘng and machine learning libraries.
- Pandas, NumPy, Scikit-learn: Used for preprocessing clinical data, handling missing values, encoding categorical features, and performing model evaluaΘons.
- Lifelines: Utilized to compute survival targets using classical survival analysis models such as:

o Cox ProporΘonal Hazards (CoxPH)
o Kaplan-Meier EsΘmator
o Nelson-Aalen EsΘmator
o Weibull FiΣer
o Event-Free Survival (EFS)

- LightGBM and CatBoost: These gradient boosΘng models were selected for their superior performance with structured and categorical data, forming the foundaΘon of the ensemble.
- Flask (Backend) & Next.js (Frontend): Combined to build the web-based interface allowing clinicians to interact with the model.
- Plotly: Employed for exploratory data analysis (EDA) and visualization of survival curves and feature distributions.

Design:

The design of the Hematopoietic Cell Transplantation (HCT) Survival Prediction System defines a comprehensive framework for processing complex clinical data, training survival models, and delivering interpretable, real-time predictions to healthcare professionals. It ensures seamless integration between data ingestion, ensemble-based machine learning, and web deployment, culminating in a robust, scalable, and equitable decision-support system for transplant clinicians.

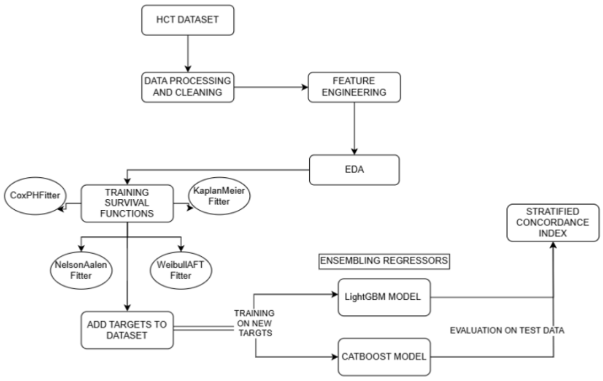


Figure 1: Design of HCT Survival Prediction

The system architecture, shown in Figure 4.1 (System Architecture Diagram), is composed of seven core components: HCT Dataset, data processing and cleaning, feature engineering, training survival functions, ensembling regressors, evaluation with stratified C-index. Each layer is optimized for accuracy, usability, and fairness, particularly for underrepresented clinical subgroups.

1. Data Input & Preprocessing

- **HCT Dataset:** The system ingests patient data containing total 54 variables-categorical (e.g., donor type, conditioning regimen) and numerical features (e.g., age, lung function scores).
- **HLA Feature Engineering:** Specialized transformation and encoding are applied to Human Leukocyte Antigen (HLA) fields to capture compatibility effects.
- **Cleaning & Encoding:** Missing values are handled using domain-specific imputation, and categorical features are encoded using target encoding or one-hot encoding where appropriate.
- **Normalization:** Continuous variables are scaled as needed to improve model convergence and interpretability.

2. Survival Modeling Engine

- **Survival Targets:** Five survival analysis targets are derived from:
  - Cox Proportional Hazards model (CoxPH)
  - Kaplan-Meier estimator
  - Nelson-Aalen estimator
  - Weibull parametric survival model
  - Event-Free Survival (EFS) metrics
- **Ensemble Learning:** For each target, two tree-based models are trained: **LightGBM** and **CatBoost**, yielding 10 models total.
- **Cross-Validation:**Each model undergoes **5-fold cross-validation** to ensure generalizability and reduce overfitting.
- **Model Evaluation:**Performance is assessed using **concordance index (C-index)**, with fairness audits applied across demographic strata.

3. Prediction & Risk Scoring System

- **Risk Aggregation:** The predictions from all 10 models are combined using a weighted average (with learned or manually tuned weights).
- **Score Normalization:** The final prediction is mapped to a **personalized survival risk score (0–100)**, where higher scores indicate higher survival probability.

4. Web-Based Interface

- **Frontend (Next.js):** A responsive, user-friendly interface allows clinicians to input patient data through dynamic forms.
- **Backend (Flask):** Provides RESTful endpoints for frontend interaction and security control.
- **Security & Validation:** Input validation, access control, and logging mechanisms are implemented to ensure clinical safety and compliance.

Process Flow Diagram

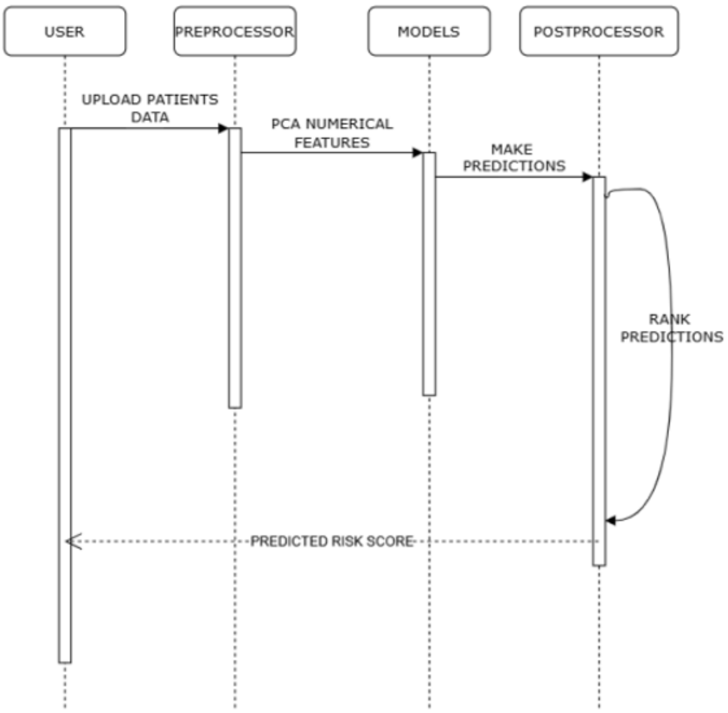


Figure 2: Process flow Diagram of HCT Survival Prediction

The flowchart (shown in **Figure 4.2**) outlines the end-to-end workflow across four core components: **User**, **Preprocessor**, **Models**, and **Postprocessor**. It captures how raw clinical data is transformed into meaningful survival risk predictions through machine learning.

Implementation:

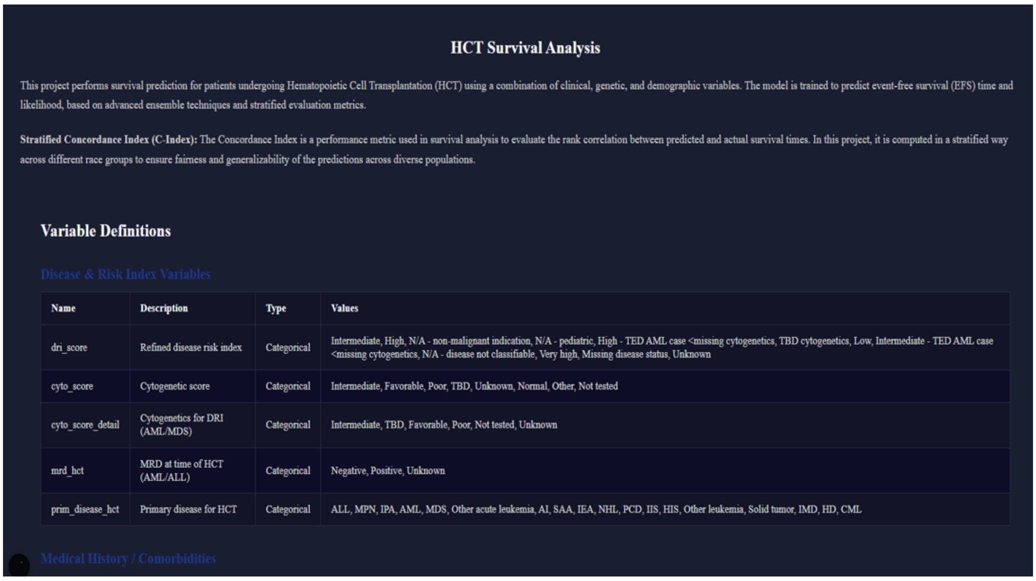
Development proceeded through modular implementation:

- **Data Cleaning & Processing:** Raw clinical data was transformed into structured formats, missing values imputed, and categorical variables encoded.
- **Survival Target Generation:** Time-to-event data was processed into 5 distinct survival targets using the Lifelines package.
- **Model Training:** Each of the 10 models (2 per survival target) was trained and tuned via 5-fold cross-validation.
- Model Ensembling: Predictions were aggregated into a single survival risk score ranging from 0 to 100 for each patient.
- **Web Interface Integration:** o Flask served the model and handled API requests. A React-based frontend (Next.js) allowed clinicians to input patient details and receive survival predictions.

RESULTS

Home Page Overview

1.Variable Definitions Table: The dashboard home page begins with a searchable, scrollable table that lists and defines every clinical variable used in the prediction system. This includes the feature name, a clinical description, value type (categorical/numerical), and sample values to aid understanding.



HCT Survival Analysis

This project performs survival prediction for patients undergoing Hematopoietic Cell Transplantation (HCT) using a combination of clinical, genetic, and demographic variables. The model is trained to predict event-free survival (EFS) time and likelihood, based on advanced ensemble techniques and stratified evaluation metrics.

Stratified Concordance Index (C-Index): The Concordance Index is a performance metric used in survival analysis to evaluate the rank correlation between predicted and actual survival times. In this project, it is computed in a stratified way across different race groups to ensure fairness and generalizability of the predictions across diverse populations.

Variable Definitions

Disease & Risk Index Variables

Name	Description	Type	Values
dri_score	Refined disease risk index	Categorical	Intermediate, High, N/A - non-malignant indication, N/A - pediatric, High - TED AML case <missing cytogenetics, TBD cytogenetics, Low, Intermediate - TED AML case <missing cytogenetics, N/A - disease not classifiable, Very high, Missing disease status, Unknown
cyto_score	Cytogenetic score	Categorical	Intermediate, Favorable, Poor, TBD, Unknown, Normal, Other, Not tested
cyto_score_detail	Cytogenetics for DRI (AML/MDS)	Categorical	Intermediate, TBD, Favorable, Poor, Not tested, Unknown
mrd_hct	MRD at time of HCT (AML/ALL)	Categorical	Negative, Positive, Unknown
prim_disease_hct	Primary disease for HCT	Categorical	ALL, MPN, IPA, AML, MDS, Other acute leukemia, AI, SAA, IEA, NHL, PCD, IIS, HIS, Other leukemia, Solid tumor, IMD, HD, CML

Medical History / Comorbidities

Figure 3: Variable Definitions Table interface

2. Data Distribution Graphs: Users can visualize the distribution of each variable across the dataset using dynamic histograms, boxplots, and count plots. This helps identify patterns such as age skewness, donor type frequency, and pulmonary function score distribution across patients.

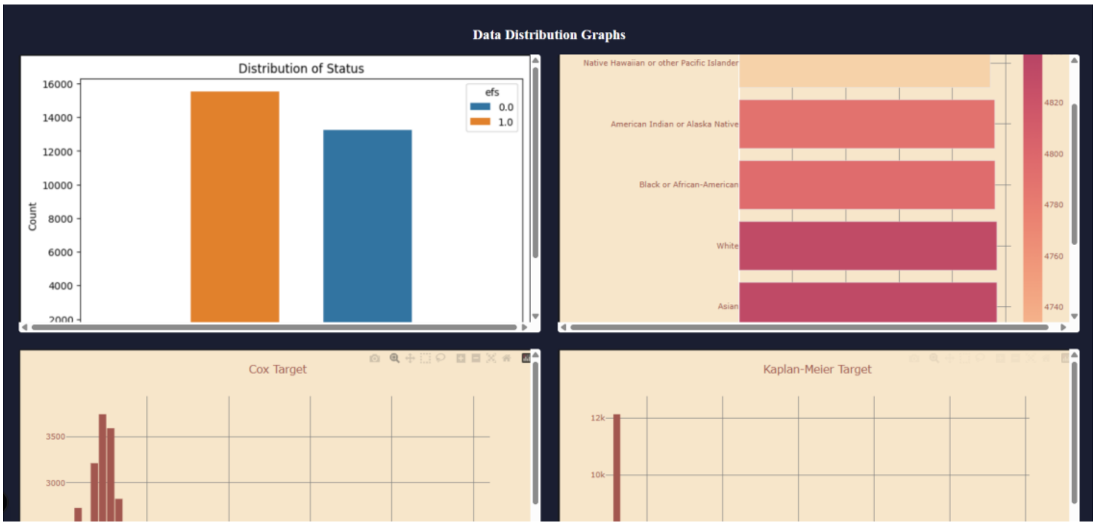


Figure 4: Interactive graphs showing variable-wise data distributions

3. Model Performance Graphs: The system displays performance metrics (e.g., Concordance Index, AUC-ROC, RMSE) for each of the 10 models (5 survival targets × 2 ML models). Bar charts and comparative plots show model effectiveness across different survival estimation approaches.

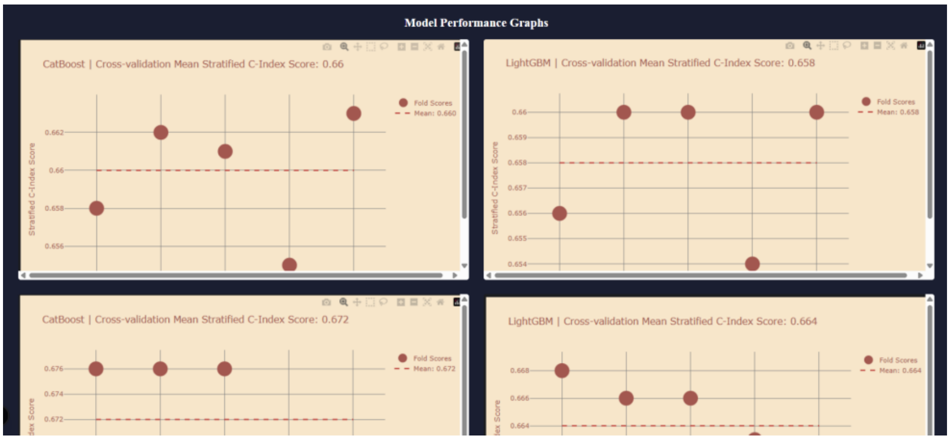


Figure 5: Model performance visualization comparing CatBoost, LightGBM etc.

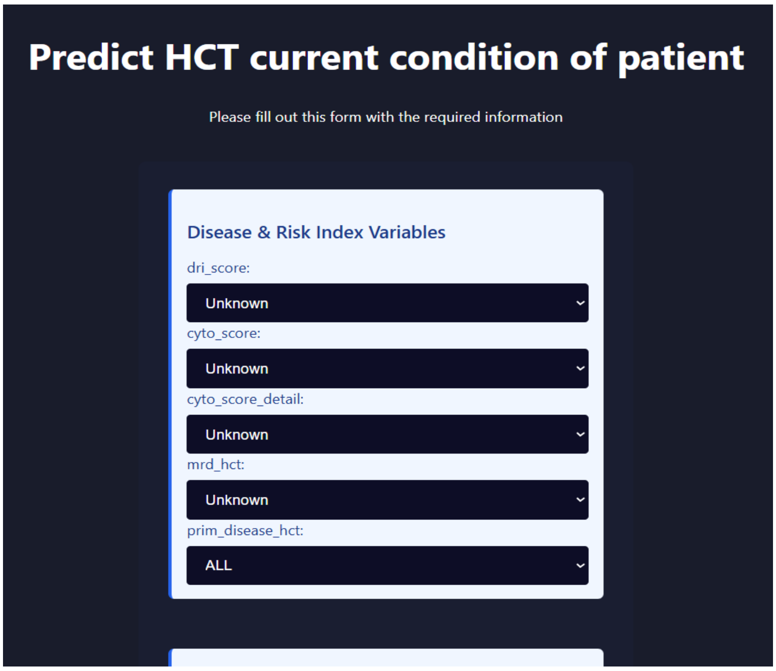
Predict Page Overview

1.Clinical Data Input Form

The Predict page features a well-organized, multi-section form where clinicians input a patient's clinical data. Variables are grouped under relevant headings for easy navigation and logical flow. The sections include:

- **Disease & Risk Index Variables**
- **Medical History / Comorbidities**
- **HLA Matching (High and Low Resolution)**
- **Transplant Conditioning / Treatment**
- **Graft & Donor Info**
- **Clinical / Functional Status**
- **Demographics**

Each field includes tooltips and input validation to minimize errors and guide users.



Predict HCT current condition of patient

Please fill out this form with the required information

Disease & Risk Index Variables

dri_score: Unknown

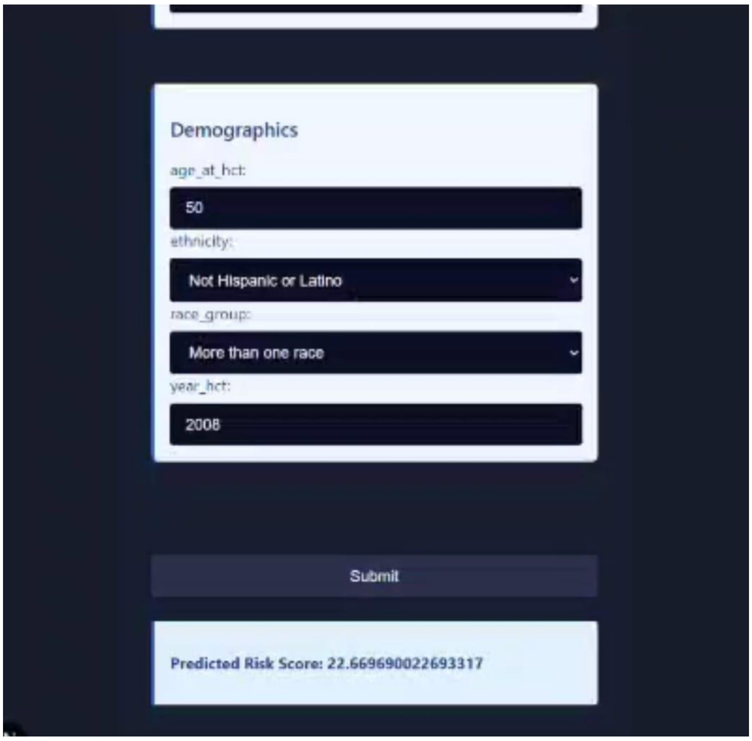
cyto_score: Unknown

cyto_score_detail: Unknown

mrd_hct: Unknown

prim_disease_hct: ALL

Figure 6: Screenshot of the clinical form with collapsible sections



Demographics

age_at_hct: 50

ethnicity: Not Hispanic or Latino

race_group: More than one race

year_hct: 2008

Predicted Risk Score: 22.669690022693317

Figure 7: Risk Score display with confidence indicators and optional breakdown

2. **Real-Time Risk Prediction Output**

After submission of the form shown in Figure 6 and 7, the backend ensemble model processes the inputs and returns a personalized survival risk score (0–100). This score estimates the probability of post-transplant survival and is presented using a color-coded bar (green to red) along with model interpretation notes.

DISCUSSION

This study introduces an ensemble-based machine learning framework for predicting post-transplant survival in Hematopoietic Cell Transplantation (HCT) patients. By combining classical survival analysis methods with gradient boosting models—LightGBM and CatBoost—the system delivers accurate, personalized survival predictions through a real-time clinical web interface.

A key contribution of this work is the integration of five survival estimators (CoxPH, Kaplan-Meier, Nelson-Aalen, Weibull, and Event-Free Survival) into an ensemble of ten models. This approach captures diverse aspects of survival risk and improves robustness through stratified 5-fold cross-validation. Predictions are aggregated into an interpretable 0–100 survival risk score, aiding clinician decision-making.

The system's modular design ensures scalability and usability. Preprocessing steps, such as domain-specific imputation and HLA feature engineering, significantly enhanced model performance. Additionally, fairness evaluations across demographic groups help mitigate bias—a critical factor in clinical applications.

Compared to prior literature, this work stands out by blending interpretability, predictive power, and deployability in a clinician-facing tool. However, limitations include reduced transparency from ensembling and the need for external validation. Future work will focus on model explainability and broader dataset integration.This project presents a practical, interpretable, and equitable solution for HCT survival prediction, bridging the gap between advanced analytics and real-world clinical utility.

CONCLUSION

We have implemented the Hematopoietic Cell Transplantation (HCT) Survival Prediction System that presents a significant advancement in the clinical decision-making landscape by offering a personalized, data-driven approach to post-transplant risk assessment. By leveraging an ensemble of ten machine learning models—built upon five distinct survival estimation frameworks and two gradient boosting algorithms (LightGBM and CatBoost)—the system accounts for the complexity and high dimensionality inherent in HCT clinical data. Through robust cross-validation, HLA-specific feature engineering, and stratified learning, the model ensures predictive accuracy while maintaining fairness across diverse patient subgroups.

Integrated into a streamlined web-based platform using a Flask backend and a Next.js frontend, the system delivers interpretable survival scores in real time, enhancing clinicians' ability to make informed, individualized treatment decisions. This approach transcends traditional survival scoring systems by embracing modern machine learning, thereby reflecting the evolving nature of clinical data and therapeutic strategies. Overall, the project illustrates the transformative potential of machine learning in complex medical domains and paves the way for smarter, fairer, and more effective transplant care.

REFERENCES

[1] Ahmad et al. (2024), *Survival Analysis in Breast Cancer: Evaluating Ensemble Learning Techniques for Prediction*.
DOI: http://doi.org/10.7717/peerj-cs.2147

[2] Altaf et al. (2024), *Survival Prediction of Children After Bone Marrow Transplant Using Machine Learning Algorithms*.
DOI: https://doi.org/10.34028/iajit/21/3/4

[3] Zhou et al. (2024), *Longitudinal clinical data improve survival prediction after hematopoietic cell transplantation using machine learning*.
DOI: https://doi.org/10.1182/bloodadvances.2023011752

[4] Woźniacki et al. (2024), *A Novel Approach for Predicting the Survival of Colorectal Cancer Patients Using Machine Learning Techniques and Advanced Parameter Optimization Methods*.
DOI: https://doi.org/10.3390/cancers16183205

[5] Alabdallah et al. (2024), *The Concordance Index decomposition: A measure for a deeper understanding of survival prediction models*.
DOI: https://doi.org/10.1016/j.artmed.2024.102781

[6] Haider et al. (2024), *The Algorithmic Divide: A Systematic Review on AI-Driven Racial Disparities in Healthcare*.
DOI: https://doi.org/10.1007/s40615-024-02237-0

[7] Shourabizadeh et al. (2023), *Machine Learning for the Prediction of Survival Post-Allogeneic Hematopoietic Cell Transplantation: A Single-Center Experience*.
DOI: https://doi.org/10.1159/000533665

[8] Dai et al. (2023), *A Flexible Ensemble Learning Method for Survival Extrapolation*.
DOI: https://doi.org/10.1007/s43441-022-00490-1

[9] Blue et al. (2023), *Racial and Socioeconomic Disparities in Long-Term Outcomes in ≥1 Year Allogeneic Hematopoietic Cell Transplantation Survivors: A CIBMTR Analysis*.
DOI: https://doi.org/10.1016/j.jtct.2023.07.013

[10] Ueda et al. (2023), *Fairness of artificial intelligence in healthcare: review and recommendations*.
DOI: https://doi.org/10.1007/s11604-023-01474-3

[11] Choi et al. (2022), *Predicting Long-term Survival After Allogeneic Hematopoietic Cell Transplantation in Patients With Hematologic Malignancies: Machine Learning-Based Model Development and Validation*.
DOI: https://doi.org/10.2196/32313

[12] Zhang et al. (2021), *The Ensembles of Machine Learning Methods for Survival Prediction After Kidney Transplantation*.
DOI: https://doi.org/10.3390/app112110380

[13] Faraggi and Simon (2008), *Bayesian Weibull Tree Models for Survival Analysis of Clinico-Genomic Data*.
DOI: https://doi.org/10.1016/j.stamet.2007.09.003