Pre-collection Prediction of Human Leukocyte Antigen Rarity in Cord Blood Banking: A Comparative Evaluation of Machine Learning Models

Amir Hossein Esmaielpour[a], Mariam Ameli[b*], Ashkan Mozdgir[c], Orod Ahmadi[d], Morteza Zarrabi[e]

a. Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, No. 43, Shahid Mofatteh Ave, Tehran, Iran, amir.esmaielpour.@khu.ac.ir

b. Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, No. 43, Shahid Mofatteh Ave, Tehran, Iran, m.ameli@khu.ac.ir

c. Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, No. 43, Shahid Mofatteh Ave, Tehran, Iran, a.mozdgir@khu.ac.ir

d. Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, No. 43, Shahid Mofatteh Ave, Tehran, Iran, orod.ahmadi@khu.ac.ir

e. Department of Stem Cells and Developmental Biology, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran, m.zarrabi@rsct.ir

Abstract

Human leukocyte antigen (HLA) compatibility is critical for the success of hematopoietic stem cell transplantation, particularly for patients without matched related donors. As cord blood units (CBUs) play a growing role in bridging donor gaps, cord blood banks must optimize donor selection and CBUs' storage to maximize HLA diversity. A prenatal, pre-collection prediction framework is developed to estimate whether a CBU's future HLA profile will be rare or common within a large banking registry. Rarity was defined as a binary label based on observed 6/6 low-resolution matches at the HLA-A, HLA-B, and HLA-DRB1 loci (0 matches = rare; ≥1 = common). Demographic, medical, and obstetric features were preprocessed via one-hot encoding, degree-2 polynomial expansion, and Principal Component Analysis (PCA); feature selection used Random Forest (RF) importance within a leakage-controlled pipeline. Five machine learning models, including K-Nearest Neighbors (KNN), RF, XGBoost, and a

* Corresponding author.
E-mail address: m.ameli@khu.ac.ir (M. Ameli)
Tel: +98-21-86072718
Fax: +98-21-88820530

deep multilayer perceptron (MLP), are used, with each evaluated with and without Genetic Algorithm (GA) hyperparameter tuning on identical cross-validation folds, and a TOPSIS multi-criteria ranking was performed to rank algorithms. The GA-tuned MLP performs best (Accuracy 0.74, ROC-AUC 0.84) and ranks first by Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), with GA-tuned XGBoost close behind. These results demonstrate that prenatal, pre-collection prediction of HLA rarity is feasible and can support prioritization for confirmatory genotyping to enhance diversity in cord-blood inventories.

Keywords

HLA Rarity Prediction, Cord blood banking, Genetic Algorithm, hematopoietic Stem cell transplantation, Deep learning

1. Introduction

Every year, hematologic cancers commonly treated with hematopoietic stem cell transplantation (HSCT) account for ~700,000 deaths worldwide [1]. HSCT remains a potentially curative therapy for many of these diseases. It is performed as autologous transplantation or allogeneic transplantation, with the graft source chosen based on urgency, disease, and donor availability [2]. Cord blood is particularly attractive because it is rapidly available, tolerates greater HLA mismatch, and has a comparatively lower risk of severe graft-versus-host disease, making it valuable when a timely matched donor cannot be secured [3]. Cord blood banks play a vital role in storing hematopoietic stem cells for future transplantation, especially when matched donors are unavailable. However, they face challenges in inventory management, such as high storage costs, limited unit volume, and inefficient prioritization of enhance HLA diversity [4].

Due to the extreme polymorphism of HLA-A, -B, and -DRB1, achieving a 6/6 or above matching often requires very large, ethnically diverse cord blood inventories [5]. Modeling studies suggest tens of thousands of units are necessary—for example, a Korean registry analysis found that 100,000 units yield ~95% of patients a 5/6 match at high resolution [6]. Traditional methods to identify and reserve rare UCB units rely on static frequency tables and manual curation—

approaches that are neither scalable nor able to leverage the rich demographic metadata associated with each unit [7].

Against this backdrop, predicting HLA rarity prenatally can materially strengthen public cord-blood banking. Banks operate under tight economic and storage constraints; larger, more diverse inventories improve match likelihoods but are costly to build and maintain [8]. A model that flags units likely to carry rare HLA profiles before birth can guide targeted collection, prioritize long-term storage of genetically valuable units, and help close match gaps for under-represented groups—improving both efficiency and equity in transplant access.

Recent advances in machine learning (ML) and deep learning (DL) have shown promise for HLA imputation from SNP arrays, prediction of transplant outcomes, and optimization of donor–recipient matching algorithms [9-12]. However, prior ML/DL work has primarily targeted genotype imputation or matching rather than the pre-collection classification of HLA rarity within cord-blood registries.

Background

Beyond molecular features, HLA profiles are shaped by demographic, familial, and medical history factors. Individuals inherit one haplotype from each parent—siblings thus have a 25% chance of a full match. Maternal–fetal HLA compatibility (e.g., at HLA-B/HLA-C) has been linked to increased fetal loss [13, 14], and elevated placental HLA-G expression correlates with higher rates of preterm birth [15]. Non-inherited maternal antigens also influence autoimmune risk, as shown in case–control studies of Type 1 diabetes [16]. Regionally isolated populations exhibit distinct HLA distributions due to founder effects and population history [17].

To maintain cost-effectiveness, cord-blood banks (CBBs) apply pre-screening criteria such as total nucleated cell (TNC) counts and $CD34^+$ cell counts to predict unit usability. Machine-learning and deep-learning models have been used to forecast these quality metrics, enabling early triage of low-potential units [18, 19]. Similarly, Haghbayan et al. applied XGBoost and Light Gradient Boosting Machine (LightGBM) models to Royan Cord Blood Bank data, successfully predicting TNC and $CD34^+$ based on delivery and birth parameters [20]. Combined with economic analyses, this work has informed revised minimum thresholds for storage to avoid banking underused grafts [21].

A central component of transplant optimization is understanding HLA allele and haplotype frequency across populations. Statistical tools such as the expectation-maximization (EM) algorithm are widely used to estimate these frequencies, guiding inventory sizing and ensuring diversity in donor pools [22]. In parallel, ML has gained traction in addressing key challenges in donor–recipient selection and cord blood resource management. ML models have been used to predict donor availability based on registry behavior, classify graft usability and impute missing genetic markers [23, 24]. For example, one study developed a predictive model to identify donors likely to respond at time of need, while others used neural networks and RF to forecast UCB unit quality from maternal and processing features [25, 26]. ML has thus supported decisions in donor prioritization, graft triage, and inventory efficiency [27]. To our knowledge, no prior study has modeled—at the pre-collection stage—whether a cord-blood unit's HLA profile will be rare or common within a banking registry, a gap with important implications for early donor selection and intelligent inventory design.

In this paper, our goal is to identify and flag units carrying rare HLA profiles before they undergo costly and time-consuming high-resolution typing. This early identification aims to guide more strategic decisions for donor collection and optimize inventory management. This study analyzes a large national dataset of 122,000 cord-blood records from the Royan Cord Blood Bank. HLA rarity is defined as a binary outcome—rare vs. common—based on whether a unit has any observed 6/6 low-resolution matches at the HLA-A, -B, and -DRB1 loci (common if ≥1 match; rare if none). A pre-collection prediction framework that uses random-forest–based feature selection, degree-2 polynomial expansion, and Principal Component Analysis (PCA) is developed, followed by classification with 4 machine-learning and deep learning models. Each model is evaluated with and without GA hyperparameter tuning on identical cross-validation folds. Overall performance is evaluated using a TOPSIS multi-criteria ranking.

Main contributions

- Pre-collection classification of HLA rarity in a large national cord-blood cohort, with a comparative evaluation of KNN, RF, XGBoost, and a deep MLP, each with and without GA tuning on identical folds.
- Operational definition of HLA rarity via observed 6/6 low-resolution match frequency at HLA-A, -B, and -DRB1.

- Leakage-controlled pipeline combining random-forest feature selection, polynomial terms, and PCA (95% variance) to capture non-linear structure while preventing train-test contamination.
- GA hyperparameter optimization protocol and paired comparisons against non-GA baselines, plus a TOPSIS multi-criteria ranking (benefit: Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC; cost: Brier) to support model selection.

These contributions will be elaborated upon in detail in the subsequent sections, and we will thoroughly explore different aspects of the study.

2. Methods

A key objective of this study is to predict HLA rarity using practical ML approaches. To this end, A comparative evaluation of four classifiers—KNN, RF, XGBoost, and a deep MLP—each assessed with and without GA hyperparameter tuning on identical cross-validation folds is developed. The study follows four main steps: data collection, feature selection, modeling, and evaluation; Figure 1 illustrates the full workflow. All DL and ML models were implemented in Python 3.8.

2.1 Dataset

The dataset comprises 122,000 samples obtained from the UCB repository of the Royan Stem Cell Technology (RSCT), Tehran, Iran. This repository provides services to parents who bank their baby's cord blood for potential future medical use. The dataset includes 28 variables recorded prior to collection and describing donor characteristics and selected pre-collection factors. Records with missing values (n = 35,747) were removed, yielding a final cohort of 86,253 entries. Owing to the presence of sensitive and confidential information, the dataset cannot be shared. Tables 1 and 2 summarize the selected variables and their characteristics; in Table 1, the columns "Positive" and "Negative" indicate the presence or absence of familial diseases, respectively. Family medical history was collected in advance via an online questionnaire facilitated by RSCT.

This retrospective, de-identified study was approved by the Kharazmi University Ethics Committee (Code is IR.KHU.REC.1402.068). All records were de-identified prior to analysis,

with all direct personal identifiers (e.g., names, national IDs, contact information) removed. Written informed consent for cord-blood banking and research use was obtained at the time of collection.

2.2 HLA Frequency and Label Definition

The dataset includes low-resolution genotyping for HLA-A, -B, and -DRB1 for each cord-blood unit. To quantify HLA rarity, A frequency-based definition guided by expert consultation was used. For each unit, it is identified whether there exists any 6/6 low-resolution match—i.e., an identical allele pair at HLA-A, -B, and -DRB1 —elsewhere in the dataset.

Following this, each cord blood unit was labeled as either:

Common: if it matched at least one other unit (i.e., frequency > 0)

Rare: if it had no 6/6 matches in the dataset (i.e., frequency = 0)

This binary, frequency-based operationalization aligns with expert judgment about what constitutes a high-utility (common) versus low-availability (rare) unit for banking. The resulting label served as the ground-truth outcome for all classification models.

Figure 2 shows the distribution of each class created for HLA rarity using a pie chart. The pie chart shows the class distribution of HLA rarity in the final cohort (n = 86,253): Rare = 44,950 (52.1%) and Common = 41,303 (47.9%). The near-balanced split indicates that no resampling was required for model training.

2.3 Feature Selection

Feature selection is a critical step in machine-learning pipelines, particularly with high-dimensional tabular data, where irrelevant or redundant variables can add noise, induce overfitting, and increase computational cost. In biomedical settings, careful selection also aids interpretability by highlighting clinically meaningful variables [28].

The original dataset included 28 features representing demographic, prenatal, and family medical history information. To make the data compatible with machine learning model, one-hot encoding was applied, which expanded the feature space to 32 columns. This transformation

allows categorical variables to be represented as binary indicators, enabling the model to treat each category independently.

2.3.1 Feature Importance

For feature importance RF selection with Recursive Feature Elimination (RFE) and L1-regularized logistic regression (LASSO) was compared. RF selector is a tree-based feature importance that captures non-linear effects and interactions, used to rank and retain predictors. RFE is a wrapper method that iteratively removes the least important features to identify a stable subset. LASSO is a linear model with an L1 penalty that drives uninformative coefficients to zero, yielding a sparse, filter-style feature set [29].

RF and RFE selected identical feature sets (Jaccard = 1.00), indicating robustness to the choice of a tree-based selector. LASSO selected a partially overlapping set (Jaccard = 0.50) that emphasized different variables. In downstream evaluation on the same outer folds, RF and RFE achieved ROC-AUC 0.823 and PR-AUC 0.803, while LASSO was slightly lower (ROC-AUC 0.819, PR-AUC 0.791). These results support RF as a practical, robust default for the main analyses. Figure 3 shows the importance of each feature using RF.

2.3.2 Polynomial Feature Generation

Following feature selection, a polynomial feature expansion of degree 2 was applied, which introduced both squared terms and pairwise interactions between features. This transformation aims to capture non-linear relationships and dependencies that simple linear models might miss [30]. Given 20 input features, polynomial expansion of degree 2 results in Eq. (1):

$$\text{Number of polynomial features} = \binom{n+d}{d} \tag{1}$$

where:

n is the number of original features,

d is the degree of the polynomial,

Number of polynomial features = $\binom{22}{2} = 231$

These 231 features were then standardized using a StandardScaler, which transforms each feature to have zero mean and unit variance. This step ensures that features are on the same scale, which is particularly important for models sensitive to feature magnitude and for subsequent dimensionality reduction.

2.3.3 Principal Component Analysis

To further reduce dimensionality while retaining as much variance as possible, a PCA to the standardized polynomial features was applied. PCA identifies orthogonal directions (principal components) that explain the variance in the dataset and projects the data into a lower-dimensional space [31]. The number of components that preserved 95% of the total variance was retained, which in this case resulted in 80 principal components. Figure 4 shows the variance-retention curve with a dashed reference line at 95% and the selected component count.

2.4 Classification Model

After dimensionality reduction via PCA, the resulting 80 principal components were used as inputs for a binary classification task predicting HLA rarity (common vs rare). The class distribution remained near-balanced (see Figure 2); therefore, no resampling was applied. A comparative evaluation of four classifiers was conducted—each assessed with and without Genetic Algorithm (GA) hyperparameter tuning on identical cross-validation fold with KNN as a distance-based voting among the k most similar samples, RF as ensemble of decision trees with bootstrap sampling and feature subsampling, XGBoost as boosted decision trees optimized by additive gradient updates, and MLP as fully connected deep neural network with nonlinear activations [32, 33]. MLP is considered a deep learning model since it uses 4 hidden layers based[34]. Models were trained within a leakage-controlled pipeline (feature selection, polynomial expansion, scaling, and PCA fit on training folds only and applied to validation/test). Accuracy, Precision, Recall, F1-score, AUC-ROC, AUPRC (area under the precision–recall curve), and Brier score (probability calibration) was reported.

2.5 Hyperparameter Tuning

Models were optimized using Genetic Algorithm, a population-based metaheuristic that efficiently explores mixed, conditional, and non-convex hyperparameter spaces. Genetic Algorithm was used because the hyperparameter space is mixed-type and conditional, yielding a rugged, non-convex objective under noisy inner-CV evaluation. A population-based GA natively handles discrete + continuous + conditional variables without reparameterization, maintains diversity to explore multiple basins of attraction, and is trivially parallelizable within a fixed budget [35, 36]. For each classifier—KNN, Random Forest (RF), XGBoost, and MLP—paired results with and without GA on the same outer cross-validation (CV) folds were reported.

Tuning was performed inside the training data of each outer fold only to prevent leakage. A fixed budget (population = 10, generations = 10), tournament selection, crossover = 0.8, mutation = 0.2, and a fixed seed for reproducibility was used.

2.6 Evaluation Metrics

Classifiers were evaluated using the confusion matrix (True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*)) and threshold-free curves. Below, one-line interpretation and the formula for each metric is given in Eqs. (2)-(10) [37].

Accuracy — overall correctness = $\frac{TN + TP}{TN + TP + FN + FP}$ (2)

Precision — among predicted positives, how many are truly positive = $\frac{TP}{TP + FP}$ (3)

Recall/ TPR — among true positives, how many are found = $\frac{TP}{TP + FN}$ (4)

Specificity — among true negatives, how many are correctly rejected = $\frac{TN}{TN + FP}$ (5)

F1 Score — harmonic mean of Precision and Recall = $2 \times \frac{\text{Pr}ecision \times \text{Re}call}{\text{Pr}ecision + \text{Re}call}$ (6)

FPR — error rates for negatives/positives = $\frac{FP}{FP + TN}$ (7)

AUC-ROC — threshold-independent separability = $\int_0^1 TPR \, d(FPR)$ (8)

AUPRC — precision–recall trade-off, informative with rarer positives =

$$\int_0^1 \Pr ec(\mathrm{Re}c)d(\Pr ec) \tag{9}$$

Brier Score — mean squared error of predicted probabilities (lower is better) =

$$\frac{1}{N}\sum_{i=1}^{N}\left(p_i - y_i\right)^2 \tag{10}$$

TP : The model correctly predicts a positive outcome.

TN : The model correctly predicts a negative outcome.

FP : The model incorrectly predicts a positive outcome when it is actually negative.

FN : The model incorrectly predicts a negative outcome when it is actually positive.

N : number of samples in the evaluation set.

p_i : model-predicted probability of the positive class for sample *i*.

y_i : true label for sample *i*.

2.7 Model Ranking via TOPSIS

To summarize performance across multiple criteria, models were ranked using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Criteria treated as benefits were Accuracy, Precision, Recall, F1, AUC-ROC, AUPRC; Brier score was treated as a cost. Scores were vector-normalized and equally weighted. Then the ideal best (max for benefits, min for cost) and ideal worst (min for benefits, max for cost) were identified and each model's closeness coefficient was computed using Eq. (11):

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{11}$$

where D_i^+ and D_i^- are the Euclidean distances from model i to the ideal best and ideal worst, respectively. Models with larger C_i rank higher. TOPSIS was chosen because it provides a single, transparent ranking that simultaneously reflects discrimination (ROC-AUC/AUPRC), thresholded performance (Precision/Recall/F1/Accuracy), and probability calibration (Brier), aligning with our operational goals without privileging any one metric.

3. Results

Exploratory data analysis was performed to gain a comprehensive understanding of the dataset's structure and feature distributions. This process involved analyzing patterns, distributions, and relationships between features and the target variable, HLA rarity. A correlation heatmap in Figure 5 was generated to visualize relationships between the 20 features selected through feature importance and the HLA_Rarity label. Warmer colors in the heatmap indicate stronger positive or negative correlations, helping to identify influential variables. Given the weak linear correlations observed between most features and HLA rarity, polynomial feature generation may help capture potential non-linear relationships that influence allele distribution.

In addition, Figure 6 presents the distribution plots for each of these 20 selected features, stratified by HLA rarity class (common vs. rare). These visualizations reveal meaningful trends and class-wise imbalances in several features. For instance, features such as infertility, diabetes, and drug use are heavily skewed toward negative cases, with very small portions of positive history recorded. Parental ages, gestational age, and the difference in parents' ages show approximately normal distributions. Some categorical variables like place of birth, type of delivery, blood group, and Rh factor display distinct frequency patterns across rarity classes, suggesting a possible role in classification. Together, these figures provide essential context for interpreting feature relevance and support the subsequent phases of model development.

Tables 3–6 summarize the hyperparameter configurations evaluated for each classifier. For every parameter, we report the search space, a short description, the default value used in non-GA baselines, and the optimal value selected by the GA during nested cross-validation. All tuning was performed inside training folds only to prevent leakage; final models were retrained on the outer-train split with selected hyperparameters and evaluated on the held-out outer-test split.

The classification performance and paired comparision of four classifiers with and without GA were compared in table 7. This table shows a reasonable performance on all metrics, especially ROC curves.

Models were ranked with TOPSIS using equal weights, treating Accuracy, Precision, Recall, F1, ROC-AUC, PR-AUC as benefits and Brier as a cost. Scores were vector-normalized and computed on identical outer CV folds. The ranking supports selecting the GA-tuned MLP as the primary model (Figure 7).

The ROC curve in figure 8(a) illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various classification thresholds. In this plot, the orange curve represents the classifier's performance, while the diagonal blue dashed line indicates the performance of a random classifier. The curve lies well above the diagonal, indicating strong discriminative ability. The Area Under the Curve (AUC) is 0.84, suggesting that the model is able to distinguish between the rare and common HLA classes with high accuracy.

The precision-recall curve in figure 8(b) provides a detailed view of the balance between precision and recall across different thresholds. In this plot, the blue curve shows that the model maintains a high precision rate even at moderate levels of recall. The area under the PR curve is 0.81, indicating that the classifier sustains strong precision at practically useful recall levels—consistent with prioritizing rare units while limiting false positives. PR is a complementary, threshold-free view to ROC and is particularly informative when the positive class is less prevalent.

The confusion matrix visualizes the performance of the binary classification model in predicting HLA rarity using GA-tunned MLP. Figure 9 provides a breakdown of true positives, true negatives, false positives, and false negatives for the two classes: "Common" and "Rare". The matrix summarizes model predictions on the test set, with actual classes on the vertical axis and predicted classes on the horizontal axis.

Table 8 reports the confusion-matrix counts and the derived class-wise rates for the GA-tuned MLP with Rare as the positive class. From the counts (TP = 6,231; FP = 1,671; TN = 6,590; FN = 2,759), TPR/Recall and TNR/Specificity were calculated. The Balanced Accuracy—the average of TPR and TNR—is 0.745, summarizing performance across both classes.

4. Discussion

This study presents a machine learning framework for predicting HLA rarity in UCB units at the pre-collection stage. Unlike prior research that primarily focused on donor availability prediction or HLA imputation, this work introduces a novel binary classification task grounded in expert-informed frequency thresholds derived from national-scale data. The model's objective—to flag units with rare HLA profiles before full-resolution genotyping—addresses a practical gap in cord blood bank management by enabling proactive inventory curation. We compared KNN, Random Forest, XGBoost, and a deep MLP, each evaluated with and without Genetic Algorithm (GA) tuning on identical folds, and summarized overall performance using TOPSIS.

The RF feature importance analysis highlights that place of birth, parental ages, and gestational age are the strongest predictors of HLA rarity, while many medical history features—such as hepatitis, anemia, or donation history—had negligible impact. This aligns with prior studies emphasizing the demographic structuring of HLA diversity; for example, regionally isolated populations like those in Crete exhibit distinct HLA distributions due to founder effects and population history [17]. The high importance of parental age difference and consanguinity (related parents) further supports genetic inheritance as a key driver, consistent with the Mendelian transmission of HLA haplotypes and the elevated homozygosity observed in consanguineous populations [13]. From the set of medical history features examined, abortion history emerged as the most important predictor in the model, underscoring its strong association with HLA rarity in this dataset.

In terms of feature engineering, the use of RF-based feature importance selection, followed by polynomial feature generation and PCA, proved effective in optimizing the model input space. The resulting dimensionality-reduced feature set retained 95% of the original variance, balancing complexity and interpretability. This preprocessing pipeline, combined with hyperparameter tuning via genetic algorithm, underscores the importance of careful pipeline design in biomedical tabular ML tasks. In a sensitivity analysis, RFE (with RF estimator) selected an identical feature set to RF (Jaccard = 1.00), while LASSO selected a partially overlapping set (Jaccard = 0.50), with downstream performance for RF/RFE slightly exceeding LASSO.

Across models, the GA-tuned MLP achieved strong discrimination and calibration (AUC-ROC 0.84; AUPRC 0.81; Brier 0.166), with GA-tuned XGBoost close behind; KNN variants

ranked lowest. The classification report and curves summarize performance in predicting HLA rarity. For the Rare class (positive), the confusion matrix yields Precision = 0.79, Recall/TPR = 0.69, and F1 = 0.74; the PR curve indicates that precision remains high over moderate recall ranges, a desirable property for prioritizing rare units while limiting false positives. For the Common class, Precision ≈ 0.71 and Recall/Specificity (TNR) ≈ 0.80, reflecting the expected trade-off at the default threshold. Threshold adjustments could further tune this trade-off to match banking capacity and genotyping budgets. A TOPSIS multi-criteria ranking, ranked MLP (GA-tuned) first, followed by XGBoost (GA-tuned).

Compared to existing literature, our approach builds on and diverges from several key strands. For example, donor availability prediction studies use machine learning to forecast whether a registered donor will follow through with donation, often using behavioral and operational variables [23, 38]. While these models improve registry logistics, they do not target the genetic diversity needs of the bank inventory. In contrast, our framework is designed for pre-collection decision support aimed at anticipating HLA rarity, making it practical for early-stage triage and resource allocation.

This study provides a proof-of-concept that could inform operational improvements in cord blood banking. By identifying and tagging units likely to be genetically rare, banks can prioritize them for long-term storage, optimize collection efforts, and potentially improve matching success rates—especially for underrepresented populations. These insights could also guide economic decision-making, reducing costs associated with typing or discarding units with redundant genetic profiles.

5. Conclusion

In conclusion, this study introduces a pre-collection prediction framework for classifying the rarity of HLA profiles in cord-blood units using a frequency-based binary label grounded in 6/6 low-resolution matches at HLA-A, -B, and -DRB1. Within a leakage-controlled pipeline (feature selection, polynomial expansion, PCA), and then KNN, Random Forest, XGBoost, and a deep MLP were compared, each with and without Genetic Algorithm (GA) tuning on identical folds. The GA-tuned MLP delivered the strongest overall performance, with GA-tuned XGBoost close

behind; this ordering was corroborated by a TOPSIS multi-criteria ranking. These findings support early, pre-collection triage of units likely to be rare, aiding strategic storage and potentially improving match opportunities. This work differs from prior studies emphasizing donor availability or genotype imputation by targeting early-stage unit selection for inventory design. While results are promising, several limitations remain: the dataset is region-specific (Iran), rarity was defined using low-resolution HLA matches, and external, multi-center validation is needed to assess generalizability over time and across populations. Future work should evaluate high-resolution definitions, test the framework on additional registries, explore alternative global optimizers for hyperparameters, and integrate calibrated predictions into operational decision systems (e.g., threshold tuning by capacity and budget) for real-time cord-blood banking support.

Declaration of competing interest

All the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' contributions

Amir Hossein Esmaielpour: Writing – original draft, Conceptualization, Methodology, Formal analysis, Data curation, Software, Visualization. Mariam Ameli: Writing – review & editing, Supervision, Validation. Ashkan Mozdgir: Writing – review & editing, Conceptualization, Investigation, Resources. Orod Ahmadi: Writing – review & editing, Supervision, Conceptualization, Software. Morteza Zarabi: Writing – review & editing, Methodology, Supervision, Conceptualization.

6. References

[1] Society A.C., "Global cancer facts & figures, 5th edition," Atlanta, GA, 2024. Accessed: 10/28/2025. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-2024.pdf

[2] Ding Y., Liu H., Wang Y., et al., "Romiplostim versus recombinant human thrombopoietin in umbilical cord blood transplantation: a single center retrospective study," Annals of Hematology, vol. 104, no. 7, pp. 3837–3842, 2025, https://doi.org/10.1007/s00277-025-06490-z.

[3] Mohammadian Behbahani Z., Karimi B., Zarrabi M., et al., "Designing an efficient supply chain network for public cord blood bank and quality prediction to improve performance: a case study in Iran," Operational Research, vol. 25, no. 3, p. 75, 2025, https://doi.org/10.1007/s12351-025-00947-9.

[4] Liedtke S., Többen S., Gressmann H., et al., "Long term stability of cord blood units after 29 years of cryopreservation: follow up data from the José Carreras cord blood bank," Stem Cells Translational Medicine, vol. 13, no. 1, pp. 30–42, 2024, https://doi.org/10.1093/stcltm/szad071.

[5] Inamoto Y., "Alternative donors: a match for matched sibling donors?," The Lancet Haematology, vol. 6, no. 11, pp. e545–e546, 2019, https://doi.org/10.1016/S2352-3026(19)30163-2.

[6] Song E. Y., Huh J. Y., Kim S. Y., et al., "Estimation of size of cord blood inventory based on high resolution typing of HLAs," Bone Marrow Transplantation, vol. 49, no. 7, pp. 977–979, 2014, https://doi.org/10.1038/bmt.2014.76.

[7] Fumarola S., Lucarini A., Lucchetti G., et al., "Predictors of cord blood unit cell content in a volume unrestricted large series collections: a chance for a fast and cheap multiparameter selection model," Stem Cell Research & Therapy, vol. 13, no. 1, p. 246, 2022, https://doi.org/10.1186/s13287-022-02915-y.

[8] Pupella S., Bianchi M., Ceccarelli A., et al., "A cost analysis of public cord blood banks belonging to the Italian cord blood network," Blood Transfusion, vol. 16, no. 3, pp. 313–320, 2018, https://doi.org/10.2450/2017.0251-16.

[9] Tanaka K., Kato K., Nonaka N., et al., "Efficient HLA imputation from sequential SNPs data by transformer," Journal of Human Genetics, vol. 69, no. 10, pp. 533–540, 2024, https://doi.org/10.1038/s10038-024-01278-x.

[10] Logan B. R., Maiers M. J., Sparapani R. A., et al., "Optimal donor selection for hematopoietic cell transplantation using Bayesian machine learning," JCO Clinical Cancer Informatics, vol. 5, pp. 494–507, 2021, https://doi.org/10.1200/cci.20.00185.

[11] Choi E. J., Jun T. J., Park H. S., et al., "Predicting long term survival after allogeneic hematopoietic cell transplantation in patients with hematologic malignancies: machine learning based model development and validation," JMIR Medical Informatics, vol. 10, no. 3, p. e32313, 2022, https://doi.org/10.2196/32313.

[12] Naito T., Suzuki K., Hirata J., et al., "A deep learning method for HLA imputation and trans ethnic MHC fine mapping of type 1 diabetes," Nature Communications, vol. 12, no. 1, p. 1639, 2021, https://doi.org/10.1038/s41467-021-21975-x.

[13] Choo S. Y., "The HLA system: genetics, immunology, clinical testing, and clinical implications," Yonsei Medical Journal, vol. 48, no. 1, pp. 11–23, 2007, https://doi.org/10.3349/ymj.2007.48.1.11.

[14] Ober C., Hyslop T., Elias S., et al., "Human leukocyte antigen matching and fetal loss: results of a 10 year prospective study," Human Reproduction, vol. 13, no. 1, pp. 33–38, 1998, https://doi.org/10.1093/humrep/13.1.33.

[15] Stout M. J., Cao B., Landeau M., et al., "Increased human leukocyte antigen G expression at the maternal fetal interface is associated with preterm birth," Journal of Maternal Fetal and Neonatal Medicine, vol. 28, no. 4, pp. 454–459, 2015, https://doi.org/10.3109/14767058.2014.921152.

[16] Akesson K., Carlsson A., Ivarsson S. A., et al., "The non inherited maternal HLA haplotype affects the risk for type 1 diabetes," International Journal of Immunogenetics, vol. 36, no. 1, pp. 1–8, 2009, https://doi.org/10.1111/j.1744-313X.2008.00802.x.

[17] Latsoudis H., Loulakaki M., Pavlidis P., et al., "The intra population HLA diversity in Crete and the importance of the regional public umbilical cord blood bank in HLA based donor selection for allogeneic hematopoietic stem cell transplantation," HemaSphere, vol. 6, pp. 1237–1238, 2022, https://doi.org/10.1097/01.HS9.0000848272.66784.75.

[18] Leung C. K., Zhu P., Loke I., et al., "Development of a quantitative prediction algorithm for human cord blood derived CD34+ hematopoietic stem progenitor cells using parametric and non parametric machine learning models," Scientific Reports, vol. 14, no. 1, p. 25085, 2024, https://doi.org/10.1038/s41598-024-75731-4.

[19] Solves P., Perales A., Moraga R., et al., "Maternal, neonatal and collection factors influencing the haematopoietic content of cord blood units," Acta Haematologica, vol. 113, no. 4, pp. 241–246, 2005, https://doi.org/10.1159/000084677.

[20] Haghbayan M., Karimi B., Mozdgir A., et al., "Increasing the efficacy of umbilical cord blood banking using machine learning algorithms: a case study from Royan cord blood bank," Scientia Iranica, 2023, https://doi.org/10.24200/sci.2023.61068.7132.

[21] Samarkanova D., Codinach M., Montemurro T., et al., "Multi component cord blood banking: a proof of concept international exercise," Blood Transfusion, vol. 21, no. 6, pp. 526–537, 2023, https://doi.org/10.2450/BloodTransfus.492.

[22] Israeli S., Gragert L., Maiers M., et al., "HLA haplotype frequency estimation for heterogeneous populations using a graph based imputation algorithm," Human Immunology, vol. 82, no. 10, pp. 746–757, 2021, https://doi.org/10.1016/j.humimm.2021.07.001.

[23] Li Y., Masiliune A., Winstone D., et al., "Predicting the availability of hematopoietic stem cell donors using machine learning," Biology of Blood and Marrow Transplantation, vol. 26, no. 8, pp. 1406–1413, 2020, https://doi.org/10.1016/j.bbmt.2020.03.026.

[24] Esmaielpour A. H., Ameli M., Mozdgir A., et al., "Predicting pre collection umbilical cord blood clotting using advanced machine learning algorithms," Blood Journal, vol. 21, no. 2, p. 151, 2024. [Online]. Available: http://bloodjournal.ir/article-1-1520-en.html [In Persian].

[25] Xiang J., Shi M., Fiala M. A., et al., "Machine learning based scoring models to predict hematopoietic stem cell mobilization in allogeneic donors," Blood Advances, vol. 6, no. 7, pp. 1991–2000, 2022, https://doi.org/10.1182/bloodadvances.2021005149.

[26] Leung C. K., Zhu P., Loke I., et al., "Development of a quantitative prediction algorithm for human cord blood derived CD34+ hematopoietic stem progenitor cells using parametric and non parametric machine learning models," Scientific Reports, vol. 14, no. 1, p. 25085, 2024, https://doi.org/10.1038/s41598-024-75731-4.

[27] Gupta V., Braun T. M., Chowdhury M., et al., "A systematic review of machine learning techniques in hematopoietic stem cell transplantation (HSCT)," Sensors, vol. 20, no. 21, 2020, https://doi.org/10.3390/s20216100.

[28] Karimi A., Mozdgir A., Tashakkori R., et al., "Predicting antenatal depression during 3rd trimester: a machine learning approach with feature selection and TOPSIS ranking," Scientia Iranica, 2024, https://doi.org/10.24200/sci.2024.63194.8288.

[29] Ebrahimi Warkiani M., Moattar M. H., "A comprehensive survey on recent feature selection methods for mixed data: challenges, solutions and future directions," Neurocomputing, vol. 623, p. 129372, 2025, https://doi.org/10.1016/j.neucom.2025.129372.

[30] Wang D., Huang Y., Ying W., et al., "Towards data centric AI: a comprehensive survey of traditional, reinforcement, and generative approaches for tabular data transformation," 2025, https://doi.org/10.48550/arXiv.2501.10555.

[31] Ma S., Dai Y., "Principal component analysis based methods in bioinformatics studies," Briefings in Bioinformatics, vol. 12, no. 6, pp. 714–722, 2011, https://doi.org/10.1093/bib/bbq090.

[32] Bringewald J., "Solar flare forecast: a comparative analysis of machine learning algorithms for solar flare class prediction," arXiv preprint, arXiv:2505.03385, 2025, https://doi.org/10.48550/arXiv.2505.03385.

[33] Patharkar A., Cai F., Al Hindawi F., et al., "Predictive modeling of biomedical temporal data in healthcare applications: review and future directions," Frontiers in Physiology, vol. 15, p. 1386760, 2024, https://doi.org/10.3389/fphys.2024.1386760.

[34] Tripathi A., Tiwari R. K., Tiwari S. P., "A deep learning multi layer perceptron and remote sensing approach for soil health based crop yield estimation," International Journal of Applied

Earth Observation and Geoinformation, vol. 113, p. 102959, 2022, https://doi.org/10.1016/j.jag.2022.102959.

[35] Xue C., Li J., Wang X., et al., "On neural architecture search and hyperparameter optimization: a max flow based approach," Neural Networks, vol. 188, p. 107507, 2025, https://doi.org/10.1016/j.neunet.2025.107507.

[36] Bakır H., Ceviz Ö., "Empirical enhancement of intrusion detection systems: a comprehensive approach with genetic algorithm based hyperparameter tuning and hybrid feature selection," Arabian Journal for Science and Engineering, vol. 49, no. 9, pp. 13025–13043, 2024, https://doi.org/10.1007/s13369-024-08949-z.

[37] Chicco D., Jurman G., "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," BioData Mining, vol. 16, no. 1, p. 4, 2023, https://doi.org/10.1186/s13040-023-00322-4.

[38] Sivasankaran A., Williams E., Albrecht M., et al., "Machine learning approach to predicting stem cell donor availability," Biology of Blood and Marrow Transplantation, vol. 24, no. 12, pp. 2425–2432, 2018, https://doi.org/10.1016/j.bbmt.2018.07.035.

[39] Nikghadam Hojjati S., Reza Rabi A., "Effects of Iranian online behavior on the acceptance of internet banking," Journal of Asia Business Studies, vol. 7, no. 2, pp. 123–139, 2013, https://doi.org/10.1108/15587891311319422.

Biographies

Amir Hossein Esmaielpour received the Master degree in Industrial Engineering from Semnan University, Semnan, Iran. He is currently a Ph.D student at the Department of Industrial Engineering, Faculty of Engineering, Kharazmi University. His research interests include machine learning, data mining, and operations research, with a specific focus on healthcare applications and predictive modeling in medical systems to optimize decision-making processes.

Mariam Ameli received the Ph.D. degree in Industrial Engineering from Amirkabir University, Tehran, Iran. She is currently an Associate Professor in the Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran. Her research focuses on applying optimization techniques, machine learning and systemic methodologies to advance sustainable development.

Ashkan Mozdgir received the Ph.D. degree in Industrial Engineering from K. N. Toosi University of Technology. He is currently an Assistant Professor in the Department of Industrial Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran. His research focuses on optimization techniques, with a particular interest in applying machine learning and expert systems to solve real-world problems in healthcare and management.

Orod Ahmadi received the Ph.D. degree in Industrial Engineering from K. N. Toosi University of Technology, Tehran, Iran. He is currently an Assistant Professor in the Department of Industrial Engineering, Faculty of Engineering, Kharazmi University. He previously served as a Quality Control Consultant for the Royan Cord Blood Bank. His primary research interests include statistical quality control, data analytics, and the implementation of machine learning models in industrial and medical systems.

Morteza Zarrabi received the M.D. degree from Iran University of Medical Sience. He is currently a faculty member at the Royan Research Institute and serves as the CEO of the Royan Stem Cell Technology Company (RSCT), Tehran, Iran. As an expert in cord blood banking and stem cell transplantation, his work focuses on the clinical applications of hematopoietic stem cells, HLA diversity, and the management of bio-repositories to improve transplantation outcomes.

List of Figure Captions.

List of Table Titles.

Table 4. XGBoost hyperparameter search, defaults, and GA-selected optima. Parameters include the number of boosting rounds (n_estimators), learning_rate, max_depth, subsample, colsample_bytree, min_child_weight, and regularization terms (reg_lambda, reg_alpha, gamma)

Table 5. Random Forest hyperparameter search, defaults, and GA-selected optima. Reported parameters include n_estimators, max_features, min_samples_split, min_samples_leaf, class_weight, and bootstrap.

Table 6. MLP hyperparameter search, defaults, and GA-selected optima. The table summarizes hidden-layer configurations (hidden_layers), activation, dropout, optimizer, learning_rate, batch_size, epochs (with early stopping), and L2 weight decay.

Table 7. the classification performance and paired comparision of four classifiers with and without GA.

Table 8. Confusion-matrix counts and derived rates for the GA-tuned MLP (positive = Rare) on the held-out test set.

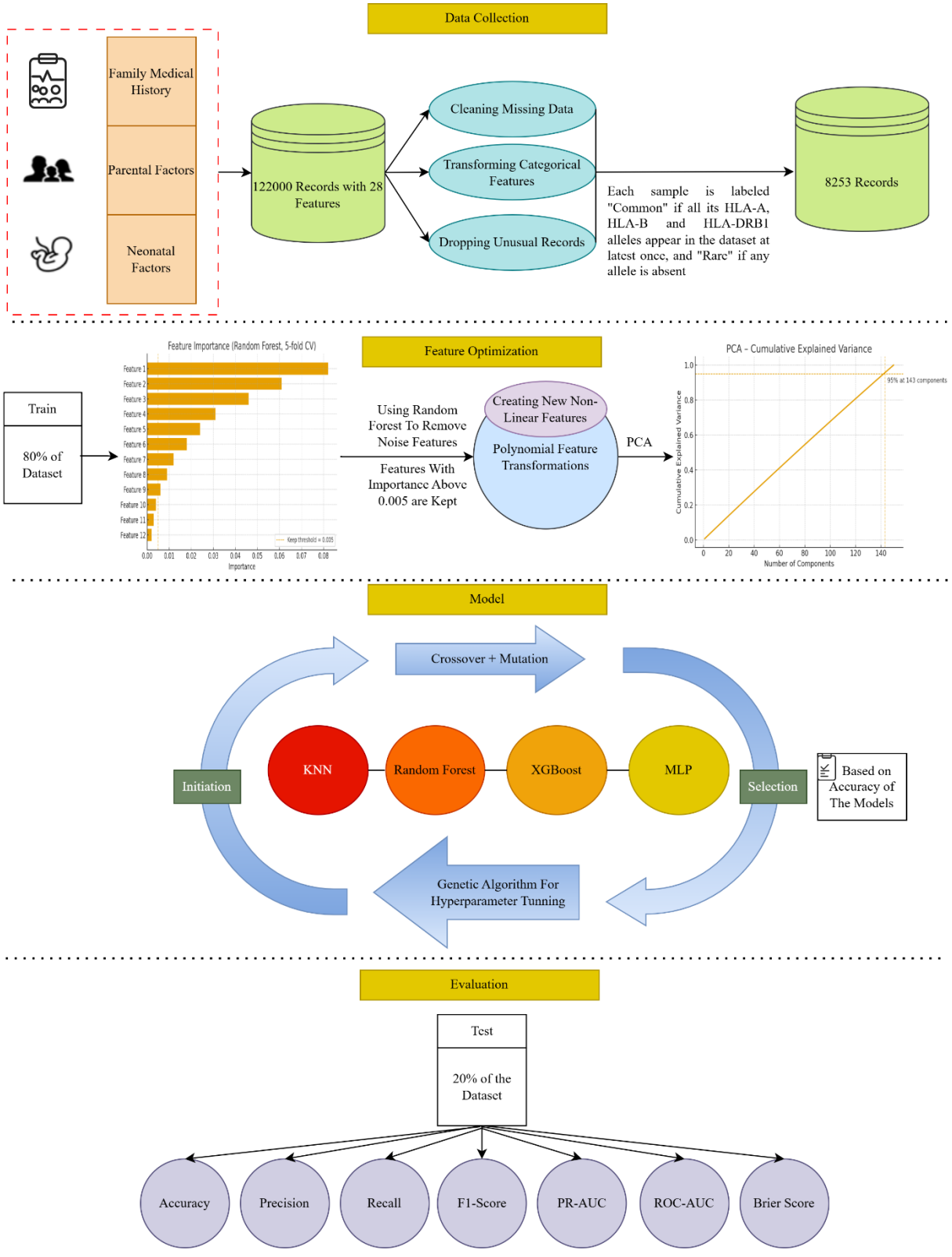
Data Collection
Family Medical History
Parental Factors
Neonatal Factors
122000 Records with 28 Features
Cleaning Missing Data
Transforming Categorical Features
Dropping Unusual Records
Each sample is labeled "Common" if all its HLA-A, HLA-B and HLA-DRB1 alleles appear in the dataset at latest once, and "Rare" if any allele is absent
8253 Records
Train
80% of Dataset
Feature Importance (Random Forest, 5-fold CV)
Importance
Feature Optimization
Using Random Forest To Remove Noise Features
Features With Importance Above 0.005 are Kept
Creating New Non-Linear Features
Polynomial Feature Transformations
PCA
PCA - Cumulative Explained Variance
Cumulative Explained Variance
Number of Components
95% at 143 components
Model
Crossover + Mutation
Initiation
KNN
Random Forest
XGBoost
MLP
Selection
Based on Accuracy of The Models
Genetic Algorithm For Hyperparameter Tunning
Evaluation
Test
20% of the Dataset
Accuracy
Precision
Recall
F1-Score
PR-AUC
ROC-AUC
Brier Score

Figure 1

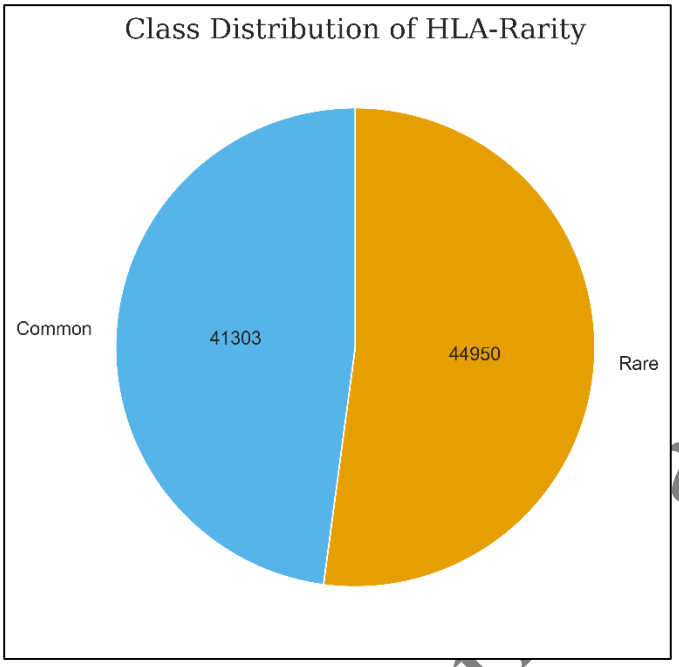

Figure 2

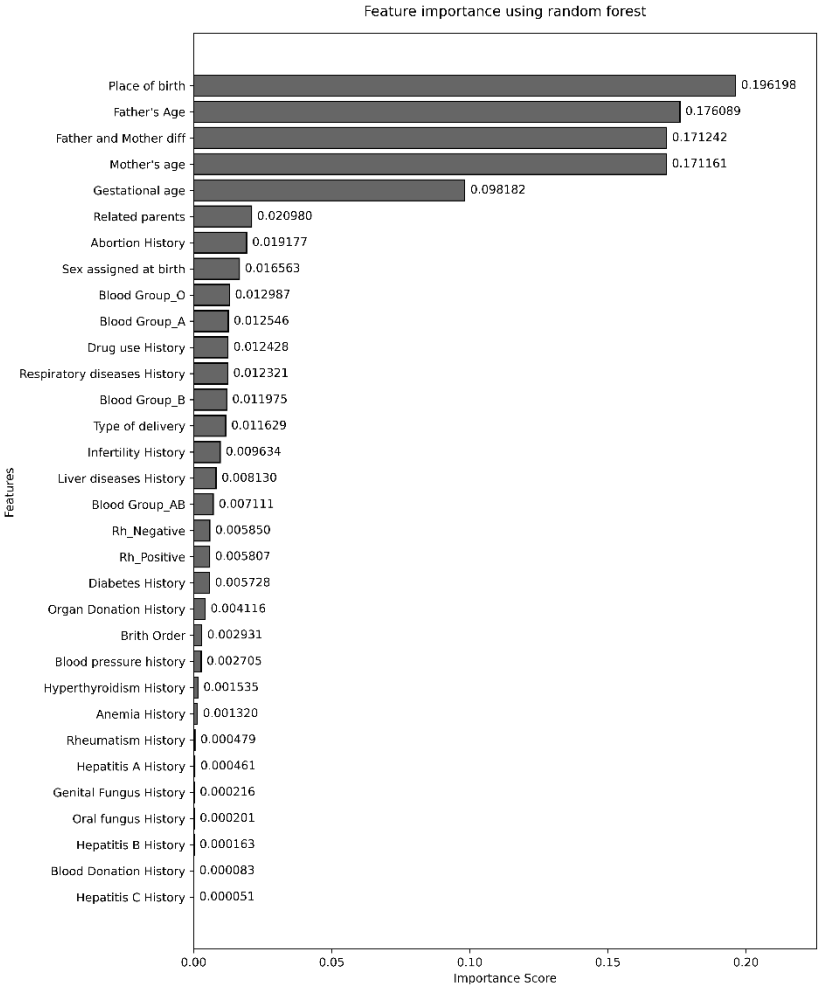

Figure 3

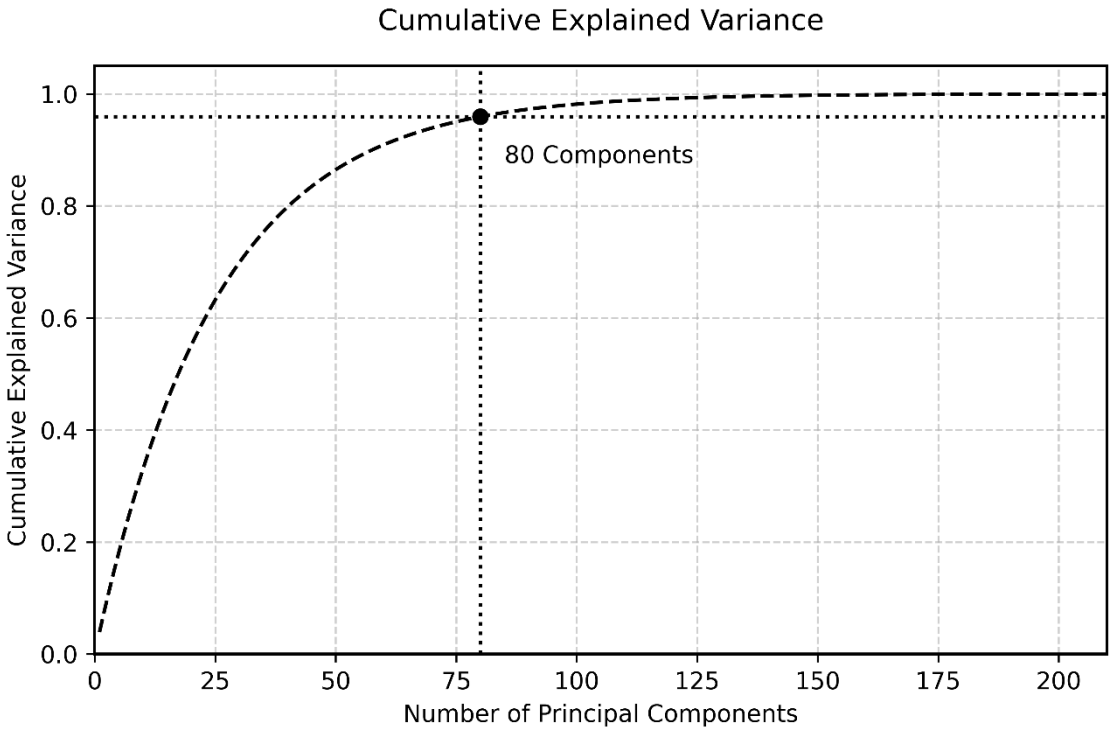

Figure 4

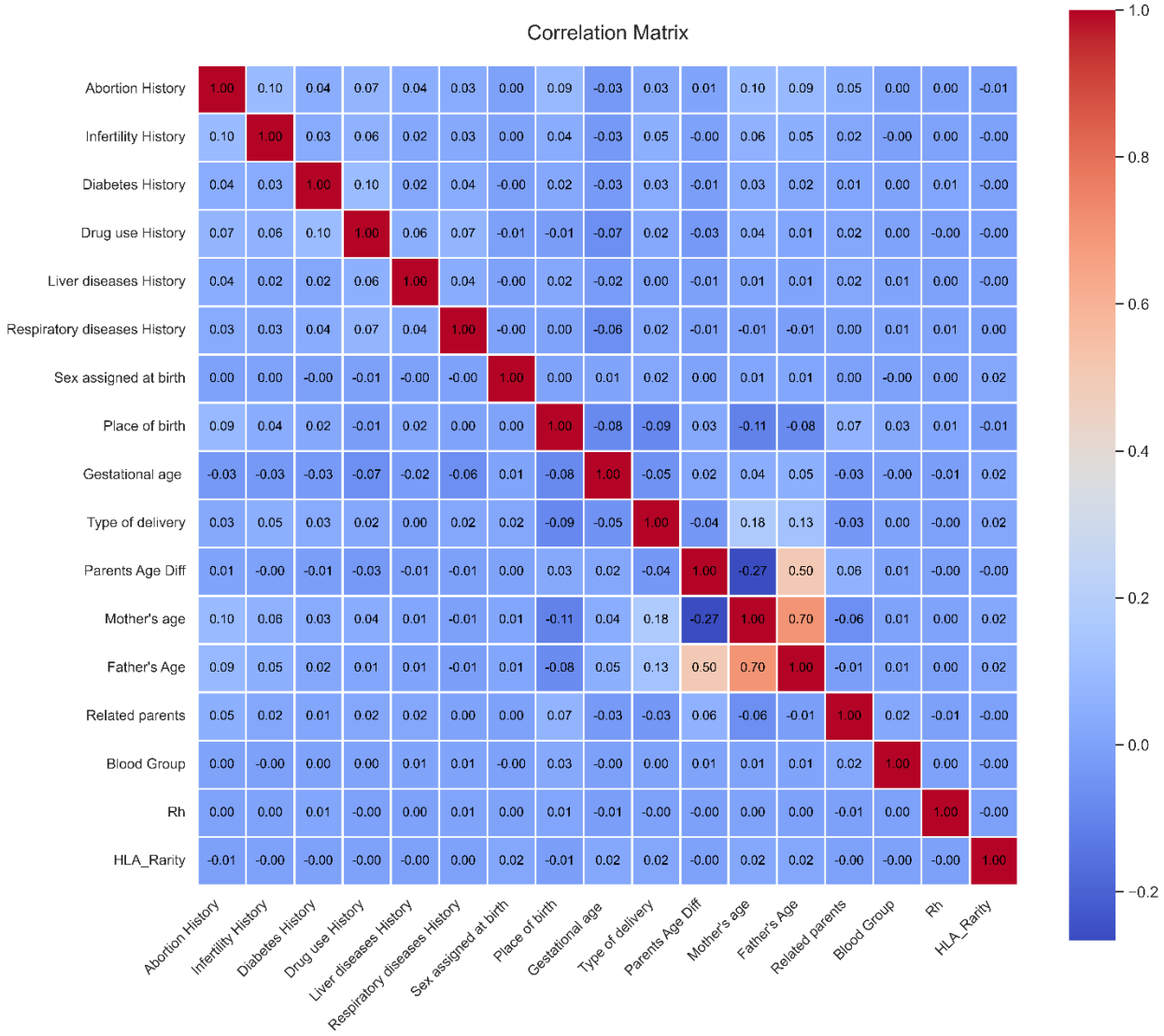

Figure 5

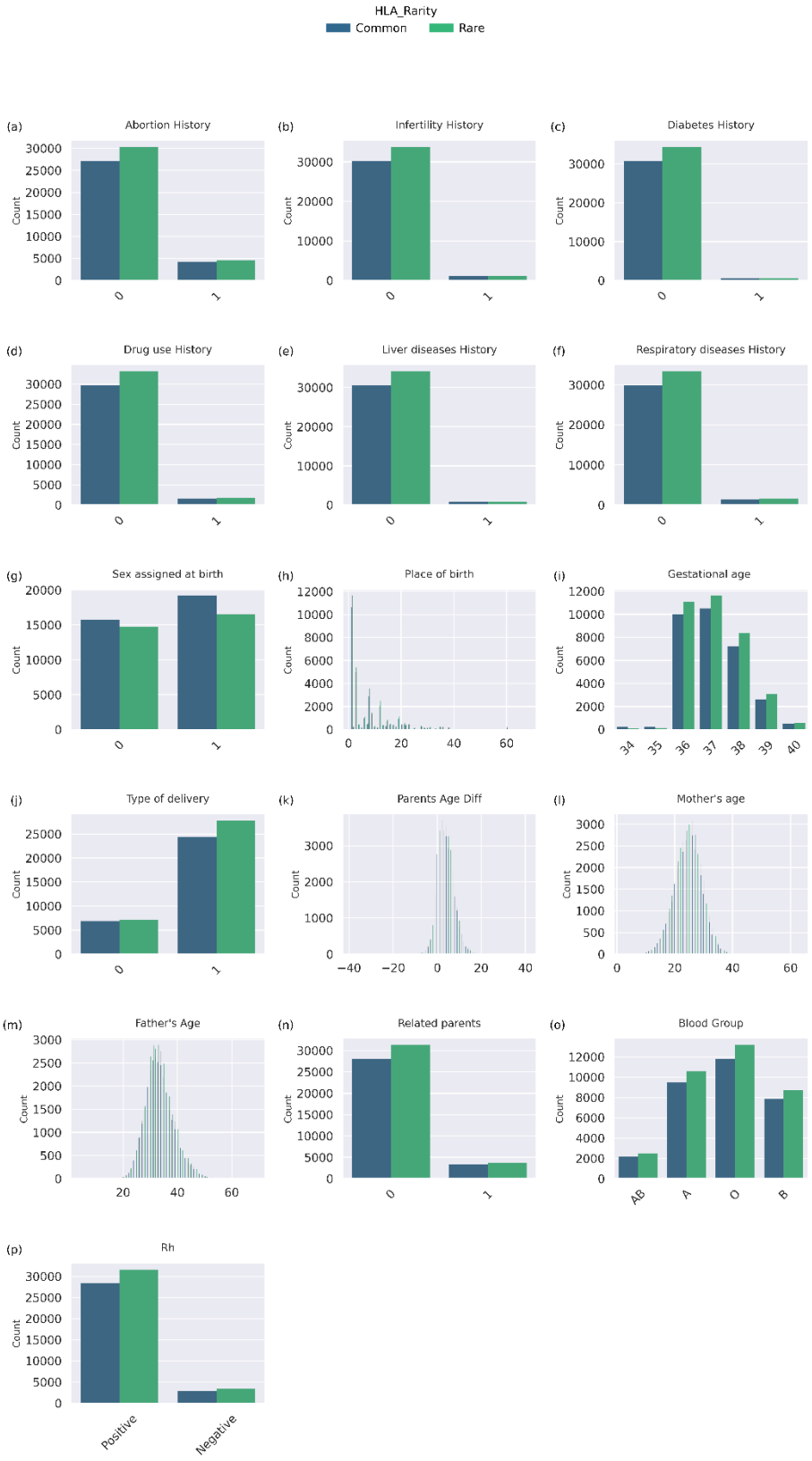
HLA_Rarity
Common
Rare
(a) Abortion History
(b) Infertility History
(c) Diabetes History
(d) Drug use History
(e) Liver diseases History
(f) Respiratory diseases History
(g) Sex assigned at birth
(h) Place of birth
(i) Gestational age
(j) Type of delivery
(k) Parents Age Diff
(l) Mother's age
(m) Father's Age
(n) Related parents
(o) Blood Group
(p) Rh
Count
Positive
Negative
AB
A
O
B

Figure 6

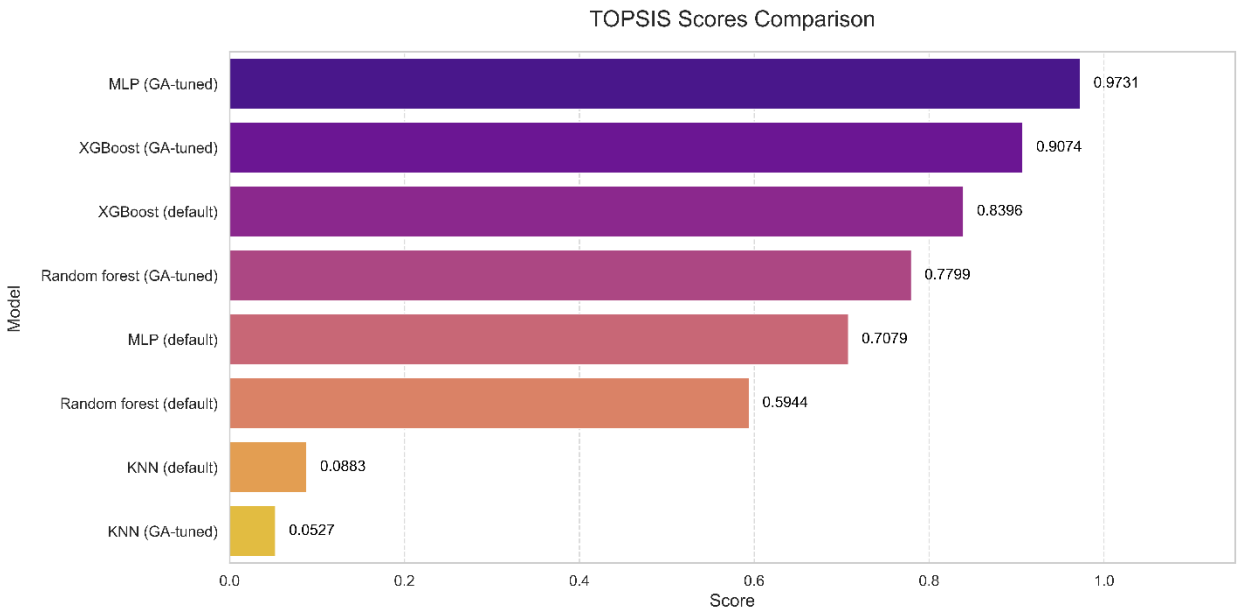

Figure 7

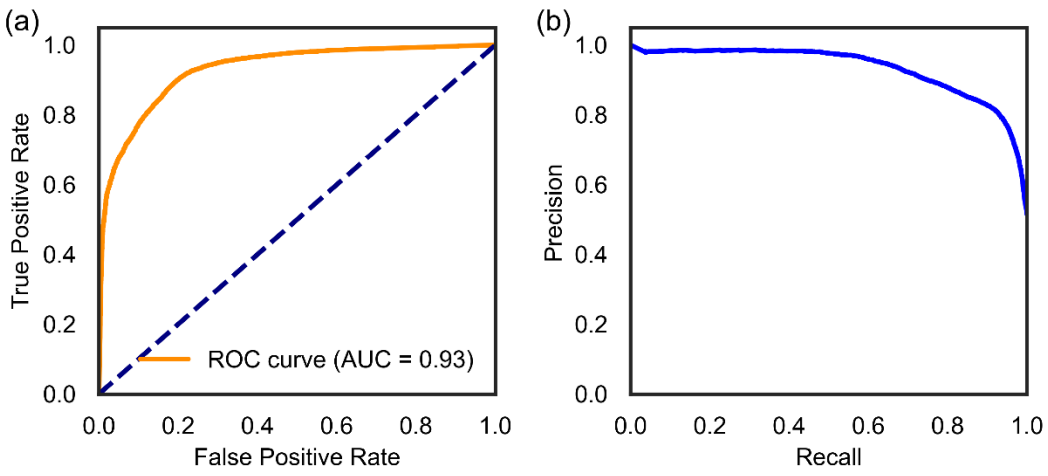

Figure 8

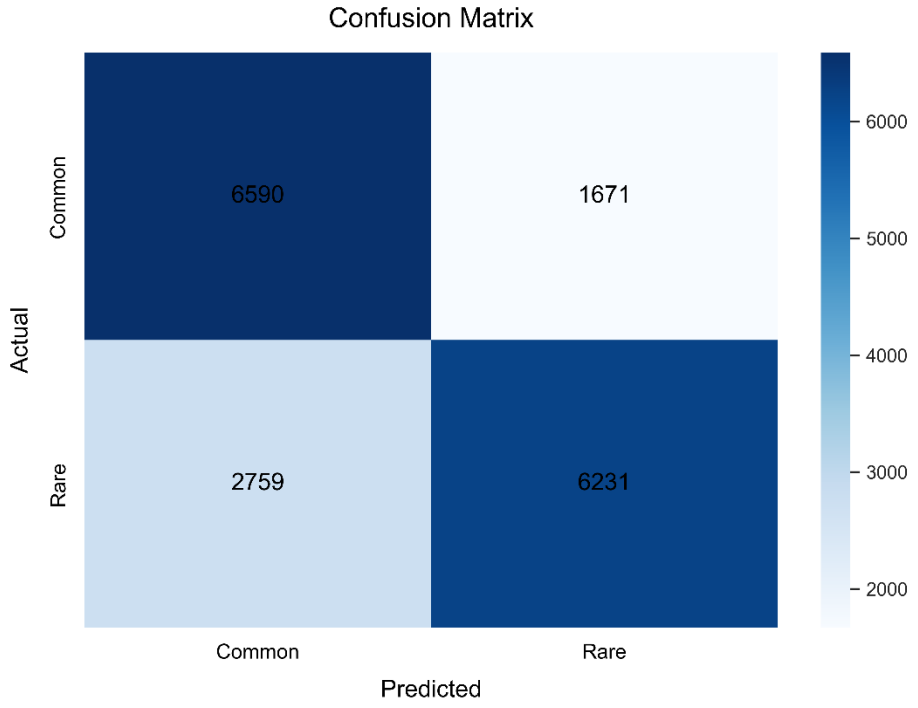

Figure 9

Table 1

Features	Data Type	Positive	Negative
Abortion History	Nominal	17644	68609
Infertility History	Nominal	4346	81907
Blood pressure history	Nominal	860	85393
Diabetes History	Nominal	2098	84155
Anemia History	Nominal	378	85875
Organ Donation History	Nominal	1104	85149
Hepatitis A History	Nominal	121	86132
Hepatitis B History	Nominal	66	86187
Hepatitis C History	Nominal	29	86224
Drug Use History	Nominal	6516	79737
Genital Fungus History	Nominal	87	86166
Blood Donation History	Nominal	40	86213
Liver diseases History	Nominal	3595	82658
Respiratory diseases History	Nominal	6722	79531
Hyperthyroidism History	Nominal	322	85931
Oral fungus History	Nominal	247	86006
Rheumatism History	Nominal	96	86157

Table 2

Features	Data Type	Numerical characteristics
Sex assigned at birth	Nominal	Male = 45693 Female = 40560
Type of delivery	Nominal	Cesarean = 62235 Natural = 24018
Related parents	Nominal	Positive = 12951 Negative =73302
Birth place	Nominal	63 different locations
Gestational age	Numerical	Average = 37.10 std = 1.04 min = 34 max = 40
Mother's age	Numerical	Average = 24.66 std = 4.67 min = 11 max = 53
Father's age	Numerical	Average = 33.5 std = 5.27 min = 14 max = 69
Parents' Age Difference	Numerical	Mean=8.86 std = 3.87

Table 3

Parameter	Search Space	Description	Default Value	Optimal Value
k (neighbors)	[5, 25]	Number of neighbors	5	15
weights	{uniform, distance}	Vote weighting	uniform	distance
metric	{minkowski(p=2), manhattan}	Distance metric	minkowski (p=2)	minkowski (p=2)
leaf_size	[20, 50]	Tree leaf size (speed/accuracy trade-off)	30	30

Table 4

Parameter	Search Space	Description	Default Value	Optimal Value
n_estimators	[100, 1000]	Number of boosting rounds	100	752
learning_rate	[0.01, 0.10]	Shrinkage	0.10	0.05
max_depth	[3, 4, 5, 6}	Tree depth	6	5
subsample	[0.5, 1.0]	Row sampling per tree	1.0	0.8
colsample_bytree	[0.5, 1.0]	Feature sampling per tree	1.0	0.8
min_child_weight	{1, 2, 3}	Min leaf Hessian (regularization)	1	2
reg_lambda	{0, 1, 5}	L2 regularization	1	1
reg_alpha	[0, 0.1]	L1 regularization	0	0.0
gamma	[0, 0.1]	Split gain threshold	0	0.0

Table 5

Parameter	Search Space	Description	Default Value	Optimal Value
n_estimators	[100, 1000]	Number of trees	500	700
max_features	{sqrt, log2}	Features per split	sqrt	sqrt
min_samples_split	{2, 4, 6}	Min samples to split	2	4
min_samples_leaf	{1, 2, 4}	Min samples per leaf	1	2
class_weight	{None, balanced_subsample}	Handle imbalance	None	balanced_subsample
bootstrap	{True, False}	Bootstrap sampling	True	True

Table 6

Parameter	Search Space	Description	Default Value	Optimal Value
Hidden layers	{1, 2, 3, 4, 5, 6}	Number of hidden layers	1	4
activation	{ReLU, tanh}	Nonlinearity	ReLU	ReLU
dropout	[0.1, 0.3]	Dropout after each hidden layer	0.3	0.3
optimizer	[39]	Optimizer	Adam	Adam
learning_rate	[0.001, 0.01]	Initial LR	0.001	0.003
batch_size	{32, 64, 128}	Mini-batch size	64	64
epochs	{30, 50, 80}	Max epochs	50	50
L2 weight decay (alpha)	[0.0, 1e-4]	Regularization	0.0	1e-3

Table 7

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Brier Score
KNN (default)	0.66	0.67	0.70	0.68	0.74	0.71	0.242
KNN (GA-tuned)	0.68	0.67	0.69	0.68	0.74	0.70	0.250
Random Forest (default)	0.72	0.72	0.70	0.71	0.77	0.75	0.195
Random Forest (GA-tuned)	0.74	0.73	0.72	0.72	0.79	0.77	0.178
XGBoost (default)	0.73	0.74	0.72	0.73	0.82	0.80	0.178
XGBoost (GA-tuned)	0.75	0.74	0.72	0.73	0.82	0.80	0.168
MLP (default)	0.71	0.73	0.70	0.71	0.80	0.79	0.187
MLP (GA-tuned)	0.74	0.74	0.74	0.74	0.84	0.81	0.166

Table 8

Metric	Value
$\boldsymbol{TP}$	6,231
$\boldsymbol{FP}$	1,671
$\boldsymbol{TN}$	6,590
$\boldsymbol{FN}$	2,759
TPR(True Positive Rate) = $\frac{TP}{TP+FN}$	0.693
TNR(True Negative Rate) = $\frac{TN}{TN+FP}$	0.798
Balanced Accuracy = $\frac{(TPR+TNR)}{2}$	0.745