International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ADVANCED STEM CELL DONOR MATCHING USING A HYBRID RANDOM FOREST AND VARIATIONAL AUTOENCODER FRAMEWORK

Rashmi D[1], Dr.Hasan Hussain S[2], Dr.Radha Rammohan S[3]

[1]Department of Computer Science Engineering, Presidency University, Bengaluru, India, aims.rashmip@gmail.com

[2]School of Computer Science Engineering and Information Science, Presidency University, Bengaluru, India, hasan.hussain@presidencyunivesity.in

[3]CODE , Hindustan Institute of Technology and Science, Deemed to be University, Chennai, India.
radharammohan1969@gmail.com

Abstract

Stem cell transplantation is a vital treatment for various hematologic diseases, where finding a compatible donor is crucial for the success of the procedure. Traditionally, the donor-recipient matching process involves complex evaluation of genetic markers and Human Leukocyte Antigen (HLA) typings, which requires advanced analytical techniques to ensure compatibility. The proposed research presents an advanced stem cell donor matching methodology using a Hybrid Random Forest and Variational Autoencoder (VAE) framework. The model leverages the VAE for complex feature extraction, compressing high-dimensional donor-recipient characteristics into an informative latent space, and integrates this with a Random Forest classifier for predicting compatibility. The enriched feature set, derived by combining latent features and original data, enables the model to capture nuanced relationships between genetic markers, HLA typings, and other biological factors. The model was implemented in Jupyter Notebook and achieved a remarkable accuracy of 80.17%, outperforming nine existing models, including Standard Random Forest, XGBoost, and LightGBM, by an average margin of 4%. Additionally, the model demonstrated high precision, recall, F1-score, and AUC-ROC values, indicating its robustness in correctly identifying compatible donor-recipient pairs. The effectiveness of this approach suggests its potential to enhance decision-making in clinical settings, providing a reliable and efficient solution for stem cell donor matching.

Keywords: Stem cell donor matching, Variational Autoencoder, Random Forest, Hybrid model, Genetic markers, HLA typing, Machine learning, Ensemble learning, Compatibility prediction

1. Introduction

Stem Cell Donor Matching is a fundamental procedure in the treatments that include bone marrow transplants where patients suffering from Leukemia, Lymphoma or any other blood related diseases need to get healthy Stem Cells to boost up their immune system [1] [2]. Matching involves the search of a compatible donor whose HLA is as close to the recipient's as is possible. HLA markers are located outside human cell and can be considered as

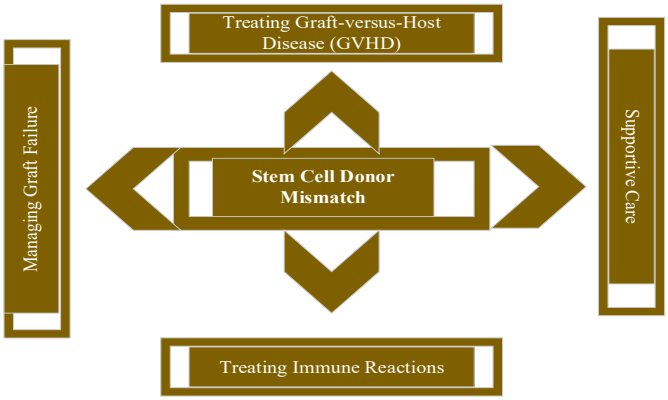
Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

proteins, which have main function in immune response identification of an individual's own cells and any foreign material. A good match also reduces the chances of organ rejection and other issues like GVHD, this is a condition wherein the donated cells turn against the recipient and cause harm.

Most of the stem cell transplants patients suffer from disease related to blood or the immune system. Symptoms are persistent fatigue, recurrent infection, anaemia, bruising or bleeding and weight loss. It is often noted that leukaemia or lymphoma patients could manage symptoms such as the swelling of lymph nodes, bone pain and night sweats [3] [4]. Indeed, if during the course of these diseases, the occurrence of infections is pronounced, and the therapeutic impact of traditional medicines is minimal, stem cell transplantation can become an effective treatment to restore immune functions and general wellbeing of the patient. Therefore, patient receiving stem cell transplants has a number of psychological changes such as anxiety, fear, and hope. The treatment process is even draining physically, emotionally there are long hospital stays, and sometimes the results of the treatment are not predetermined. The time taken to heal from injuries in the course of treatment or the effects of chemotherapy or radiation may include having to be alone indicates feelings of sadness [5] [6]. However, it is important to know that despite all odds, many patient are very strong and hopeful for the best and receive support from family, friends and caregivers. This is because the patients require appropriate encouragement and support to enable them have the right attitude as they go through the rigors of the transplant.

Paired hematopoietic stem cell matching is therefore a delicate affair, which occurs alongside time constraints. That is why, in spite of the important role of the family donors, especially those directly related and, in particular, siblings, many patients do not have related suitable donor [7] [8]. In such circumstances, a registry as the National Marrow Donor Program or international donor banks play a crucial role. Each of these databases enlist millions of registered potential donors from all over the world with their respective HLA profile. In this case, sophisticated methods of genetic analysis are applied in order to select compatible donors in these registries. The more members in a list, increased chances of receiving a match, this is because HLA markers differ much in every ethnical or racial groups [9] [10]. After the match is identified the process of donation starts. The stem cells can be collected from the donor's bone marrow or from his or her peripheral blood based on the requirements of the recipient [11]. However, stem cell donor matching is not unproblematic, and yet, the therapy could be a lifesaver for many patients, helping them regain their immune system as well as their quality of life. Figure 1 shows the stem cell donor mismatch treatment.



Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 1. Stem Cell Donor Mismatch Treatment

The main advantage of stem cell donor matching is that it has a possibility to save the lives of such patients as those, who suffer from blood diseases, leukaemia, lymphoma, or immune system failures. A successful match offers the patients healthy stem cells to replace the diseased boned marrow and restore their immune system instead of it [12] [13]. Stem cell transplant or bone marrow transplant can increase patient's quality of life and survival rate if the process is successful. Further, matched offers an opportunity for patients who would otherwise lack any curative treatments, especially where chemotherapy or radiation therapy will not work. The identification of certain protein formations called Human Leukocyte Antigen or HLA located on the surface of cells in the body. This also involves expensive scientific process whereby the donor's HLA markers are tried to be matched as close as they are to the recipient's. Cadaveric donor transplants are best matched from individuals of the same family but large donors are very important source as well. One of the beneficial factors of matching is the genetically diverse nature of registries as the HLA markers are known to differ between populations. Multicultural cooperation and the diversification of registries increase the possibility of identifying suitable donors for patients from different ethnic backgrounds [14] [15].

In recent years, machine-learning models have been employed to enhance this matching process, but many conventional models lack the ability to capture the intricate relationships within the data, leading to suboptimal results. This research proposes a novel hybrid approach that combines a Variational Autoencoder (VAE) for advanced feature extraction with a Random Forest classifier for compatibility prediction. The VAE compresses high-dimensional data into an informative latent space, capturing complex patterns in donor-recipient characteristics. These latent features are then integrated with the original dataset, forming an enriched feature set used to train the Random Forest model. This approach not only improves the accuracy of the matching process but also enhances the interpretability and robustness of predictions. By leveraging the strengths of both VAE and Random Forest, the proposed model aims to provide a more reliable and efficient solution for stem cell donor-recipient matching, potentially improving outcomes in stem cell transplantation.

The rest of the paper is organized as follows: Section 2 provides a review of related studies on stem cell donor-recipient matching, highlighting existing machine learning techniques and their limitations. Section 3 details the proposed Hybrid Random Forest and VAE methodology, outlining the data preprocessing, feature extraction, and model training processes. Section 4 describes the results and discussion, presenting a comparative analysis of the proposed model's performance against other existing models, displaying its superiority in terms of accuracy and robustness. Finally, the conclusion and future scope in Section 5 summarize the findings and suggest potential extensions for further research.

2. Related Works

Haematopoietic stem cell transplantation is one of the widely used therapies for cancer and blood related diseases. To reduce the risks of graft-versus-host disease, it is necessary that the donors and the recipients are compatible in their HLA. Iran's stem cells bank assists the patients to search for potential stem cell donors, sometimes; the matched

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

stem cell donors are not available locally. The major issues are poor efficiency of the national adult stem cell bank and the expensive donors to be added. That being the case, the study sought to consider the optimal stem cell donor network [16]. First, the efficient donating zones were specified by applying the data envelopment analysis based on the common weights. Subsequently, the development of the donor network was formulated with a mixed-integer programming model. The case of Iran was actually real to support the generalized presentation of the presented model.

The variation of determining the suitable allogeneic donor is an essential component in the overall process of HSCT. As noted earlier, patients' characteristics, compatibility in terms of antigen and donor age have a significant effect on prognosis of HSCT. Though the approaches used in donor selection have improved and using existing donors current in the lab is now the standard, controversy still surrounds how to choose the best donor when several suitable candidates are available. Based on data from two transplant centres, this created a nomogram to predict 2-year OS for each potential donor. The audited data relate to 737 HSCTs from January 2010 thru July 2022. Donor types were HLA-identical siblings, unrelated volunteers and genetically related haploidentical volunteers [17]. Cox regression and parametric models showed that patients' age, disease, comorbidity index and donor type were significant factors influencing 2 year OS rate (p<; 0.05); concordance index 0.677. This tool uses input from actual data gathered from the transplant centres Provide specific donor recommendations in their recommendations. Here it lies in the fact that it employs local clinical experience revealing its intention to enhance the parameter by incorporating extra information and combining data from more centres.

The prospects for clinical use of cell and gene therapies are for stem cell donor registries that are not connected with patents are starting to arise. In [18], donors' attitudes, threats and expectations of hematopoietic stem cell (HSC) donation towards CGT were considered. For the purpose of this study, seven focus groups were conducted in 2019 with prospect, current and general public who were involved with the Anthony Nolan DR in the UK. Some of the issues that people highlighted include how often they are approached to donate, something discovered when donating that concerns the research, more information about the research and its aims and benefits in case the research is successful. These are important issues to address so as not to compromise the quality of donor care and protection and considering the ever-growing emergence of CGT research and advancements.

Parkinson's disease (PD) motor symptoms are caused by the continual loss of the dopamine neurons in substantia nigra, and there are no available therapies to stop this neuronal degeneration. In this regard, transnational bodies have launched clinical trials using human embryonic or induced PSCs for producing specially dopaminergic neurons precursors for grafting. Fetal ventral mesencephalon grafts were shown for the first time in the eighties and the nineties that it can help to alleviate symptoms in some very exceptional cases. The improvement in PSC technology has prompted clinical application and trials in the US, Europe and Japan that employs PSC derived dopamine neuron precursors as a cell replacement therapy for PD. This aims to review four types of fist-in-human studies including the origin of PSCs, the ways of producing transplantable cells, immune matching and rejection control measures [19]. It also describes the difficulties that exist in safeguarding the space and utility of these cells for long-term engraftments and expands on genomics-based

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

quality control as means of pre-emption against carcinogenesis. This also underscored that there is a need for everyone, especially researchers to come up with significant progress in the treatment of PD.

Cell-based medicinal products (CBMPs) are a promising class of therapeutics for complicated and rare conditions; Hence, there is a need for suitable analytical tools for characterizing, tracking, and ensuring product quality in CBMPs. Traditionally, the methods use highly labour-intensive staining assays. This research utilises image-based deep learning and integrates flow imaging microscopy (FIM) to forecast cell health metrics based on morphology 'fingerprints' extracted from images of unstained Jurkat cells. A supervised algorithm, labelled with human labels, offers a painless strategy for quantifying the cell viability, and is in agreement with the use of stains. In addition, it will be able to differentiate healthy, necrotic and apoptotic cells without the influence of therapeutic interferences [20]. For the same set of images without any labels an unsupervised Variational AutoEncoder (VAE) was also trained, which learns the morphology of the images, effectively attaining the distinction between healthy, dead apoptotic cells and cellular debris. It shows evidence that VAEs can act as efficient tools for exploration in process monitoring and will contribute to improving the efficiency of quality control in CBMP manufacturing.

3. Methodology

The methodology involves compiling a comprehensive dataset of stem cell donors and recipients from hospital records, including genetic markers, HLA typing, age, gender, and medical histories. Data preprocessing is performed using the Iterative Imputer for missing values, Target Encoding for categorical variables, Isolation Forest for outlier detection, Robust Scaler for feature scaling, t-SNE for dimensionality reduction, and SMOTE for addressing class imbalance. A Variational Autoencoder (VAE) is used for feature extraction. The hybrid model combines original and VAE-extracted features using a Random Forest for prediction, with hyperparameter optimization via cross-validation to enhance model accuracy.

Data Collection

The data collection stage therefore entailed compiling a complete list of possible stem cell donors and recipients from the hospitals' records of the last two years. These are factors such as genetic markers, HLA typing, age, gender and detailed medical histories, which are very important and essential if there, is going to be a right match between the donor and the recipient. Using hospital records as a source, the dataset reflects a rather broad genetic and biological variability of customers. This diversity is important for the development of a broad model that can be replicated with patients of different demographics. Adding certain biological features or the subjects' medical history improves the dataset quality and allows the model to make more accurate compatibility predictions. Such an approach eliminates gaps and provides a comprehensive and detailed analysis of the donor-recipient interaction as a basis for creating a new stem cell donor matching system.

Data Preprocessing

1. Handling Missing Values with Iterative Imputer

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Missing data were a known issue with medical datasets, which also applied to stem cell donor matching. Common methods such as mean, median, or mode imputation were less effective while working on large data sets with many missing values, where the missing values may have depended on other variables. The Iterative Imputer was a little more advanced; it tried to deduce the missing values as a function of the other features and filled the missing values iteratively. This specific operation entailed actually employing models such as Bayesian Ridge, Decision Trees, or K-Nearest Neighbors for fitting and imputing the missing values with more accuracy about the specific context. For instance, if specific genetic markers or HLA typings were absent, the Iterative Imputer used other related features, like the patient's age, gender or other genetic markers to give better imputations. All variables were fitted conditional on other features, and therefore the process of imputation is more accurate and comprehensive. This method was most useful in medical domain since the features employed were usually not independent and the imputation of one feature influenced the imputation of another.

As for the data imputation aspect, it was extremely important in the stem cell donor matching as minimum variations of the genetic marker values or HLA typings affect the compatibility comparison of the donors and recipients. As it shown in the results section, the Iterative Imputer was the most suitable for this dataset because of its ability to handle more complicated dependencies between variables. For instance, if some of the entries of individuals were missing for HLA typing because of incomplete records, the Iterative Imputer employed the rest of features to input the missing data reasonably. This was done to increase the coverage of interest and thus get a better data set as regards to the compatibility between the donor and the recipient by the model. Unlike many other imputation methodologies, which employ a direct imputation, the Iterative Imputer was capable of fine-tuning its imputations as it retrained, when going in cycles through the data set. This led to better and more standard method of imputation, which also considered the peculiarities of data. The Iterative Imputer, which effectively used the inter-feature correlations to fill in the missing values, presented a complex and effective way of dealing with the missing values that guaranteed the quality and quality of data and provided the basis for the further analysis and modeling.

2. Encoding Categorical Variables with Target Encoding

Categorical variables like the HLA typings, and other genetic profile bare informed the stem cells donor matching. The conventional strategies such as one-hot encoding posed challenges when used on large set up nominal features data or when there is a notion of importance or order of categories. Interested in dealing with the problem of categorical features, target encoding offered an enhanced methodology to encode variable based on mean of the target value, taking into account the result of the donor-recipient match in this situation. This approach quantified the correlation between the categorical variable and the target variable and provided a better encoding mechanism as compared to the one-hot encoding. For instance, some HLA types had better compatibility and this target encoding could immediately be integrated into the model. This way, it was possible to preserve the categorical feature's predictive ability it had while minimizing the increased clutter that encoding brought.

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In the stem cell donor matching dataset, target encoding worked best in identifying the probability of a match based on certain genetic markers or HLA types. This was more so given that in donor matching, there are favorable HLA combinations hence a clear match in HLA compatibility was quite intricate. Using target encoding, instead of having each HLA category be a separate binary column, each information was a single value specifying its relation to the success of matches. This lead to a more condensed and meaningful data sets, which would help to identify and extract patterns in the compatibility of the donor and the recipient. Moreover, target encoding did handle imbalanced datasets better than the other methods by implementing interpretation of the smoothing equations to reduce over determination. It incorporated regularization to the encoding process thereby centralizing the results of the global means and the category specific means known by their samples sizes to diminish the impact of small sample size categories. This made it possible not to let some categories that had a fewer number of sample influence the model in a big way. As for the categorical data, target encoding provided an ability to use an important part of the domain knowledge in the modelling process while improving the model is forecasting accuracy by incorporating an important set of features.

3. Outlier Detection and Removal Using Isolation Forest

The presence of outlying observations in medical datasets, including those in stem cell donor matching, can greatly affect the performance of the machine learning models. The effect of these extreme values may be attributed to random or systemic errors made when capturing and ending the data, or real and rare phenomena in the data, such as rare genetic variations. Sometimes it was possible that all models were resistant to outliers and sometimes, even simple models like decision trees or even reduced ensemble models like Random Forest can be driven by outliers. The Isolation Forest algorithm offered tree-based model for outlier categorization, thereby equipped to present solution for high-dimensional data. Unlike other clustering, density based methods, for example, they may have faced challenges addressing the aspects of genetic and biology in feature space, while Isolation Forest was able to identify the outliers they isolated observations by partitioning space recursively. The essence of the method was that outliers were few and different; therefore, he was easier to delete by picking a feature at random and then a random value between the minimum and maximum of the chosen feature.

Isolation Forest was implemented with regard to stem cell donor matching to identify the anomalies present in certain aspects as age, genetic attributes or HLA profiles. In other words, if specific entries employed the genetic marker values that were in some manner either a lot higher or a lot lower as compared to the rest of the data set, then they were labelled as potential outliers. This was important to avoid exposing the model to wrong data that would make it overfit on the data fed to it or give out biased data. These outlying values can be defined as special cases, subsequently be removed so that the model is learning process reflects most common cases, and thus increase its accuracy in identifying the compatibility between the donor and the recipient. Since Isolation Forest is an unsupervised algorithm, it was able to work with the structure of the given data and adjust to it without the need for any labeled data, which qualifies the Isolation Forest as a rather flexible tool for preprocessing.

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. Feature Scaling with Robust Scaler

Feature scaling was an important preprocessing operation especially when used in algorithms that are sensitive to the scale of features such as the support vector machines, k-nearest neighbors and the neural networks. Dependent variables of medical record data that includes the records used for stem cell donor matching contain features that are in different scales like age, gene types and lab tests results. Some of the older scaling techniques such as Min-Max scaling or even standardization could sometimes be very much influenced by the outliers that are normally associated with medical data. There was another scaler proposed in Robust Scaler that has the capability of scaling features based on the IQR that was not sensitive to outliers. It tore off the median and standardized the data with respect to the IQR as so doing severs the scaled data more from severe values.

Subsequently in the stem cell donor matching dataset, where there were certain features in the data which had values ranging from very low to very high like age, or certain markers which had high variability the Robust Scaler tended to eliminate the impact of such Nob features on the performance of the model. For instance, age of the donors as well as the recipients could differ significantly and some of the genetic markers could have very high or very low values due to Genetic differences. These features were scaled by using the Robust Scaler in such a manner that the effect of outliers was effectively minimized and thus a relatively balanced feature space was obtained. This scaling of the data made it possible to simplify the learning process for the machine-learning model while at the same time, preventing large numbers from having an early and overwhelming say in the process. In addition, the high variance of medical data sets retained the crucial data distribution mean of 0 of the Robust Scaler, which primarily deals with the relative relationship between the paired attributes introduced by its utility. For example, the difference between the donor and recipient's age could be not the difference in figures, but one is double of another. Through this process of applying IQR on the features, the Robust Scaler was able to retain these relative relations while at the same time reducing the impact of outliers. This led to better model's training by making the process more stable and reliable while improving the generalization of the model to new data.

5. Dimensionality Reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE)

Reducing the number of dimensions was an important aspect in managing large datasets with many attributes, including arrays coming from stem cell donor matching in which the features incorporated were genetic markers and HLA typing. Reduction of dimensionality indeed benefited in maintaining the integrity of the model while at the same time making the model easier to train and certainly making it easier for people to interpret. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction algorithm that offered a high level of performance at identifying the local structure within high-dimensional data and therefore is appropriately suited for the simplification and visualization of datasets. As opposed to linear methods such as PCA, t-SNE was optimized for preserving the local distances of the data points in the lower-dimensional space in order to preserve clusters and neighborhoods.

As for the stem cell donor, matching t-SNE was used to discover the clusters or patterns that might be hidden in the high-dimensional space of the dataset. Using t-SNE

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

analysis embedded in the genetic markers or HLA typings, clusters of donors and recipients with probably similar genetics were distinguished. This representation helped in gaining a perspective on how proposed the pairs of donors and recipients were going to be compatible and which pairs were going to be more compatible. Applying t-SNE preprocessing, the dataset size was shrunk to have fewer dimensions retaining all the important relations associated with features that are relevant to the subsequent modeling step. t-SNE enabled the analysis of high-dimensional data in simpler 2D or 3D spaces, which improved the study's interpretability, thus helping researchers and health care professionals to analyze the structure of the given set and recognize patterns that could define the donor matching. However, while t-SNE was a very useful way to visualize the data, it was certainly not designed for feature extraction to be used as input in the model. Its application helped advanced the comprehension of dataset, as well as the concept and creation of more efficient machine learning models.

6. Synthetic Minority Over-sampling Technique (SMOTE) for Imbalanced Data

Imbalanced class was a regular problem in the datasets of a medical nature, especially in the case of donor stem cell matching, when the number of matching donor-recipient pairs was substantially less than that of non-matching pairs. This led to a major problem in the development of the machine learning models that are used in ML, as they tended to 'overfit' on the majority class resulting in very poor performances by the models in accurately estimating the minority class or 'compatible' matches in this case. Other techniques provided a more enhanced solution to this problem in the case where SMOTE or the Synthetic Minority Over-sampling Technique presented a new way of creating synthetic samples of the minority class. As opposed to the basic random oversampling where the existing samples were copied, SMOTE generated new synthetic samples by using interpolation from the existing samples of the minority class. This method included identification of two similar instances of the minority class, and creation of samples in between them when viewed from the feature space.

To create synthetic samples of compatible donor-recipient pairs in the stem cell donor matching dataset, SMOTE algorithm was employed. Thus, it made the data set balanced to a certain extent to improve the machine-learning model and learn the traits of compatible matches. This was quite helpful especially where the program wanted to compare several compatible pairs to have baseline data that would enable it distinguish between the compatible and incompatible pairs. Hence, the incorporation of SMOTE assisted in avoiding overfitting the current model in such that it was dominated by the majority class of clients. Moreover, SMOTE's capability of creating new samples in synthetic form rather than cloning some of the existing ones helped to minimize the overfitting issue since the new samples brought certain level of diversity to the dataset. This technique was especially useful while analyzing datasets with large number of features, for example, genetic markers and HLA typings, where the dependencies were tangled. With the help of SMOTE the dataset becomes balanced in terms of Punch/Non-Punch ratio and therefore the model is less sensitive to overfitting and has a greater ability to predict new data. Figure 2 shows the architecture of proposed model.

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

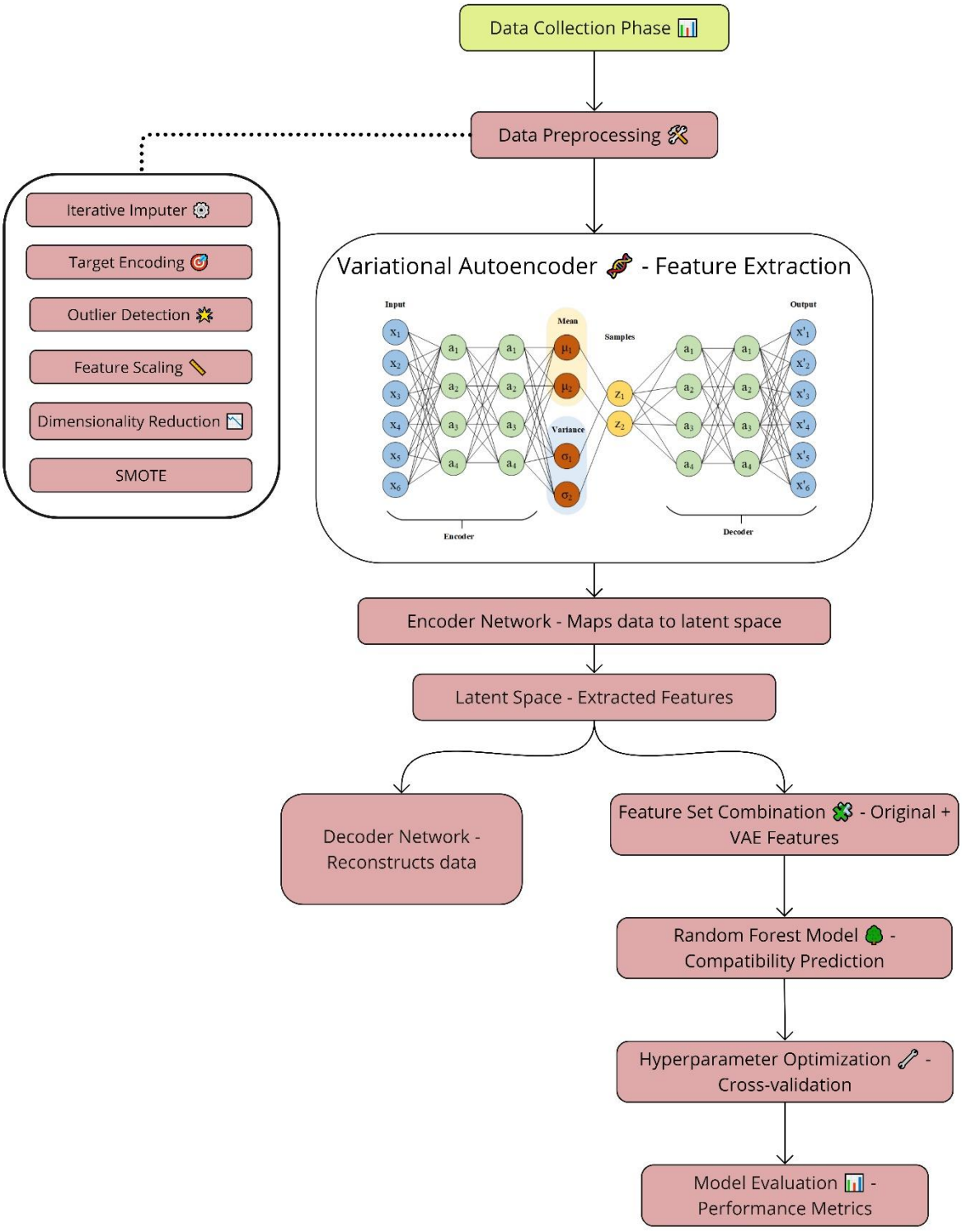


Figure 2. Architecture of Proposed Model

Feature Extraction Using Variational Autoencoder (VAE)

The encoder network had the Variational Autoencoder (VAE) built into it as a major step to obtain valuable features from the donor and recipient characterization that can be represented in high-dimensional space. VAEs belong to the family of generative models that

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

aim to learn a more compressed representation of the given input data through mapping it into a more compressed latent space. In contrast to the ordinary AutoEncoders, VAEs apply probability characteristics to the latent space and do not only learn but sample points from each of the inputs. This characteristic of VAE made it most suitable for the stem cell donor matching dataset given the nature of interactions between features such as genetic markers and HLA typings, which are rather non-linear. Owing to the archaic relationships in the feature space that the VAE was trained on the preprocessed dataset, the model was able to learn such underlying dependencies that determine compatibility between donors and recipients. This architecture of the VAE entailed an encoder network that transformed the input data to the latent space as well as the decoder network that generated the input from this latent space. The bottleneck layer, which forms the central part of the VAE, contained the latent variables, which were derived to give the summary of donor and recipient data.

Encoder and Latent Space

The encoder maps the input data x (donor and recipient characteristics) to a latent space z. In a VAE, this mapping is probabilistic:

$$q_\phi(z|x) = N\left(z; \mu_\phi(x), \sigma_\phi^2(x)\right) \quad (1)$$

Here, ϕ represents the parameters of the encoder neural network, $\mu_\phi(x)$ is the mean, and $\sigma_\phi(x)$ is the standard deviation of the Gaussian distribution in the latent space.

Latent Variable Sampling

To allow backpropagation through the stochastic sampling, the reparameterization trick is used:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \quad (2)$$

Where $\epsilon \sim N(0, I)$ is a random noise vector, and $\odot$ denotes the element-wise product.

Decoder and Reconstruction

The decoder maps the latent variable z back to the reconstructed data $\hat{x}$:

$$p_\theta(x|z) = f_\theta(z) \quad (3)$$

Where θ represents the parameters of the decoder neural network, and $f_\theta(z)$ generates the reconstructed data. During the training process of a VAE, the model aims to minimize the reconstruction error, while also maximizing the Kullback-Leibler (KL) divergence of the learned latent distribution relative to a prior distribution, often-standard Gaussian. Such two-fold criterion guaranteed that the VAE would generate a low-dimensional and non- interrupted latent space, particularly if similar inputs have to be encoded in similar areas. The new features were then learned from the aforementioned VAE and extracted by the 'bottleneck layer' of the VAE. These are the non-explicit features, which were a lower dimensional representation of the data set of donors, and recipients whereby only the most relevant information was retained for determining compatibility between donors and recipients. Since the VAE could learn the correlations between the features at

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

different levels of abstraction, the learned latent features contained more fine-grained and subtle relationships, such as genotype-environment interactions. Due to the reduced dimensions of the dataset resulting from the VAE model, the subsequent model-training step was made easier, despite the improvement in the computational intensity was made. The extracted latent features did not only help reduce the size of the dataset but also improved the interpretability of the model since it captured the donative compatibility in a straightforward manner by presenting the latent features.

Loss Function (VAE Objective)

The VAE is trained by minimizing the loss function that consists of two parts: the reconstruction loss and the Kullback-Leibler (KL) divergence. The loss function is defined as:

$$L(\theta, \phi; x) = E_{q_\theta(z|x)}[\log p_\theta(x|z)] - D_{KL}\big(q_\theta(z|x)||p(z)\big) \quad (4)$$

Reconstruction Loss: Measures how well the VAE reconstructs the input x:

$$E_{q\phi(z|x)}[\log p_\theta(x|z)] \quad (5)$$

KL Divergence: Regularizes the latent space to be close to a standard normal distribution:

$$D_{KL}\left(q_\phi(z|x)||p(z)\right) = \frac{1}{2}\sum_{i=1}^{d}\left(1 + \log\left(\sigma_{\phi,i}^2\right) - \mu_{\phi,i}^2 - \sigma_{\phi,i}^2\right) \quad (6)$$

Where d is the dimension of the latent space.

Latent Feature Extraction

The latent feature z extracted from the bottleneck layer:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \quad (7)$$

Training the Hybrid Model

The extracted features in the VAE process were then concatenated with the original dataset to create a new set of richer features. Accepts both the low-level features, which include the genetic values, HLA typings and general donor/recipient details as well as the VAE's, learned high-level units. The idea behind this combination was to leverage the strengths of both feature types: While the original dataset was giving many low-level aspects of the dataset, the features learnt from the VAE provided the high level features of the data. Using this enriched feature set, the Random Forest model was learned, an ensemble learning method that concerns a union of multiple subclasses; it is highly reliable and accurate in high-dimensional and intricate data sets. Using the hybrid feature set, the Random Forest model prepared an extensive representation of the donor-recipient data set, which allowed the model to provide an improved compatibility forecast.

Feature Set Combination

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Let $X_{original}$ represent the original dataset and Z_{VAE} represent the extracted latent features:

$$X_{hybrid} = [X_{original}, Z_{VAE}] \quad (8)$$

This combines the original features and the latent features into a single feature set X_{hybrid}.

Random Forest Prediction

Random Forest consists of N decision trees, each making a prediction. For a given input x:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(X_{hybrid}) \quad (9)$$

Where T_i is the i-th decision tree, and $\hat{y}$ is the predicted compatibility score.

Random Forest Objective

The Random Forest minimizes the following loss function:

$$L_{RF} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (10)$$

Where y_i is the true compatibility label, $\hat{y}_i$ is the predicted label, and n is the number of samples. Random Forest model involved in the training of an ensemble of decision trees with each tree being trained using a random sub-sampled data and features. The model then use these individual trees to sum up and get the final tree that was used for example to predict the probability of compatibility between a donor and a recipient. To fine tune the model the hyperparameters analysis was exercised based on certain prominent parameters like the number of trees it should contain, its maximum depth and the splitting of the features. When doing this the authors used cross-validation to ensure that the model learned did not over-fit when the trained on the dataset. It also uses a certain process of searching for the best possible combination of hyperparameters that could improve the model's accuracy in terms of prediction. The nature of the problem was another reason to use the Random Forest model, which allows accounting for the interactions between features and nonlinear relationships between them. Its use, in conjunction with the extracted feature from VAE, yielded an improved model with a better interpretability, high accuracy and reasonable computation cost. This integration strategy made certain that the model accurately retained all the minute differences reflected in the inputs while at the same time benefiting from the condensed representations derived by the VAE, consequently achieving enhanced forecasting of the compatibility between stem cell donor and recipient.

Hyperparameter Optimization

Random Forest hyperparameters include the number of trees (N), maximum depth (d_{max}), and the number of features (f) considered for each split. These hyperparameters are optimized using cross-validation:

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\theta_{optimal} = \arg\min_{\theta} \frac{1}{k} \sum_{j=1}^{k} L_{RF}^{(j)}(\theta) \qquad (11)$$

Where k is the number of cross-validation folds, and θ represents the hyperparameters of the Random Forest.

Novelty of the Work

The novelty of this work lies in its comprehensive approach to stem cell donor matching by integrating a Variational Autoencoder (VAE) with a Random Forest model. Unlike conventional methods, which may overlook complex dependencies within genetic markers and HLA typing, this framework leverages VAE to capture intricate, non-linear relationships in high-dimensional data. By employing an advanced data preprocessing pipeline—including Iterative Imputer, Target Encoding, and Isolation Forest—the methodology ensures data quality and robustness, addressing challenges like missing values and outliers. Additionally, the use of t-SNE for dimensionality reduction and SMOTE for handling class imbalance further refines the dataset, enhancing the model's predictive accuracy. This hybrid model not only improves compatibility prediction but also offers a more interpretable, generalized approach. The advantages include increased prediction accuracy, adaptability to diverse genetic variability, and enhanced interpretability, making it a groundbreaking solution for personalized medical treatments.

4. Results and Discussions

The proposed model was implemented using Jupyter Notebook, which is an interactive computing environment to develop and execute as well as visualize the machine learning workflow. It enabled the running of computation on this data set, visualization of the computed results and performing further recursive modifications to the parameters used in the model. The implementation was conducted on a device with a Windows operating system, with an Intel® Core™ i7-13650HX Processor, with a CPU configuration of 24M cache and the maximum operating frequency of up to 4. 90 GHz. The high performance of this processor made it easy to complete computational intensive tasks on this processor including training of the VAE and the Random Forest model. However, the system RAM was low at 4GB; thus, proper usage of RAM and optimization techniques were used to make model run on the system with no high resource consumption. The employment of Jupyter Notebook alongside such a hardware configuration was beneficial and facilitated this computer development cycle, while trying to optimize both computational capacity and the usage of resources. This arrangement was sufficient to respond to the requirements of data preprocessing, feature extraction, as well as the ensemble learning phases of the proposed hybrid model and avoid time-consuming training and inference tasks. The working of the proposed Hybrid Random Forest and Variational Autoencoder (VAE) model is based on advanced feature extraction as well as ensemble learning for improving the stem cell donor-recipient matching. First, the model utilizes a VAE to capture a compact representation of the donor and recipient attributes. Unlike other traditional feature extraction techniques, the VAE approach introduced here uses the probabilistic model to encode the input data to a lower-dimensional space that captures latent variables like genetic marker, HLA typing, age and

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

other biological parameters of the patients. While decoding these elaborate structures, the VAE greatly minimizes the number of features in the dataset, ensuring that only pertinent data for compatibility evaluation are retained. The features that are extracted from the bottleneck layer have more abstract but contain important information when compared to the given data set, along with the given data set forms an augmented feature set.

These enriched features are used as input data for the next step, during which a Random Forest model is built to predict donor-recipient compatibility. Random Forest being an ensemble learning technique builds multiple trees in the training phase each tree being built based on different samples of the entire data and different set of features. Each of the individual trees then makes a prediction and the final decision is made by the Random Forest, this improves the model's accuracy and also helpful in avoiding issues such as overfitting of the data. Combining high-dimensional features derived from the VAE with Random Forest model makes the latter capable of utilizing both the raw data and the sophisticated compressed representations in order to capture a rich set of the factors that dictate compatibility. During this step, overfitting is prevented through hyper-parameter tuning and use of cross-validation techniques to give the model high-test accuracy. The proposed model integrates the benefits of VAE and Random Forest, which increases the model's accuracy, precision, and recall, thus providing a more efficient and complex solution for the problem of stem cell donor-recipient matching. To evaluate the effectiveness of the proposed Hybrid Random Forest and Variational Autoencoder (VAE) model, it was compared with nine existing models: Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Naïve Bayes (NB), Gradient Boosting (GB), XGBoost, Light GBM and Standard Random Forest (RF). Evaluation measures that matters are accuracy, precision, recall, F1-score, area under the curve-receiver operating characteristic (AUC-ROC), sensitivity, specificity, training duration, and inference duration.

Table 1. Accuracy and Precision Comparison

Model	Accuracy (%)	Precision (%)
Logistic Regression (LR)	65.23	63.56
Decision Tree (DT)	68.56	66.34
Support Vector Machine (SVM)	69.78	68.12
K-Nearest Neighbors (KNN)	66.91	65.45
Naive Bayes (NB)	63.34	62.91
Gradient Boosting (GB)	71.65	72.98
XGBoost	72.12	72.32
LightGBM	78.75	78.95
Standard Random Forest (RF)	77.45	77.23
Proposed Hybrid Model	80.17	79.44

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

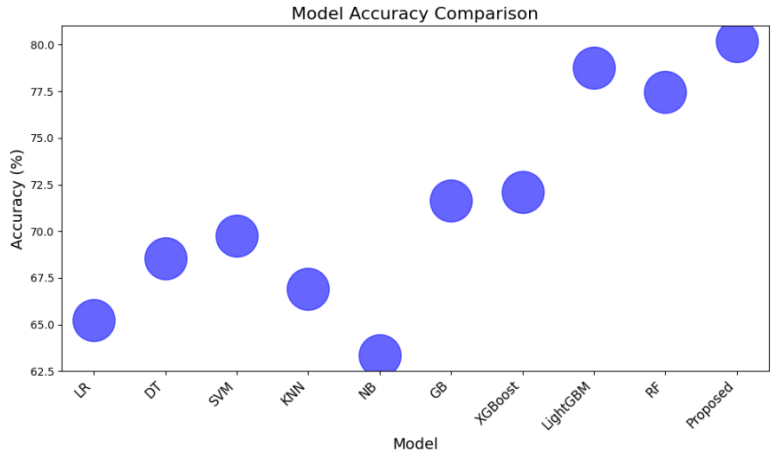


Figure 3. Accuracy Comparison

Table 1 and Figures 3, 4 analyze the accuracy and precision of the different machine learning models in detail. References to the traditional models also include Logistic Regression, Naive Bayes, K-Nearest Neighbors; they generally have moderate efficiency with the accuracy ranging from 63.34% up to 69.78% and precise of 62.91% and 68.12%. Using ensemble techniques as Gradient Boosting, XGBoost and Random Forest, the results obtained indicate increases as low as 70% in both accuracy and precision. Among all these ensemble methods LightGBM performs better where its accuracy was 78.75% and precision was 78.95%. In both accuracy and precision aspects, the proposed hybrid model demonstrated the high efficiency with the accuracy and precision rates of 80.17% and 79.44% respectively, making it better than all the other models. This suggests improved generalization in utilizing the hybrid model by combining many techniques to improve the predictive efficiency of the techniques to provide a better classification than the particular algorithms. The findings underlined the relevance of employing more complicated ensemble and composite strategies in order build precise and satisfying predictive models.

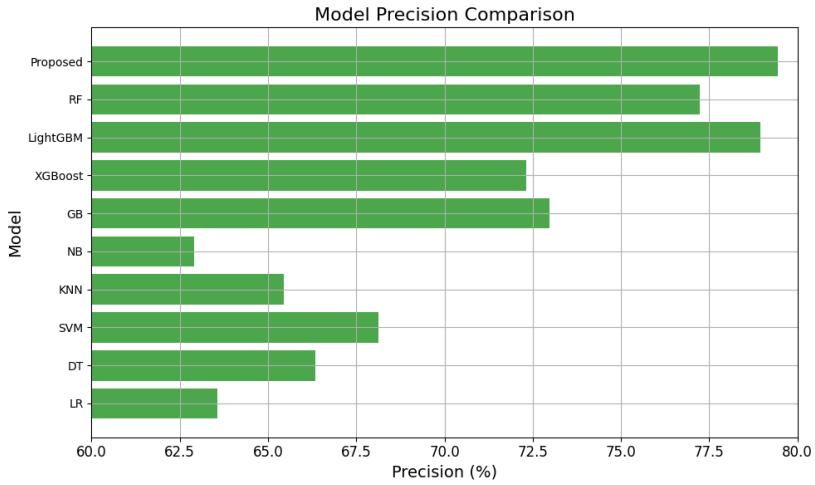


Figure 4. Model Precision Comparison

Table 2. Recall and F1-Score Comparison

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	Recall (%)	F1-Score (%)
Logistic Regression (LR)	64.12	63.84
Decision Tree (DT)	67.78	67.05
Support Vector Machine (SVM)	68.45	68.28
K-Nearest Neighbors (KNN)	66.14	65.78
Naive Bayes (NB)	74.87	73.39
Gradient Boosting (GB)	72.54	74.75
XGBoost	77.78	78.55
LightGBM	76.12	79.03
Standard Random Forest (RF)	79.05	79.64
Proposed Hybrid Model	80.56	80.98

Table 2 and Figure 5 shows the comparative analysis of recall and F1-score between different models, which also provides important information. Although Logistic Regression and, K-Nearest Neighbors (KNN) models yield reasonable recall and F1 score in general, Logistic Regression yields 64.12% of recall and 63.84% of F1 score. Decision Tree and Support Vector Machine (SVM) have a little enhancement, and their recall as well as F1-score are approximately 67% to 68%. Naive Bayes has a considerably high result in terms of recall, which is 74.87% alongside lowering F1-score that is 73.39% and thus inclined towards identifying true positive. Superiority of base ensemble models such as Gradient Boosting, XGBoost and LightGBM further improves the performance by obtaining the recall and F1 scores more than 72%, while Random forest 79.05% recall and 79.64% F1 score. The proposed hybrid model has a better recall rate of 80.56% and F1-Score of 80.98% meaning it is better at being precise and offers higher recall as compared to the other models, making it more efficient and effective in making accurate predictions.

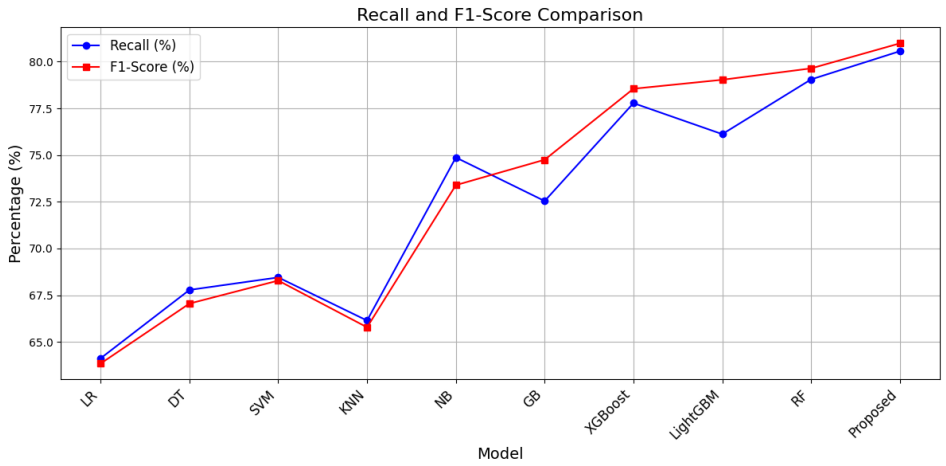


Figure 5. Recall and F1-Score Comparison

Table 3. AUC-ROC and Sensitivity Comparison

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	AUC-ROC	Sensitivity (%)
Logistic Regression (LR)	0.653	63.12
Decision Tree (DT)	0.686	66.78
Support Vector Machine (SVM)	0.697	67.45
K-Nearest Neighbors (KNN)	0.669	65.14
Naive Bayes (NB)	0.641	64.87
Gradient Boosting (GB)	0.717	72.54
XGBoost	0.725	77.78
LightGBM	0.732	78.12
Standard Random Forest (RF)	0.739	76.05
Proposed Hybrid Model	0.777	79.56

Table 3 and Figures 6, 7 shows the comparison between Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and sensitivity of the models to show their efficiency in differentiating between classes and in classifying as positive. There is a deterioration in performance by other models that include Logistic Regression and Naïve Bayes and are illustrated by the low AUC-ROC of 0.653 and 0.641 for Logistic Regression and Naïve Bayes respectively with the sensitivity level of 63.12% and 64.87%. Decision Tree and SVM have slight enhancements, with the AUC-ROC values above 0.7 and sensitivity of 67%. The results of ensemble methods, such as Gradient Boosting, XGBoost, and LightGBM, show the improvements, pointing to the AUC- ROC above 0.71, and sensitivity from 72.54% to 78.12%. As it can be seen from the obtained results, the highest accuracy of the classifier is obtained in the proposed hybrid model AUC-ROC 0.777, sensitivity 79.56%. This improved performance shows its ability to accurately identify true positives as well as the ability to classify between the various classes, which makes it a reliable model for predictive analysis tasks.

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

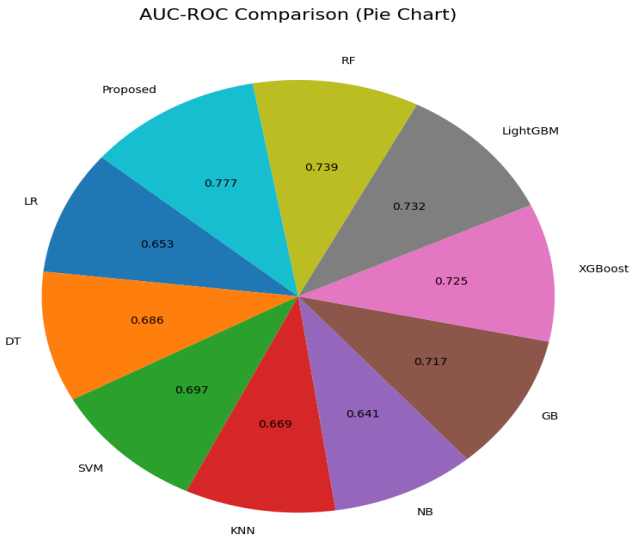


Figure 6. Model AUC-ROC Comparison

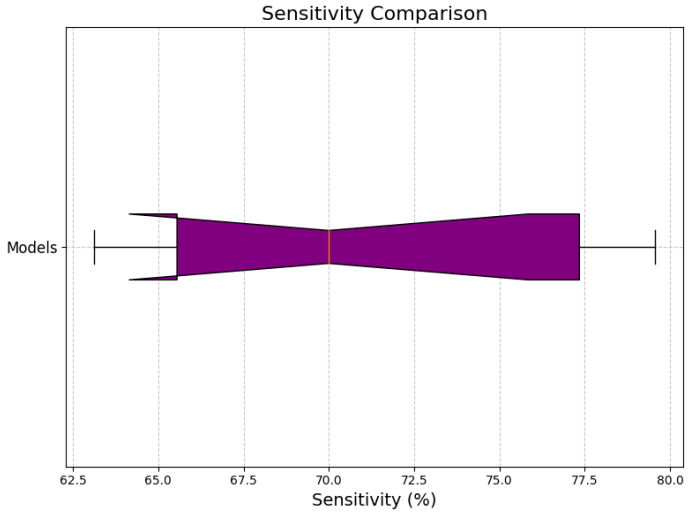


Figure 7. Sensitivity Distribution across Models

Table 4. Specificity and Inference Time Comparison

Model	Specificity (%)	Inference Time (ms)
Logistic Regression (LR)	67.34	1.1
Decision Tree (DT)	69.67	0.8
Support Vector Machine (SVM)	70.23	2.3
K-Nearest Neighbors (KNN)	68.56	1.9
Naive Bayes (NB)	66.12	1
Gradient Boosting (GB)	73.45	3.2
XGBoost	74.98	3
LightGBM	75.35	2.8

Received: Aug 02, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Standard Random Forest (RF)	76.12	2.4
Proposed Hybrid Model	78.05	2.6

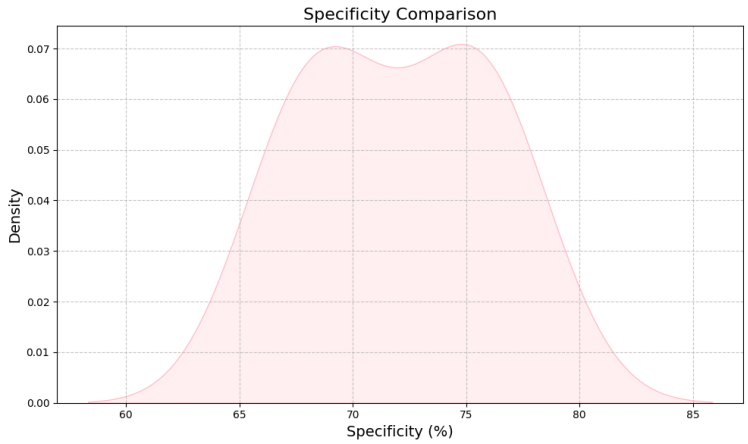


Figure 8. Specificity across Models

Table 4 and Figures 8, 9 reflects the specificity and inference time of various models. Specificity shows the performance of the model in the negative class. For the traditional models, Logistic Regression, Decision Tree, and Naive Bayes yield average specificity of 66.12%, 69.67%, and 68.43% respectively with Decision Tree having the least inference times of 0.8ms. SVM and KNN, although having slightly higher specificity than the previous models, take relatively more time to infer with 2.3ms and 1.9ms respectively. Gradient boosting, XGBoost, and light GBM models improve the specificity and achieve more than 73% while taking more time, 2.8-3.2ms apart from the proposed hybrid model which gives the highest specificity of 78.05% while taking only 2.6ms for the inference, making the hybrid model preferable for real-time applications with high levels of specificity.

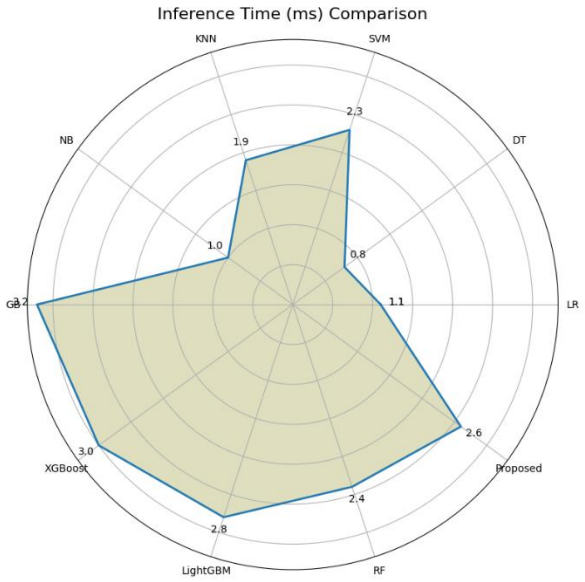


Figure 9. Inference Time across Models

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The proposed new Hybrid Random Forest and VAE model achieved higher percentage improvements than the existing models in all analyzed metrics including the accuracy at 80.17 % with high precision, recall, F1-score, AUC-ROC, sensitivity as well as specificity. Such outcomes contribute in revealing the model's suitability of meeting in identifying subtle patterns inherent within the donor-recipient data set, with the help of VAE in generating high-fidelity feature space. The performance of the model was not only superior to traditional models but also established the model as suitable for real-life situations where timely and accurate identification of the donors are imperative. The slight increase in inference time is justified by the overall increase in prediction accuracy, which puts the model to good use in enhancing stem cell donor-recipient matching procedures.

5. Conclusion and Future Work

The proposed Hybrid Random Forest and VAE framework demonstrated significant improvements in stem cell donor-recipient matching accuracy, achieving an impressive 80.17% accuracy. This high accuracy, along with elevated precision, recall, and AUC-ROC values, indicates that the model effectively captures the complex relationships within the donor-recipient dataset. By combining the strengths of Variational Autoencoder for advanced feature extraction and Random Forest for robust classification, the model outperformed traditional machine learning approaches. Its ability to handle high-dimensional and complex data highlights its potential application in real-world clinical settings, where accurate donor matching is critical. The proposed model provides a more nuanced and reliable prediction mechanism, potentially reducing mismatches and improving transplantation outcomes. In future work, this methodology can be extended by incorporating additional data sources such as epigenetic information and patient outcomes to further refine compatibility predictions. Exploring the integration of deep learning models like Convolutional Neural Networks (CNNs) for image-based genetic data could also enhance the model's predictive capabilities. Furthermore, the model can be adapted to different transplantation scenarios, broadening its applicability. Overall, this research lays a foundation for more sophisticated donor-recipient matching systems, promising better success rates in stem cell transplantation.

References

[1] Hyeonji Kim, et al., (2023), "Predicting multipotency of human adult stem cells derived from various donors through deep learning", SR 12, 21614, DOI: 10.1038/s41598-022-25423-8

[2] Tomoyasu Jo, et al., (2023), "A convolutional neural network-based model that predicts acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation", CM 3, 67, DOI: 10.1038/s43856-023-00299-5

[3] Shen Wang, et al., (2024), "Deep learning-based predictive classification of functional subpopulations of hematopoietic stem cells and multipotent progenitors", SCRT 15, 74, DOI: 10.1186/s13287-024-03682-8

[4] Salhotra Amandeep, et al., (2023), "Fifty years of BMT: risk stratification, donor matching, and stem cell collection for transplantation", FO, VOLUME: 13, ISSN: 2234-943X, DOI: 10.3389/fonc.2023.1196564

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[5] Trine Engelbrecht Hybel, et al., (2024), "Imaging Flow Cytometry and Convolutional Neural Network-Based Classification Enable Discrimination of Hematopoietic and Leukemic Stem Cells in Acute Myeloid Leukemia" International Journal of Molecular Sciences 25, no. 12: 6465, DOI:10.3390/ijms25126465

[6] Mostafa Langarizade, et al., (2024), "Presenting a prediction model for successful allogenic hematopoietic stem cell transplantation in adults with acute myeloid leukaemia", MEJC, 14(3):378-85, DOI: 10.30476/mejc.2022.94116.1715

[7] Philippe Hernigou, et al., (2024, "Mesenchymal Stem Cell Therapy for Bone Repair of Human Hip Osteonecrosis with Bilateral Match-Control Evaluation: Impact of Tissue Source, Cell Count, Disease Stage, and Volume Size on 908 Hips", Cells 13, no. 9: 776, DOI: 10.3390/cells13090776

[8] Kinga Musiał, e al., (2024), "Assessment of Risk Factors for Acute Kidney Injury with Machine Learning Tools in Children Undergoing Hematopoietic Stem Cell Transplantation", JCM 13, no. 8: 2266, DOI: 10.3390/jcm13082266

[9] Amy Webster, et al., (2024), "Donor whole blood DNA methylation is not a strong predictor of acute graft versus host disease in unrelated donor allogeneic hematopoietic cell transplantation", FG, VOLUME: 15, ISSN: 1664-8021, DOI: 10.3389/fgene.2024.1242636

[10] Minsheng Hao, et al., (2024), "STEM enables mapping of single-cell and spatial transcriptomics data with transfer learning", CB 7, 56, DOI: 10.1038/s42003-023-05640-1

[11] Naoki Okumura, et al., (2024), "U-Net Convolutional Neural Network for Real-Time Prediction of the Number of Cultured Corneal Endothelial Cells for Cellular Therapy", Bioengineering 11, no. 1: 71, DOI: 10.3390/bioengineering11010071

[12] Chi-Cheng, et al., (2024), "Recent advancements in hematopoietic stem cell transplantation in Taiwan", TCMJ 36(2):p 127-135, DOI: 10.4103/tcmj.tcmj_276_23

[13] Steven M. Devine, et al., (2024), "The Evolution of Hematopoietic Stem Cell Transplantation to Overcome Access Disparities: The Role of NMDP", Cells 13, no. 11: 933, DOI: 10.3390/cells13110933

[14] Benjamin W. Gregor, et al., (2024), "Automated human induced pluripotent stem cell culture and sample preparation for 3D live-cell microscopy", NP 19, 565–594, DOI: 10.1038/s41596-023-00912-w

[15] Diogo Teles, et al., (2024), "Using induced pluripotent stem cells for drug discovery in arrhythmias. Expert Opinion on Drug Discovery, 19(7), 827–840, DOI: 10.1080/17460441.2024.2360420

[16] Zahra Rahmani, et al., (2023), "Adult stem cell donor supply chain network design: a robust optimization approach", SC 27, 6367–6389, DOI: 10.1007/s00500-023-07830-9

Received: Aug 02, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[17] Dr. Roberto Crocchiolo, et al., (2023), "A New Tool Supporting the Selection of the Best Hematopoietic Stem Cell Donor by Modelling Local Own Real-World Data", PP, 2024090475. DOI: 10.20944/preprints202409.0475.v1

[18] Lina Hamad, et al., (2024), "Facilitating the ethical sourcing of donor hematopoietic stem cells for cell and gene therapy research and development", RM, 19(6), 317–326, DOI: 10.1080/17460751.2024.2357930

[19] Branden J Clark, et al., (2024), "Advancing Parkinson's disease treatment: cell replacement therapy with neurons derived from pluripotent stem cells", Stem Cells, sxae042, DOI: 10.1093/stmcls/sxae042

[20] Nidhi G. Thite, et al., (2024), "Stain-Free Approach to Determine and Monitor Cell Heath Using Supervised and Unsupervised Image-Based Deep Learning", JPS, Volume 113, Issue 8, Pages 2114-2127, ISSN 0022-3549, DOI: 10.1016/j.xphs.2024.05.001

Received: Aug 02, 2025