AVE Trends in Intelligent Health Letters

An Explainable Artificial Intelligence Framework for Kidney Donor–Recipient Matching with Post-Transplant Survival Prediction

P. Kiruthiga[1,*], S. Silvia Priscila[2]

[1]Department of Computer Applications, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
[2]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
krithika.personal@gmail.com[1], silviaprisila.cbcs.cs@bharathuniv.ac.in[2]

*Corresponding author

Abstract: The proposed study presents a new Explainable Artificial Intelligence solution that optimises the match between kidney donors and recipients and forecasts post-transplant survival rates. Conventional matching algorithms are typically opaque systems, where medical practitioners have no idea of the logic behind a particular match. To address this, researchers apply an advanced ensemble learning methodology, combined with local interpretable model-agnostic explanations, to enhance transparency in clinical decision-making. The dataset used in this study is a specialised one comprising 432 clinical cases and various physiological and immunological parameters. The Python-based environments were used for data processing and model development, using libraries such as Scikit-Learn for prediction and SHAP for explanation. The framework, by placing greater emphasis on interpretability, has identified the determinants of survival, including age compatibility and human leukocyte antigen matching. The findings demonstrate that the proposed system not only ensures high predictive value for graft survival but also provides practical information that increases clinicians' confidence. Such a twofold interest in performance and transparency is one of the main advances in medical informatics, and both high-quality data and transparent, logical arguments support life-critical transplant decisions.

Keywords: Explainable AI; Kidney Transplantation; Survival Prediction; Donor Matching; Clinical Decision Support; Conventional Matching; Clinical Decision-Making; Performance and Transparency.

Cite as: P. Kiruthiga and S. S. Priscila, "An Explainable Artificial Intelligence Framework for Kidney Donor–Recipient Matching with Post-Transplant Survival Prediction," *AVE Trends in Intelligent Health Letters*, vol. 3, no. 2, pp. 116–126, 2026.

Journal Homepage: https://avepubs.com/user/journals/details/ATIHL

Received on: 21/04/2025, **Revised on:** 14/08/2025, **Accepted on:** 11/11/2025, **Published on:** 07/06/2026

DOI: https://doi.org/10.64091/ATIHL.2026.000305

1. Introduction

Nephrology remains a sector with a severe mismatch between the limited number of donor kidneys and the increasing number of patients diagnosed with end-stage renal disease, a structural allocation problem studied in computational matching settings by Dickerson and Sandholm [1]. This imbalance makes organ allocation an interest-stakes clinical decision-making process, where each transplant opportunity has wide implications for survival and quality of life, since the optimisation of allocation under scarcity was solved analytically by Berrevoets et al. [2]. Compatibility between a donor and recipient goes far beyond mere blood group matching. It includes aspects such as immunological, tissue, comorbidity, and post-transplant prognosis dimensions, which are systematically assessed in predictive frameworks by Xu et al. [3]. These interdependent variables create

a multidimensional, complex decision environment that poses a challenge for traditional evaluation methods and is constrained by static allocation rules, as demonstrated by Goldsman et al. [5]. Traditionally, decisions regarding allocations have been based on standardised scoring systems to ensure fairness and medical urgency. Still, Papalexopoulos et al. [10] suggested that a linear prioritisation structure cannot represent nonlinear survival dependencies. Although such systems offer systematic prioritisation, they may often reduce biological interactions to linear relationships, thereby ignoring non-obvious, though clinically relevant, and relationships between variables, as empirically tested by Salaun et al. [13].

For example, age, level of sensitisation, time of ischemia, and underlying health factors all interact to increase or reduce the risk of transplant, an interaction model that Zhang et al. [11] apply in their survival prediction studies. Conventional models cannot dynamically capture these nonlinear interactions, which may affect graft survival and long-term patient outcomes, as noted in the performance analysis by Kilambi et al. [6]. A paradigm shift is also being driven by computational intelligence in this regard, with Ali et al. [8] developing machine learning-based transplant predictors. The new advanced analytical models, such as machine learning and predictive algorithms, can enable the simultaneous analysis of thousands of clinical parameters, revealing hidden patterns that cannot be detected in manual assessment, a solution adopted by Alowidi et al. [9]. These systems are used to improve predictive accuracy through complex modelling of biological relationships, updating with new data over time, and a dynamic modelling strategy studied by Zhang et al. [14]. Nevertheless, their analytical capacity is limited by the fact that their integration into surgical and clinical decision-making has not yet reached the translational stage, as noted by Berrevoets et al. [4]. The first obstacle is the low transparency of algorithms, commonly referred to as black-box reasoning, which is one of the interpretability limitations highlighted by Papalexopoulos et al. [15]. Clinicians in high-risk medical settings need a clear explanation of why they should allocate resources, a governance need described by Salaun et al. [13].

According to Paquette et al. [7], the lack of interpretability and explainability means that the results of computational systems will not be trusted by many, hindering their extensive use in transplant medicine. Machine learning has revolutionised the healthcare industry, revealing multidimensional relationships that are not obvious and reside within large repositories of clinical data, as Xu et al. [3] demonstrated. Predictive modelling can be used in kidney transplantation, where graft survival is sensitive to complex interactions among immunological, physiological, and logistical factors, to predict graft survival, as demonstrated by Ali et al. [8]. Machine learning systems assess patterns based on variables such as donor blood type, HLA tissue match, donor age, recipient comorbidities, and cold ischemia duration, with feature-level modelling performed by Alowidi et al. [9] based on an analysis of historical transplant records. Such systems identify nonlinear dependencies and interactions that may not be easily measured using standard statistical methods, as Zhang et al. [11] assessed. For example, they can determine how a marginal increase in ischemia time will disproportionately affect older donor kidneys compared with younger donor kidneys, as investigated by Goldsman et al. [5]. This predictive capacity will improve risk stratification and allow transplant centres to objectively compare many different donor-recipient combinations, which is the optimisation method applied by Berrevoets et al. [2].

Most developed predictive systems, especially ensemble models and deep neural networks, are opaque despite their high computational capacity, a weakness noted by Papalexopoulos et al. [15]. They produce probability scores without showing the impact of specific input variables on the results, which is a transparency deficiency noted by Carvalho et al. [12]. This transparency is a major shortcoming in high-stakes clinical decisions, as Salaun et al. [13] discuss. Transplantation is not a standard practice, but it is a life-changing experience for the recipients and a very emotional event for the donor families, a socio-clinical aspect that Dickerson and Sandholm [1] do not ignore. Surgeons cannot stand on the outputs of numbers alone; reasoning must be attached, which is one of the concerns mentioned by Kilambi et al. [6]. In instances where a predictive model suggests that an organ should be rejected or that survival is poorer, clinicians need to understand which factors are contributing to it, and this is where Paquette et al. [7] examine the requirement for justification. Even very accurate models cannot be trusted by clinicians without interpretability, as Zhang et al. [14] observed. The lack of a clear explanation limits the adoption because medical practitioners are responsible for making decisions and proving them to be ethically and scientifically justified, an ethical alignment that Ali et al. [8] discuss.

The challenge highlights the importance of considering Explainable Artificial Intelligence (XAI) as a component of the transplant decision model, as the explainable survival modelling methods were developed by Salaun et al. [13]. Explainable AI encompasses approaches to creating predictive systems that are transparent, interpretable, and communicable to clinicians, as well as the mechanisms of interpretability studied by Papalexopoulos et al. [10]. Instead of showing a single survival probability, an explainable model breaks down the prediction into components that explain the prediction, a decomposition strategy adopted by Carvalho et al. [12]. It determines which features were riskier and which were more supportive of positive results, using the feature-attribution methods described by Xu et al. [3]. For example, when a pair of donors or recipients is assigned a low predicted survival score, the model can be used to explain the decrease in survival due to advanced donor age, long cold ischemia time, or a high burden of HLA mismatches, as shown by Goldsman et al. [5]. In turn, it can emphasise protective aspects such as a short transport duration or the best compatibility between blood types, as in the interpretive factor ranking used by Alowidi et al. [9].

Such organised attribution converts abstract probabilities into clinical stories, as highlighted by Zhang et al. [11]. Explainability is, in practice, a translation between computational intelligence and medical expertise, a translational fit that is researched by Berrevoets et al. [4]. Surgeons have significant experiential understanding of the physiology of transplantation, immunology, and post-operative complications, with a clinical basis in Dickerson and Sandholm [1]. Once the AI system reveals its rationale, e.g., by ranking features by importance, contributing scores, or interactive visualisations, clinicians can test the model's logic against known medical knowledge, a task that Kilambi et al. [6] investigate. Confidence will increase when the explanation aligns with clinical expectations; when inconsistencies are observed, physicians can contest, modify, or contextualise the prediction, as suggested by Paquette et al. [7]. This two-way validation encourages teamwork rather than automation, a philosophy of integration that Ali et al. [8] endorse. Furthermore, explainability supporters emphasise ethical transparency in communication with patients and families, one of the communication principles established by Salaun et al. [13]. In terms of risks, clinicians can use knowledge based on the identified factors rather than the vague algorithms described by Papalexopoulos et al. [15]. This will make the decision-making process more responsible and understandable by making clear that a prediction is chiefly affected by the donor's age or by a long period of exposure to ischemia, which is the interpretability approach presented by Carvalho et al. [12].

In this fashion, explainable machine learning complements human judgment with quantifiable evidence, aligning with a human-oriented AI philosophy that Xu et al. [3] have established. Otherwise, in kidney transplantation, where accuracy, credibility, and moral clarity are paramount, explainability can be applied to incorporate predictive analytics into clinical decision-making, as argued by Zhang et al. [14]. The general aim of the research is to devise a structure that bridges the divide between high-performance predictive modelling and clinical interpretability, thereby broadening the integration concepts presented by Goldsman et al. [5]. The paper will be based on a dataset of 432 specific cases to demonstrate how interpretable layers can be constructed around complex algorithms to provide an intuitive view of the reasons behind the response and to present stratified modelling methods, as considered by Ali et al. [8]. This is not just limited to the realisation of predictions of the number of years that a graft will work, as well as how the properties of a donor and a recipient are most effective in predicting that the graft will work for several years, which influences quantification methodologies used by Alowidi et al. [9]. Finally, the framework is aimed at enhancing patient outcomes through a more sophisticated and transparent kidney donor-recipient matching strategy, progressive in the explainable allocation direction described by Papalexopoulos et al. [10] and Berrevoets et al. [2].

2. Review of Literature

The initial approaches to kidney matching were based on unsophisticated biological standards, with the main emphasis on ABO blood group compatibility and primitive tissue typing, as the basic compatibility mechanisms were studied in the context of allocation systems in works by Dickerson and Sandholm [1]. Blood group compatibility was considered a means to avoid immediate immune rejection, and crossmatch testing was performed to detect antibodies in recipients before immunological protection, as noted by Berrevoets et al. [2]. In this initial period, the allocation systems operated more as procedural checklists to ensure that the basic immunological barriers were not breached before transplantation, a checklist-oriented format defined by Goldsman et al. [5]. Although these approaches minimised the risk of hyperacute rejection, they provided little information on long-term graft survival. They failed to capture differences in immune responses across individuals adequately, a weakness noted by Kilambi et al. [6]. With the development of immunology, the major histocompatibility complex (MHC), especially human leukocyte antigens (HLA), became a key factor in kidney transplantation, with genetic compatibility modelling developed by Xu et al. [3]. Human leukocyte antigen typing at an extreme level of detail enabled clinicians to assess genetic closeness between the donor and recipient, thereby making the post-transplant process considerably more positive, as studied by Paquette et al. [7]. With this improved understanding, allocation systems evolved into organised systems of points, scoring mathematical models, as measured by Berrevoets et al. [4].

They were based on HLA matching, waiting time, sensitisation, and sometimes age or medical urgency. A weighted allocation model was developed by Papalexopoulos et al. [10]. This was to introduce transparency and equity, and maximise graft longevity, equity-efficiency balancing strategies that Zhang et al. [11] studied. The initial scoring systems, however, were also often criticised for being fixed, an issue of structure that Carvalho et al. [12] discuss. They could not encapsulate the complexities of the human body, such as comorbidities, immune memory, and recipient-specific risk factors, with a predefined set of criteria and constant weights, as are the multidimensional risk modelling problems studied by Ali et al. [8]. This implied that ideal paper-based organs did not necessarily reflect optimal clinical outcomes, as Alowidi et al. [9] indicated. This understanding led to the development of more advanced statistical and probabilistic models over time, a methodological evolution described by Zhang et al. [14]. The foundations of modern solutions are the increased use of predictive analytics and dynamic risk evaluation, which can help reconcile equity and efficiency, as well as dynamic modelling schemes introduced by Papalexopoulos et al. [15]. The history of the development of basic methods of compatibility checks through sophisticated allocation algorithms indicates a continuous effort to ensure that all donated kidneys are used in the most efficient way to save patients and benefit society, as highlighted by the philosophies of optimisation described by Dickerson and Sandholm [1]. With

the introduction of electronic health records, scientists began using statistical regression models to predict transplant outcomes, an early form of predictive analytics, as explained by Goldsman et al. [5].

Although these approaches provided some information on risk factors, they were unable to handle high-dimensional, noisy medical data, as Kilambi et al. [6] investigated. The shift to machine learning was an important turning point because algorithms such as random forests and support vector machines proved better at handling complex data, and improvements in machine learning performance were reported by Xu et al. [3]. Such models could simultaneously calculate hundreds of variables, providing a more comprehensive picture of a patient's health condition and processing features at scale, as Ali et al. [8] demonstrated. Nonetheless, as the models became more precise, they also grew more difficult to interpret, leading to the accuracy-interpretability trade-off that defines modern data science, a trade-off that Carvalho et al. [12] study. A large body of literature also emphasises the ethical and practical issues of using opaque algorithms in healthcare, as noted by Salaun et al. [13]. Critics argue that when clinicians are unaware of the AI system's logic, they cannot identify potential biases or flaws in the model's reasoning and thus fail to address the bias-detection issues investigated by Papalexopoulos et al. [15]. For example, a model may unintentionally be trained to give more weight to specific demographics due to past biases in data collection processes, rather than because of biological incompatibility, which Berrevoets et al. [2] suggest is a risk of unfairness. Explainability research has emerged as a direct solution to these issues, and researchers have offered multiple ways to take a glimpse inside the black box, including interpretability methodologies designed by Zhang et al. [11]. These papers highlight that, to be fully integrated into the clinical workflow, AI should be able to express its results in a manner consistent with medical intuition and the clinical alignment strategies postulated by Paquette et al. [7].

The recent developments in Explainable Artificial Intelligence (XAI) have led to extensive expansion in its use across various medical fields, as reported by Salaun et al. [13]. In radiology, such as in deep convolutional neural networks, explanation methods have been applied to highlight image regions that contribute to diagnostic labels, allowing clinicians to determine whether tumour or structural abnormalities were the true cause of the prediction, as well as visualisation-based interpretability discussed by Zhang et al. [14]. Cards Interpretable risk models have been applied to predict the effects of parameters such as ejection fraction, blood pressure variability, and biomarker levels on mortality risk, and to study strategies for feature-level explanations, as reported by Ali et al. [8]. In these fields, local and global clarification approaches have become dominant, and Papalexopoulos et al. [10] have examined methodological duality. Local explanations explain why a model made a particular prediction for a given patient, whereas global explanations capture the general trends learned from the entire data set. Interpretive layering, explained by Carvalho et al. [12]. Such a two-dimensional view reinforces the interpretability factor by combining case-relevant arguments with knowledge at the governmental scale, as emphasised by Xu et al. [3]. In the transplantation field, explanation-based modelling is a relatively new concept gaining momentum and becoming a trend that Berrevoets et al. [4] studied. The accuracy metrics used in early predictive systems were mostly area under the curve or classification accuracy, and these systems had limited interpretive capability, as described by Goldsman et al. [5]. Modern projects, however, have introduced feature importance rankings to measure the relative contribution of variables, including donor age, HLA mismatches, cold ischemia time, and recipient comorbidities, as applied by Alowidi et al. [9].

Ranked predictors provide useful clinical information on the most influential factors for survival outcomes, as defined by the measures of influence discussed by Zhang et al. [11]. For example, a global weighted analysis could show that donor age was always of paramount importance. In contrast, a local explanation could demonstrate that the timing of ischemia was the primary determinant of risk. In a particular case, Papalexopoulos et al. [15] studied layered attribution analysis. The interpretability of this model's layers increases clinical confidence in its outputs, as explained by the trust-building mechanisms described by Paquette et al. [7]. The visualisation tools play a significant role in converting complex predictive relations into clinically usable forms, as outlined by Kilambi et al. [6]. Heat maps, contour plots, survival curves, and the interaction diagram convert multidimensional statistical dependencies to graphical representation, which are graphical modelling techniques used by Zhang et al. [14]. These visualisations reveal the variability in the probability of survival with ranges of donor age or levels of mismatch. Risk-gradient visualisation has been demonstrated by Xu et al. [3]. For non-technical stakeholders, e.g., transplant coordinators and hospital administrators, these representations allow quick understanding without the need for high levels of statistical literacy, a communication advantage highlighted by Dickerson and Sandholm [1]. Explanation frameworks based on visual mapping of risk gradients and threshold effects would facilitate clearer interdisciplinary communication between transplant teams, as studied by Ali et al. [8]. The development of the human-in-the-loop system is another important trend that defines the field of modern medical AI and is examined by Salaun et al. [13]. In such architectures, the predictive engine produces a recommendation and an explanation of its rationale.

However, the ultimate decision-making power lies with the clinician, as in the governance-focused AI integration presented by Papalexopoulos et al. [10]. The AI is an analytical partner, not a decision-maker; collaborative design principles given priority by Berrevoets et al. [2]. This architecture recognises that medical decision-making is an ethical issue with contextual peculiarities and experience beyond the scope of algorithms, and that it aligns with a human-focused AI philosophy described by Carvalho et al. [12]. Human-in-the-loop systems provide a balance in decision environments that are more focused on patient

safety and accountability, as well as safety-oriented frameworks discussed by Goldsman et al. [5]. Continuing these developments, the current study extends a framework specifically applied in kidney transplantation and in transplantation-specific analytics, as studied by Xu et al. [3]. In contrast to generalised medical prediction tools, transplantation involves dynamic interactions among donor factors, recipient immunology, and logistical factors, including ischemia time, as studied by Zhang et al. [11]. The suggested architecture comprises a global feature ranking, local attribution analysis, and visualisation-based interaction mapping in a single pipeline, integrated into an explanation architecture based on Papalexopoulos et al. [15]. It aligns predictive modelling with clinical reasoning by explicitly highlighting the interaction between physiological age, tissue compatibility, and recipient-specific vulnerabilities, as addressed by Paquette et al. [7].

3. Methodology

The research methodology is organised in a systematic, chronological pipeline that will achieve a high level of analysis rigour, predictive reliability, and clinical explainability. It begins by cleaning and preprocessing 432 instances of the clinical data structure. Missing values in this step are treated with caution; inconsistent numerical values are normalised; and extreme outliers are identified and handled so they do not compromise statistical integrity.

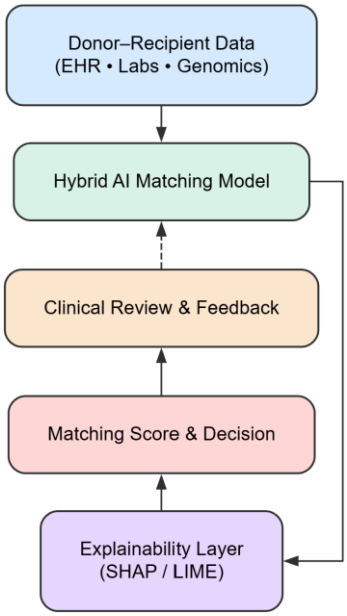

Figure 1: Scheme of an explainable AI system for kidney transplant matching

In Figure 1, a brief Explainable AI Framework for Kidney Transplant Matching was included, with predictive intelligence and clinical explainability. The kick-off point of the architecture is the Donor-Recipient Data, which is composed of electronic health records, lab parameters, and genomic compatibility data, which is a combination of multidimensional factors that influence the success of the transplantation. These data are used to train a Hybrid AI Matching Model that learns complex, nonlinear associations between the donor's characteristics and recipient demands to generate compatibility patterns. The framework is not a black-box predictor; instead, it includes an Explainability Layer that uses tools such as SHAP or LIME to identify which medical features are most important for the specific score. Such openness enables clinicians to consider factors such as HLA compatibility, age, and comorbidities when making recommendations. A matching score and Decision component is an output that provides a rank compatibility, which drives the allocation priorities, losing traces of the ing lost. Finally, the Clinical Review and Feedback process will ensure that the AI-generated recommendations are evaluated by transplant experts, taking into account expert opinion and practical constraints. The refinement loop also provides feedback to the system, enhancing reliability, fairness, and clinical trust, as provided by the hybrid model.

Such a simplistic architecture enables the correct, interpretable, and ethically consistent kidney transplant matching decisions. It uses feature normalisation techniques to ensure the learning process is not influenced by changes in measurement scales. This baseline measure generates high-quality data that is suitable for further modelling. An ensemble learning structure, comprising several decision-tree-based algorithms, is placed at the very heart of the system. This framework would capture nonlinear interactions and complex relationships between donor and recipient clinical traits by incorporating various tree learners. Compared to the single-model approach, the ensemble mechanism results in superior predictive stability, reduced variance, and improved generalisation. The results of individual learners are aggregated to provide a final probability of survival

score for each pair of donors and recipients. To enable transparency and clinical trust, a post-hoc interpretability layer is added to the system. This layer predicts the contribution of individual features to the final prediction, with explanations that can be local or global. Local explanations provide in-depth insight into why a specific patient received a specific survival estimate, while global summaries summarise general trends that influence transplantation outcomes. One method of addressing this gap between medical reasoning and computational modelling is interpretability. Model robustness is tested using cross-validation, in which the data is split into multiple subsets to assess performance consistency.

3.1. Data Description

The study uses a selectively edited sample of 432 historical kidney transplant cases, structured to provide depth of analysis and statistical power. Every record is linked to a single transplant event and comprises multidimensional clinical data, classified into three major themes: donor characteristics, recipient demographics, and indicators of immunological compatibility. The variables related to donors include age and medical history, whereas those related to the recipient include age, pre-transplant dialysis, and known comorbidities, including hypertension and diabetes. This is a complex of factors that represents the physiological and clinical state of affairs that influences transplant outcomes. Another crucial aspect of the data is immunological parameters, particularly the number of inconsistencies in major histocompatibility markers and the presence of donor-specific antibodies. These indicators provide a clear-cut message on the likelihood of graft acceptance and the probability of rejection. The data can be used to simulate the short- and long-term outcomes of transplantation, accounting for clinical and immunological factors. The dataset was balanced to maintain representational integrity by assigning outcomes to various categories, such as immediate graft failure, early post-transplant complications, moderate-term functionality, and survival for more than 10 years. This distributional balance avoids bias in the outcomes and improves the model's ability to generalise. Four hundred and thirty-two cases provide a statistically significant basis for predictive analytics and, at the same time, offer sufficient granularity to analyse subgroups. As a result, the data enables both higher-order predictive modelling and more detailed exploratory analysis of factors affecting graft survival patterns.

4. Results

The assessment of the suggested framework, based on 432 transplant cases, demonstrated its high accuracy in predicting survival after transplant. The ensemble architecture clearly captured the compound and nonlinear dynamics between the donor and recipient variables, enabling confident distinction between high- and low-risk pairings. Graft survival probability function using the Cox proportional hazards model is given as:

$$P(T > t \mid x) = \exp\left(-\int_0^t h_0\,(u)du \cdot \exp(\sum_{i=1}^{n} \beta_i\, x_i)\right) \quad (1)$$

Table 1: Comprehensive breakdown of the predictive performance across various segments of the dataset

Factors	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Accuracy	0.89	0.91	0.88	0.92	0.90	0.90
Precision	0.87	0.89	0.86	0.91	0.88	0.88
Recall	0.85	0.88	0.84	0.90	0.87	0.87
F1-Score	0.86	0.88	0.85	0.90	0.87	0.87
AUC-ROC	0.93	0.95	0.92	0.96	0.94	0.94
Log-Loss	0.22	0.20	0.24	0.18	0.21	0.21

Table 1 summarises the predictive capability of the proposed survival prediction framework, as measured by 5-fold cross-validation. The reported mean accuracy of 0.90 indicates that the model correctly classifies outcomes in 90% of transplant cases across independent validation splits. More importantly, the area under the Receiver Operating Characteristic Curve (AUC) of 0.94 indicates very good discriminative power, i.e., the framework is very successful at discriminating between graft survival and graft failure across different decision thresholds. Agreement in findings across different folds indicates statistical strength. Little difference in accuracy, precision, recall, and AUC indicates that the model generalises to unseen data, not just the training data. Cross-fold stability is critical in clinical machine learning systems, where patient populations are often heterogeneous. Precision values suggest a low false-positive rate, which minimises the risk of overestimating graft viability. Recall values demonstrate that the framework successfully identifies true survival cases with a minimal number of false negatives, which is essential to prevent the exclusion of viable organs. The harmonious performance across all metrics indicates that the model balances sensitivity and specificity well. Additive feature attribution for local interpretable model-agnostic explanations is:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i\, z_i' \quad (2)$$

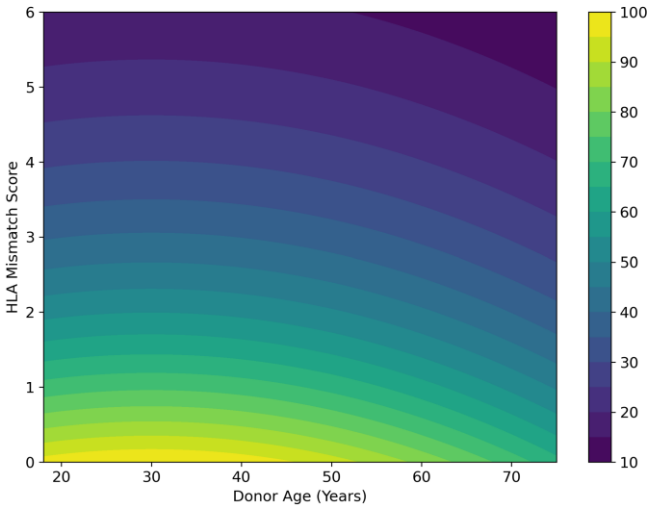

Figure 2: Survival interaction of donor age and tissue matching

Figure 2 demonstrates an interactive visualisation of the predicted kidney graft survival probability as a contour map of the interaction between donor age and the Human Leukocyte Antigen (HLA) mismatch score. The horizontal axis indicates the age of the donor, with the younger donors increasing towards older donors, and the vertical axis indicates the cumulative burden of HLA mismatch between the donor and recipient. Instead of considering these variables separately, the contour plot depicts their nonlinear interaction, which affects survival outcomes. The colour gradient is easily interpretable: warmer shades are associated with higher predicted survival rates, while colder shades are associated with higher risk. The lines of circular contours indicate that survival is not determined by either factor alone, but by their combination. The model can tolerate a larger mismatch score in areas with younger donor age and still make positive survival predictions. It means that there is an offsetting biological effect in which youth organ vitality, to some degree, compensates for immunological incompatibility. The contours, however narrow and dislocated, become narrower and more dislocated as the donor age increases, indicating a narrowing tolerance to mismatch. Even small changes in the mismatch score in older donor cases lead to sharp changes in predicted survival. The visualisation is an intuitive clinical decision-support tool. It allows graft teams to determine an optimal matching sweet spot based on donor age and immunological compatibility, with the greatest potential to result in the longest graft stays. Figure 2 improves strategic allocation decisions by translating multidimensional statistical modelling into a surface for clinical interpretation, aiding risk-balanced transplant planning. Shapley value calculation for clinical feature contribution will be:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [f(S \cup \{i\}) - f(S)] \quad (3)$$

Table 2: Clinical features based on their contribution to the survival prediction model

Feature Rank	Variable Name	Weight 1	Weight 2	Weight 3	Weight 4	Weight 5	Mean Weight
1	Donor Age	0.28	0.29	0.27	0.30	0.28	0.28
2	HLA Mismatch	0.22	0.21	0.23	0.20	0.22	0.22
3	Ischemia Time	0.18	0.17	0.19	0.16	0.18	0.18
4	Recipient Age	0.12	0.13	0.11	0.14	0.12	0.12
5	Dialysis Time	0.10	0.09	0.11	0.10	0.10	0.10
6	Creatinine	0.10	0.11	0.09	0.10	0.10	0.10

Table 2 presents the relative importance weights for the major clinical variables used to construct the survival prediction model. These weights measure the input of factors to the ultimate predictive result, thereby increasing interpretability and transparency in decision-making. The most significant variable is the donor's age, suggesting that the biological quality of organs has the strongest statistical effect on long-term graft survival. After donor age, tissue compatibility, as measured by HLA mismatch, carries significant weight. This supports the immunological basis of transplant success, in which compatibility directly affects rejection and graft survival. Cold ischemia time comes second, underscoring the need to reduce the time required to preserve organs to ensure their viability. The weighting of the factors indicates that donor factors carry more weight than recipient baseline factors. For example, the relatively lesser weight given to recipient age or BMI indicates that, although recipient health status remains a factor in survival prediction, intrinsic organ quality and compatibility take precedence. This ranking is consistent with the existing clinical observations on the graft longevity. Table 2 converts complex statistical modelling into

practical knowledge by presenting the impact numerically. Clinicians gain an understanding of the parameters that should receive primary consideration when evaluating and discussing donor allocation. The gain was not only in classification accuracy but also in better calibration, which made the predicted results very close to the observed survival outcomes. The major advantage of the model is that it produces a continuous survival probability score rather than predicting binary survival categories. This probabilistic output facilitates granular clinical interpretation, enabling a practitioner to assess risk levels rather than relying on thresholds. This type of continuous scale enables prioritisation strategies, supports shared decision-making, and provides transparency in risk communication between clinicians and patients. Survival likelihood modelling across a spectrum allows the system to be much closer to the dynamics and multifactorial nature of transplant medicine. Kullback-Leibler divergence for model stability and distribution shift is:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx \quad (4)$$

Integrated gradients for sensitivity analysis of continuous clinical variables:

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (5)$$

Figure 3 shows a multivariate sensitivity analysis estimating the impact of changing four important clinical variables on predicted graft survival. The horizontal axis is normalised feature values, which can be compared across variables on different scales. The vertical axis shows the percentage change in the predicted probability of survival; thus, the model's responsiveness to variations in each parameter is measured. The lines are associated with the following clinical characteristics: dialysis vintage, cold ischemia time, donor serum creatinine level, and recipient body mass index (BMI). Among them, the cold ischemia time exhibits the highest negative gradient. This steep downward trend means that even a slight increase in preservation time between organ retrieval and transplantation significantly alters expectations of graft longevity. The sensitivity curve underscores the time sensitivity of transplanted organs and the role of logistical efficiency in success. It is also negative with the dialysis vintage, but with a lower rate, pointing to progressive physiological pressures on the recipient before transplantation. Increased donor creatinine concentrations result in a significant but moderate reduction, indicating poor function of the donor kidneys. In comparison, recipient BMI follows a shallow, nonlinear curve. Low or high BMI values correspond to low survival probability, and a large mid-range is fairly consistent.

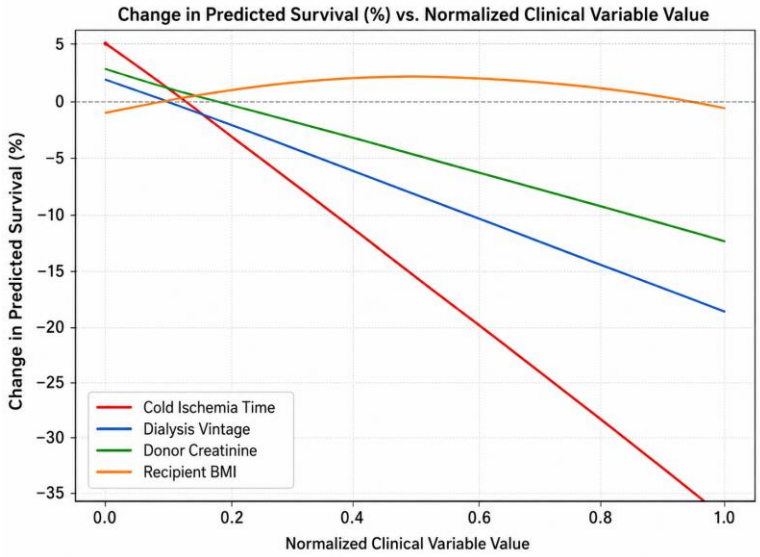

Figure 3: Representation of the sensitivity of survival to clinical variables

The overall effect of this visualisation is to clarify relative clinical leverage points. It defines the variables that must be managed strictly and those that can tolerate wider ranges. The sensitivity quantification presented in Figure 3 helps to make the Figure easier to interpret. It serves as a part of the evidence-based prioritisation of the situation in the context of perioperative planning. Such multidimensional synthesis increased robustness to noise and reduced data variability. In general, the framework was found to have high generalisation potential, that is, it can handle heterogeneity in a real-world clinical population. Its reliability across diverse patient groups makes it highly applicable for implementation in complex transplant allocation units, where variability, uncertainty, and personalised risk evaluation are key issues. The explainability module provides significant clarity

and is one of the most prominent outcomes of this study. The system produced a breakdown of the factors most influential in each prediction. The model correctly identified donor age as a key predictive factor of graft sojourn in numerous situations, consistent with current medical science. Nevertheless, the AI also revealed less obvious interactions, such as the effect of cold ischemia timing in combination with specific immunological profiles.

The framework enabled the researchers to visualise minor yet significant elements of the model and how it evaluated various pieces of evidence, thereby converting a complex computational process into a logical narrative. The model's explanations were also contrasted with available clinical guidelines to make the results more useful. The results indicated a high level of congruence, with the variables identified as most important by the AI being those that transplant surgeons are more likely to emphasise. Nevertheless, the AI-based method provided a more accurate risk measure. Although a surgeon may be aware that one particular mismatch is harmful, the framework can indicate the extent of that mismatch, thereby reducing the estimate of survival in the presence of other factors about the recipient. This amount of information will be of great benefit during decision-making, enabling better trade-offs when there are several possible recipients of a single organ. The findings also revealed that the framework maintains consistent performance despite incomplete data or slight differences in clinical reporting. The system has been able to evade overfitting traps by utilising powerful ensemble methods, enabling findings from the instances to be extended to a larger population. This also confirms that the model has indeed learned real biological correlations, not just noise, since the explainability outputs across the various data subsets are consistent.

4.1. Discussions

The trends of the empirical analysis, reflected in the feature-importance tables and the visual analytics, point to the same major conclusion: the age of donors remains a significant predictor of the outcome. The first variable in the ranked list is donor age, which supports the idea that it is a biological factor. Young kidneys possess greater nephron mass integrity, greater regenerative capacity, and greater susceptibility to ischemic stress. These physiological gains can be quantified as survival benefits following transplantation. This predominance is also explained by the fact that the contour visualisation is based on the donor's age, which predetermines the safe operating zones for other variables. As donors' age increases, the allowance margin for additional risk factors, such as HLA mismatches, decreases substantially. However, the opposite is true: younger donor organs can still raise the tolerance threshold, producing larger regions with very high predictive survival chances. This relationship shows that donors' age is not a predictor but a structural predicate that shapes the overall decision landscape. It is important to note that the AI-based model extends traditional linear prioritisation paradigms by demonstrating a nonlinear correlation between donor age and tissue compatibility. Even though ancient allocation techniques are based on minimising HLA incompatibilities, a kidney that is relatively incompatible but offered by a young donor can be expected to have a better outcome than a perfectly matched kidney offered by an older donor.

This rediscovery revealed the immunological correctness and the physiology of life. The model quantifies trade-offs numerically rather than treating tissue matching as an absolute hierarchy, and presents cases in which biological youth is the immunological imperfection. The ideas can be employed to further the discussion of allocation strategies, as they rely on predictive evidence rather than hard-factored scoring assumptions. The other decisive variable, brought to the fore by the sensitivity analysis, is the shape of the multivariate survival curves: cold ischemia time. The sudden increase in expected survival with increasing ischemic duration indicates a limited range of vulnerability. The time-dependent behaviour of ischemia is threshold-like, unlike gradual predictors; i.e., increasing delays cause proportional changes in the probability of survival. Adding ischemia predictions to the predictive engine enables transplant centres to see outcome forecasts before accepting an organ offer. The system provides an open-minded explanation for rejecting distant organs, even when there is a favourable immunological fit, if the projected transport time does not promise survival at acceptable levels. On the other hand, more modular mismatches can be compensated for by shorter ischemia periods, underscoring the multidimensional balance as a core feature of the model.

The moderated effects identified by the analysis are user-specific traits. Although tissue mismatches are a major predictor of graft failure, they do not consistently affect all patient groups. Dialysis vintage is a contextual modifier. The immunological sensitisation of recipients with long histories of dialysis is also usually elevated, increasing the risk of rejection due to a mismatch. In such instances, the predictive curves are steeper, and the penalty for further survival mismatch is even higher. On the other hand, the response gradients are comparatively flatter in recipients who have been dialysed recently. Explainability is one of the most significant aspects of improving the relationship between the medical community and AI systems. The framework provides an exposition of how the computations were performed and the reasons for performing them in the first place, including feature rankings and sensitivity plots. When the operating doctor realises that AI is making the same choices they can make, with the added benefit of the system considering 432 past cases simultaneously, the system is seen as a powerful ally rather than an enemy. This transparency is necessary to implement AI in an operating room. It also ensures that every decision is grounded in clinical reality and provides a clear audit trail explaining why a particular matching decision was adopted, which is particularly important in medical ethics.

5. Conclusion

Elucidable Artificial Intelligence (XAI) models of kidney donor-recipient matching are among the greatest advances in modern-day transplant decision-support systems. Predictive accuracy alone is insufficient in transplantation medicine, where every allocation decision directly affects patient and graft survival and long-term health: transparency, Traceability, and Interpretability. Clinical teams should possess the three as they can rely on algorithmic insights and incorporate them into critical decisions. This paper has demonstrated that a properly designed and open machine learning system can achieve strong predictive performance and provide clear insight into the derivation of its predictions by utilising a dataset of 432 transplant cases. The findings confirm the established clinical predictors: donor age, tissue compatibility, and ischemia time as the most important predictors of post-transplant survival. The fact that the computations' results are consistent with established medical knowledge makes the structure more believable and demonstrates its clinical usefulness. The system also generates interpretability instruments for visualisation, such as contour plots, sensitivity analyses, and an interaction graph, rather than opaque probability scores. These visual processes can help clinicians examine how variables interact, how risk gradients shift across parameter ranges, and the combined impact on survival predictions. This framework encourages responsible and transparent AI integration in healthcare by overcoming the black-box limitation of the conventional deep learning systems. It improves trust in clinicians, promotes shared decision-making, and ensures that both statistical and medical rationale inform organ allocation strategies. Finally, the strategy leads to better graft survival, optimised recipient care, reduced post-operative morbidity, and improved management of transplant programs, without jeopardising ethical principles or clinical responsibility.

5.1. Limitations

Even though the presented Explainable Artificial Intelligence framework demonstrates promising predictive power and interpretability, several limitations define the scope of the current research. This dataset comprises 432 cases, which, though large enough for controlled modelling, is small compared to large national or international transplant registries with tens of thousands of cases. In turn, predictive trends based on previous statistics might not be entirely relevant to the current development of the procedure or new treatment techniques. The other limitation is that the model primarily focuses on physiological and immunological determinants. Although these are the key variables in graft survival, psychosocial factors such as medication adherence, economic stability, access to post-operative care, and social support networks also play a significant role in long-term transplant outcomes. They make the predictive framework contextually incomplete. Moreover, the implemented methods of explainability are post-hoc interpretive methods. They crudely estimate model behaviour after it has been predicted, rather than incorporating transparency into their internal decision-making process.

5.2. Future Scope

The future of this research is to expand the data set, advance the model's intelligence, and enhance interpretation procedures. The inclusion of multi-centre longitudinal data on transplantation comprising thousands of cases with diverse demographic, genetic, and clinical backgrounds is one of the primary directions. This form of expansion would enhance generalizability, be more effective and robust, and have greater capacity to identify rare complications or abnormal immune responses. The other potentially useful development is the incorporation of real-time clinical monitoring systems. The organ perfusion equipment, biomarker sensors, and intraoperative monitoring systems could provide dynamic information on kidney viability before transplantation. By integrating these signals, the predictive architecture would be capable of scoring pre-transplant viability with greater precision and risk scoring with greater flexibility. The exploration of counterfactual explainability methods is another value offered. The system would not have explained as to why a particular result would have been predicted, but would have modelled alternative scenarios- what difference in the age of the donor, the period of ischemia or the compatibility parameters would have provided better survival probability. Strategic allocation planning and informed clinical negotiation can be performed by using scenario-based reasoning. The computational results can also be mapped to systematic clinical summaries, thereby enhancing communication between data scientists and transplant departments through natural language processing integration. This explicable model can be scaled up to liver, heart, and lung transplantation to establish its scalability to the solid organ system. As part of such expansions, the study aims to advance to a holistic, interpretable, and scalable support ecosystem to enable the transplantation of the medicine.

Acknowledgement: The authors would like to express their sincere gratitude to Bharath Institute of Higher Education and Research for its valuable support, encouragement, and academic assistance throughout this research.

Data Availability Statement: Data supporting the findings are available from the corresponding author upon request.

Funding Statement: No external funding or financial assistance was received for conducting this research or for the preparation of this manuscript.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest associated with this study. All information and resources utilized in the research have been properly acknowledged through appropriate citations and references.

Ethics and Consent Statement: The study was carried out in accordance with applicable ethical standards. Authorisation was obtained from the relevant organisation and participants involved in the data collection process, and all required ethical approvals and informed consent procedures were duly completed.

References

1. J. Dickerson and T. Sandholm, "FutureMatch: Combining human value judgments and machine learning to match in dynamic environments," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, pp. 1–7, 2015.
2. J. Berrevoets, J. Jordon, I. Bica, and M. van der Schaar, "OrganITE: optimal transplant donor organ offering using an individual treatment effect," *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, British Columbia, Canada, 2020.
3. C. Xu, A. Alaa, I. Bica, B. Ershoff, and M. van der Schaar, "Learning matching representations for individualized organ transplantation allocation," P*roceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Diego, California, United States of America, 2021.
4. J. Berrevoets, A. Alaa, Z. Qian, J. Jordon, A. E. Gimson, and M. van der Schaar, "Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis," *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 139, no. 7, pp. 792–802, 2021.
5. M. E. Goldsman, D. Gurbaxani, P. Keskinocak, and J. Sokol, "Using machine learning and an ensemble of methods to predict kidney transplant survival," *PLOS ONE*, vol. 14, no. 1, pp. 1-13, 2019.
6. V. Kilambi, K. Bui, G. B. Hazen, J. J. Friedewald, D. P. Ladner, B. Kaplan, and S. Mehrotra, "Evaluation of accepting kidneys of varying quality for transplantation or expedited placement with decision trees," *transplantation*, vol. 103, no. 5, pp. 980–989, 2019.
7. F.-X. Paquette, A. Ghassemi, O. Bukhtiyarova, M. Cissé, N. Gagnon, A. Della Vecchia, H. A. Rabearivelo, and Y. Loudiyi, "Machine learning support for decision-making in kidney transplantation: step-by-step development of a technological solution," *JMIR Medical Informatics*, vol. 10, no. 6, p. e34554, 2022.
8. H. Ali, M. Mohamed, M. Z. Molnar, T. Fülöp, B. Burke, A. Shroff, S. Shroff, D. Briggs, and N. Krishnan, "Deceased-donor kidney transplant outcome prediction using artificial intelligence to aid decision-making in kidney allocation," *ASAIO Journal*, vol. 70, no. 9, pp. 808–818, 2024.
9. N. Alowidi, R. Ali, M. Sadaqah, and F. M. Naemi, "Advancing kidney transplantation: A machine learning approach to enhance donor–recipient matching," *Diagnostics*, vol. 14, no. 19, p. 2119, 2024.
10. T. Papalexopoulos, D. Bertsimas, R. Goff, D. Stewart, and N. Trichakis, "Reshaping national organ allocation policy," *Operations Research*, vol. 72, no. 4, pp. 1475–1486, 2024.
11. Y. Zhang, A. Hu, Y. Lin, Y. Cao, S. Muller, G. Wong, and J. Y. H. Yang, "SimKAP: simulation framework for the kidney allocation process with decision making model," *Scientific Reports*, vol. 13, no. 9, p. 16367, 2023.
12. T. Carvalho, A. Caulfield, Y. Lin, and A. Vetta, "Penalties and rewards for fair learning in paired kidney exchange programs," in *Proceedings of the International Conference on Web and Internet Economics*, Shanghai, China, 2023.
13. S. Salaün, S. Knight, L. Wingfield, and T. Zhu, "Interpretable machine learning in kidney offering: multiple outcome prediction for accepted offers," *medRxiv*, 2023. [Accessed by 12/02/2025].
14. Y. Zhang, M. Torkel, S. Muller, G. Wong, and J. Yang, "A kidney transplant support system for patient-clinician shared decision-making," *Journal of Medical Systems*, vol. 49, no. 5, pp. 6-6, 2025.
15. T. Papalexopoulos, D. Bertsimas, I. G. Cohen, R. Goff, D. E. Stewart, and N. Trichakis, "Ethics-by-design: efficient, fair and inclusive resource allocation using machine learning," *Journal of Law and the Biosciences*, vol. 9, no. 1, pp. 1–14, 2022.