Machine Learning Does Not Outperform Traditional Statistical Methods When Predicting Kidney Transplant Survival

Aoife McDaid[1], Diarmuid Brady[2], and Tomas Ward[2]

Dublin City University, Dublin, Ireland
aoife.mcdaid3@mail.dcu.ie, diarmuid.brady35@mail.dcu.ie,
tomas.ward@dcu.ie
[1] 19369973 [2] 19479166

Abstract. Kidney transplantation is the most preferred treatment for end-stage renal disease (ESRD) as it provides patients with the best chance of long-term survival and improved quality of life. Previous studies have shown that machine learning (ML) methods have emerged as powerful tools for predicting clinical outcomes, particularly for recognising patterns, relationships, and correlations that may not be apparent using traditional statistical methods. Our data consisted of 1613 donor-recipient pairs with clinical and genetic data. We performed clustering using pre-transplant variables that were available at the time of transplant using the R package *KAMILA* to identify subgroups within the data. Once the clusters were evaluated, we ran separate prediction models on each of these clusters to obtain the most accurate prediction scores. Finally, a comparison was made between traditional statistical methods and ML methods. One statistical model and three machine-learning (ML) models were evaluated on the dataset. These models were evaluated by calculating the Root Mean Squared Error (RMSE) and the R-squared scores. There were two aims of this study. The first was to compare the performance of a statistical model and several ML models in predicting eGFR one year after kidney transplantation. And the second aim of this study was to implement and evaluate the cluster-then-predict ML method for predicting this same target variable. From our research, we found there to be no significant difference between the results produced by the two approaches.

Keywords: Linear Regression · Random Forest Regression · Support Vector Regression · XGBoost Regression.

1 Introduction and Background

Kidney transplantation is the preferred treatment option for end-stage renal disease (ESRD), providing patients with the best chance of long-term survival and improved quality of life. However, the success of kidney transplant outcomes is influenced by several well-established clinical factors including donor age. Genetic predictors of transplant outcome have traditionally been limited to human

leukocyte antigen (HLA) mismatch, but recent work has shown that Polygenic Risk Scores (PRSs) for traits related to hypertension and the estimated glomerular filtration rate (eGFR) in both donors and recipients can have a significant impact on graft outcome. A PRS estimates the cumulative effect of common genetic variation on an individual's disease status.

eGFR is a measure of kidney function that estimates the rate at which blood is filtered through the kidneys. It is measured on a scale from 0 to 120 with higher values associated with greater function, and it is a strong indicator of long-term graft survival and patient outcomes. In recent years, machine-learning (ML) methods have emerged as powerful tools for predicting clinical outcomes. ML models are well suited to recognising patterns, relationships, and correlations that may not be apparent using traditional statistical methods and have achieved broad success across various domains. Previous studies have shown that ML techniques have been used to predict kidney graft failure based on donor and recipient characteristics. We developed ML models to predict recipient eGFR at one-year post-transplant using donor and recipient clinical and genetic factors.

Our research aims to provide insights into the factors that influence kidney transplant outcomes and to develop accurate models for predicting eGFR one year after transplantation by implementing two approaches. The first approach compares statistical methods to ML methods and the second evaluates a cluster-then-predict framework. The rest of the paper is organised into five sections: Literature Review, Methodology, Results, Conclusion and Limitations. The Literature Review section will provide an overview of the existing research, while the Methodology section will describe the research design and procedures. The Results section will present the study findings and their implications for the field. The Conclusion section will provide an overview of the takeaways from the paper and the Limitations section will present some of the shortcomings we have identified in our work.

2 Literature Review

2.1 Imputation Methods

The quality and quantity of data used when training ML models have a profound effect on the model's performance and the insights obtained. Handling missing data is a vital step of the processing phase. However, choosing a suitable missing data method remains a key challenge faced by researchers and it depends on the type of data. [1]

Tsvetanova et al. reviewed the missing data handling methods used in 23 Clinical Prediction Models (CPMs). They found Complete Case Analysis (CCA), the exclusion of patients with any missing values, was the most commonly used method during development (40%) [2]. While this is a simple and easy-to-use method, it does not make the most of the data available. Mean/Median/Mode Imputation, Multiple Imputation and K-Nearest Neighbour (KNN) Imputation were among the five other methods identified.

Tripathi et al. performed a taxonomy of the different types of missing data and several methods to overcome the missing data problem. Multiple Imputation by Chained Equations (MICE) was the preferred option over single imputation methods such as Mean Imputation (strongly biases the dataset, particularly for categorical data) and KNN-Imputation (sensitive to outliers) [3]. MICE was chosen on the basis that it could effectively balance the contribution of continuous and categorical data types.

When handling missing data, it is important to apply the appropriate techniques to achieve the desired results. MICE is the most sensible method to use when imputing data from a mixed dataset due to its flexibility.

2.2 Clustering Methods

Clustering is an unsupervised ML technique that aims to identify patterns or natural groupings called phenotypes within raw data. Phenotyping refers to the process of categorising patients into groups based on their clinical characteristics or outcomes [4]. Phenotyping can be used to inform clinical decision-making, identify risk factors for poor outcomes, and inform future research into the underlying mechanisms of transplant complications.

Thongprayoon et al. performed multi-modal consensus clustering on 3204 obese ($BMI > 40kg/m^2$) kidney transplant recipients. They discovered each cluster possessed distinct clinical profiles, which could help aid decision-making, based on several factors including age, race, and primary kidney disease. The groups also exhibited variable post-transplant outcomes for graft survival. Finally, due to the promising graft survival in all clusters, they proposed reconsidering obesity as a hard barrier to kidney transplantation [5].

Jalakam used multi-modal consensus clustering on 25,824 kidney transplants to find three clusters of kidney recipients that possessed distinct, interpretable characteristics. He discovered there was little research on clustering that effectively handled mixed-type data. He implemented KAy-means for MIxed LArge datasets (*KAMILA*) along with Mixture Models [6].

Mixture models are statistical models that assume a cohort is composed of multiple subgroups. They estimate the parameters of these subgroups' probability distribution and assign each observation to the most likely subgroup based on its characteristics [7]. The *KAMILA* algorithm extends the well-known k-means clustering algorithm to include mixed-type data by using a weighted semi-parametric procedure. The weighting method minimises within-cluster dispersion and maximises between-cluster dispersion using a brute-force search. It balances the contribution of quantitative and qualitative variables [8].

Clustering is an effective exploratory analysis technique used to understand subgroups within a dataset. When applied carefully, it can produce novel insights that can inform clinical decision-making. When working with mixed-typed data it is important to balance the contributions of all data types to achieve optimal clusters. Mixture models and *KAMILA* are two methods that can handle mixed-type data. *KAMILA* in particular, can generate clusters while balancing several data types without specifying weights due to its semi-parametric method.

2.3 Clustering Evaluation

The goal of clustering is to generate subgroups of kidney transplants that exhibit distinct features. Therefore, the evaluation metrics we choose should focus on this goal.

Jalakam evaluated his clusters using several internal indices, comparing the average values of variables for each cluster as well as visualising a t-Distributed Stochastic Neighbor Embedding (t-SNE) [6].

The Silhouette score, and Calinski-Harabasz (CH) index were two of the evaluation metrics used. The silhouette score measures the average difference between the distance of a data point to its cluster versus the distance of that same data point to the nearest neighbouring cluster. The silhouette score is best suited for cases where the number of clusters is not known in advance, as it does not assume a specific number of clusters. The CH index measures the ratio of the between-cluster variance to the within-cluster variance, multiplied by the ratio of the total number of observations to the total number of clusters minus one. The CH index is best used when the number of clusters is known and emphasises the ratio of between-cluster to within-cluster variance [9].

t-SNE is a dimensionality reduction technique commonly used for visualising high-dimensional data in two dimensions. It works by mapping high-dimensional data to a lower dimension while maintaining overall structure. This means that observations (kidney transplants) close/similar to each other in high dimensions will remain close to each other when projected to the lower dimension, thus, maintaining cluster formation. Perplexity and step size are two key parameters of a t-SNE plot. Perplexity sets the number of neighbors while the step parameter sets the learning rate.

Evaluating the quality of a clustering technique requires a variety of measures. In this section, we have reviewed some of the common techniques used which will be implemented in the methodology.

2.4 Predictive Models

Our research focused on kidney transplant outcomes. We conducted a comprehensive review of numerous studies that employed both traditional statistical and ML approaches to analyse and predict the outcomes of kidney transplantation.

Riley et al. reviewed kidney transplant survival prediction models including regression models (Cox proportional hazards (CPH) and LR) and ML models such as Decision Trees (DTs), Random Forests (RFs), and Support Vector Machines (SVMs). The authors noted that many of these predictive models had limitations, such as inadequate sample sizes, limited external validation, and inconsistent reporting of model performance. They emphasised the need for further research to develop and validate more accurate prediction models for kidney transplant outcomes. They also found that while many studies reported good performance of their prediction models, they had inconsistent reporting of their performance metrics. From this paper, we gained a better understanding of the importance of evaluating our predictive models in a way that is most

suitable to our data and our target outcome [10]. These points were also found by Senanayake et al. who aimed to evaluate the ability to use ML algorithms to predict graft failure following kidney transplantation. We found they elaborated on the points made previously, that the studies they reviewed housed some restrictions, including small sample sizes, poor reporting of model performance, and a lack of external validation. They also found that most studies did not consider the impact of missing data [11].

We see that Bae et al. compared the performance of traditional regression models with ML models to predict kidney transplant outcomes. The regression models included logistic regression and CPH regression and the ML models included RFs, SVMs, and neural networks. They used data from the United Network for Organ Sharing (UNOS) which contains data on adult deceased-donor kidney transplant recipients between 2000 and 2016. The authors found that ML models did not outperform traditional regression models in predicting transplant outcomes [12]. Again the use of RFs and SVMs in this paper helped influence the ML models we decided to practice. These traditional methods were again compared to ML methods by Yoo et al., who found that the random survival forest model outperformed CPH in predicting graft survival in kidney transplant recipients.

Truchot et al. investigated the ability of ML models to outperform traditional approaches when predicting kidney allograft outcomes. The data comprised of 8422 kidney transplant recipients gathered from 14 transplant centres in Europe, South America and the USA. The authors implemented ML models such as SVMs, DTs, Gradient boosting, and the traditional approach of a Cox-based prognostication system (CBPS). They compared the ML models to the traditional method by evaluating Harrell's C-index [13]. The authors reported that ML models did not exhibit better performance than traditional statistical methods. However, this finding intrigued us to further investigate this approach using our data. Inspired by this paper, we delved deeper into the ML models used, such as SVMs, DTs, and Gradient Boosting, to explore their potential in improving the accuracy of our predictions.

2.5 Evaluation of Predictive Models

Chicco et al. found that the coefficient of determination (R-squared) is a more informative measure of evaluation than the Symmetric Mean Absolute Percentage Error (SMAPE). They also reported that R-squared does not have the problems surrounding interpretability that Mean Square Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) have. These problems are due to the fact they do not provide any insight into why the model is making certain predictions or how the input variables are contributing to the output. The authors suggest using R-squared as the evaluation metric for regression models [14]. This suggestion greatly impacted our decision on what evaluation metrics to perform to evaluate and compare the results of our models.

After carrying out research into predictive models, we discovered RF, SVMs, and gradient boosting models were the most commonly used. We decided to evaluate each model using the R-squared score and RMSE.

2.6 Cluster-Then-Predict Framework

Trivedi et al. proposed a novel technique for implementing clustering in prediction tasks. Their cluster-then-predict framework involved performing k-means clustering and training a model on each cluster produced. The models predicted a test set to produce k sets of predictions and these were combined in a naive ensemble. This produced superior accuracy in many of the datasets they worked on [15]. Mark et al. implemented a basic version of the cluster-then-predict framework by dividing the dataset using recipient age. A model was created for older (recipient age greater than 50) and younger kidney recipients. They predicted graft survival and discovered differences in feature importance between the two models and cohorts. Age and diabetes diagnosis had the highest predictive power in the older and younger groups, respectively [16].

Based on these findings, we believe there is a gap in the literature for combining the capabilities of both clustering and predictive models to gain greater insight into the factors that influence kidney transplant outcomes.

3 Methodology

3.1 Dataset and Technology

The kidney transplant dataset used in this paper was obtained through the Royal College of Surgeons Ireland. The dataset contains information on 2104 kidney transplants and includes 157 variables, such as pre-transplant factors (e.g., donor and recipient age) and post-transplant factors (e.g., medications taken after surgery). The oldest and most recent transplants took place in 1981 and 2007, respectively and 89.89% were first-time transplant recipients. Recipient age ranged from 18 to 79 (mean equal to 45.93 ± 13.21 standard deviation), 36.26% of which were female. Donor age ranged from 3 to 81 (43.16 ± 15.43), 42.33% of which were female. eGFR at one year had a mean 51.57 ± 18.5 standard deviation. The mean and standard deviation for other key variables included Cold Ischemia Time (93.85 ± 61.95), and HLA mismatches (2.04 ± 1.16). A high proportion of patients had certain binary clinical variables, such as patients on dialysis (95.74%), parenchymal intracranial haemorrhage (PIH) (59.5%), and diabetes (type I or II) (88.98%).

To access the dataset, we utilised King's College London's High-Performance Cluster (HPC). We set up a secure shell (SSH) protocol to connect to the HPC, which allowed us to access a Linux environment. We then created an Anaconda environment containing various Python packages required for analysis. Additionally, we set up Jupyter Labs and RStudio to work with Python and R, respectively.

To run our analyses, we utilised the Slurm Workload Manager to schedule tasks and request compute nodes to run Jupyter Labs or RStudio. By using SSH tunnelling, we redirected the localhost on the Linux cluster to the localhost on our local machine, allowing us to access these interfaces.

3.2 Data Cleaning, Processing and Engineering

The dataset consisted of 2104 observations and 157 variables, however, the majority were not used because either they were post-transplant variables, single-valued, duplicate, or contained too many missing values (greater than 40%). We will refer to variables using their shortened version and a data dictionary of the main variables is provided in the appendix. We inspected the data types of each variable, making conversions where necessary. For example, we set *GraftDate* as a DateTime variable and created numeric versions of categorical variables such as *DonSex* and *RecSex* for analysis.

After checking data types, we performed a visual sanity check by plotting the distribution of each continuous variable and a count plot of the categorical variables using the Seaborn library.

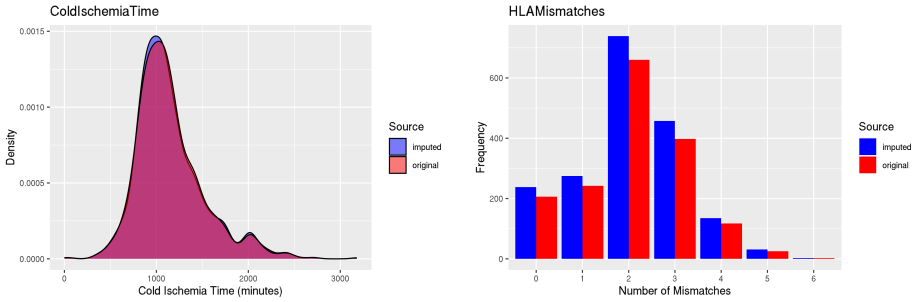
We addressed missing values in three variables: *PrimaryRenalDisease* (34.19%), *HLAMismatches* (12.09%), and *ColdIschemiaTime* (31.15%). To impute the missing data, we used MICE (Multivariate Imputation By Chained Equations) in RStudio. MICE was chosen because it can handle both continuous (using Logistic Regression) and categorical (using Multinomial Logistic Regression) variables.

To evaluate our imputation methods, we used the Kolmogorov-Smirnov test for the continuous variables, *ColdIschemiaTime* and *HLAMismatches*, and the Chi-Squared test for the categorical variable, *PrimaryRenalDisease*. We found that the imputation of the continuous variables worked well, as evidenced by a p-value of 1. However, there was a significant difference in *PrimaryRenalDisease* (p-value $<$ 2.2e-16) due to its long-tail distribution, which included one highly common disease (Polycystic Kidney Disease) and many diseases with low frequencies. Given the challenges with computing this variable, we had to devise an alternative solution. Figure 1 illustrates the distributions of the original and imputed continuous variables.

After cleaning the data, we were left with 1613 observations and 46 variables (including outcome variables). We engineered new features, including squared and cubed variables, to capture non-linear relationships. We created an age difference and sex mismatch variable to capture donor-recipient dependence and extracted the ten most common diseases from the *PrimaryRenalDisease* variable to create binary variables for each disease. We also extracted the year, season, month, and day from the *GraftDate*.

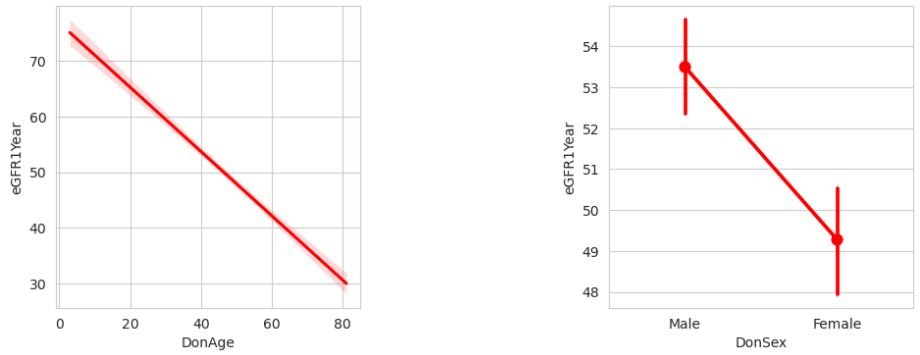
We performed feature selection using LR. We selected features with a significant relationship with *eGFR1Year* (p-value $<$ 0.05). Using *regplot* and *pointplot* from Seaborn, we visualised continuous and categorical variables, respectively as shown in Figure 2.

From this analysis, we selected fifteen variables including donor and recipient age and sex, sex mismatch, year of transplant, intracranial haemorrhage, recipi-



(a) Cold Ischemia Time Distributions. (b) HLA Mismatches Distributions.

Fig. 1: Comparison of original and imputed distributions.



(a) Donor Age Regression Plot. (b) Donor Sex Regression Plot.

Fig. 2: Regression Plots of eGFR with Continuous and Categorical Variables.

ent first principal component, recipient PRSs (hypertension and height-adjusted kidney volume), donor PRSs (hypertension, eGFR, and stroke), cold ischemia time, and graft number.

3.3 Data Modelling - Clustering

Clustering analysis was performed to identify subgroups within the data that shared characteristics that were useful in a clinical setting. Two clustering approaches were applied using the *mclust* and *KAMILA* libraries in r. We chose these models because they could effectively handle mixed-type data.

We ran *mclust* using the default parameters. We determined the number of clusters using the Bayesian Information Criterion (BIC) for values of k from two to ten. Three clusters had the lowest BIC score indicating the best fit for the model.

Next, we ran *KAMILA* and the number of continuous and categorical variables had to be the same. Therefore we chose the features with the highest

predictive value of *eGFR1Year* based on LR. Categorical variables included *DonSex*, *RecSex*, *SexMismatch* and *IntracranialHaemorrhage*. Continuous variables included *DonAge*, *RecAge*, *RecPC1*, and *RecipientHypertensionPRS*. We ran ten-fold cross-validation to determine the number of clusters and set the number of initialisations to ten.

We evaluated models using internal evaluation indices, statistical testing of mean differences, and t-SNE plots. We chose the silhouette score and the CH index to evaluate the compactness and separateness of the clusters. Next, we investigated cluster differences using a 95% confidence interval of each variable mean across all clusters and placed the results in a table. Finally, we ran t-SNE algorithms in both R and Python. We ran hyperparameter tuning on perplexity and step size and evaluated each run using Kullback-Liebler (KL) divergence. The KL divergence is a measure of the difference between two probability distributions, in this case, the high-dimensional data distribution and the lower-dimensional embedding distribution. The lower the KL divergence between the two distributions, the better the quality of the visualisation.

3.4 Data Modelling - Prediction

To predict the values of the eGFR at 1 year for each donor and recipient pairing, we implemented two approaches. The first approach consisted of running the regression models on the data as a whole and the second approach followed a cluster-then-predict outline.

Implementing Regression Models on Data as a Whole. We used four different regression models to predict the eGFR1Year variable: Linear Regression (LR), eXtreme Gradient Boosting Regression (XGBR), Random Forest Regression (RFR), and Support Vector Regression (SVR). LR is a traditional statistical model that aims to find a linear relationship between the predictor variables and the target variable. XGBR and RFR are both ML models that use DTs to make predictions. SVR is an ML model that uses support vectors to find the best decision boundary between the predictor variables and the target variable. By comparing the performance of these models, we gained insights into which model is best for predicting the eGFR 1 year post-transplant.

We used an 80/20 split on our dataset to evaluate the performance of our models. This approach allowed us to avoid overfitting, and assess the generalisability of our models.

Hyperparameter tuning optimises a model's performance by finding the best combination of hyperparameters. We conducted hyperparameter tuning for each of our models. We utilised Grid Search, an imported tool from the Python library, Scikit-learn (sklearn), to evaluate various hyperparameter combinations and identify the optimal parameters for each model [17].

Implementing Regression Models on each of the Clusters. We implemented the cluster-then-predict framework. Using the previously generated clusters, we assigned each donor-recipient pair to their respective groups resulting in

five dataframes. We evaluated these same models (LR, XGBR, RFR, SVR) on each dataframe. We carried out the same approach as seen above, implementing an 80/20 split, hyperparameter tuning and regularisation on each of the models to produce the best possible results. Overall, the cluster-then-predict framework we used allowed us to take advantage of the subgroups within our data and tailor our predictions to each one.

Evaluation of Predictive Models. We assessed the performance of our regression models using RMSE, MAE, and R-squared. RMSE measures the square root of the average of the squared differences between predicted and actual eGFR values, while MAE measures the average absolute difference between predicted and actual values. A lower RMSE and MAE indicate better performance. R-squared is a measure of how much of the dependent variable's variance is explained by the independent variables. It ranges from 0 to 1, with a value of 1 indicating that all the variance in the depended is explained by the independent variables.

We evaluated the statistical and ML regression models using these metrics and we compared the two approaches using the Student's t-test to determine statistical significance. This test is used to determine the statistical significance of the difference between two sets of data by calculating a t-value based on the means, standard deviation, and sample size of each group.

We then compared running the regression models on the entire dataset with a cluster-then-predict approach, evaluating each model's performance on each cluster using the same metrics. We analysed the cluster results separately to identify the best models for each cluster, then calculated the average RMSE, MAE, and R-squared scores across all clusters to assess overall performance. We used the average scores to compare the cluster-then-predict results with the results generated by running regression models on the entire dataset. We used a Student's t-test to compare these scores for statistical significance.

4 Results

4.1 Cluster Analysis

Internal evaluation indices A silhouette score of -0.094 and 0.029 and a CH index of 21.571 and 47.482 was calculated for *KAMILA* and *mclust*, respectively. These results indicate *mclust* produced more compact and separate clusters as higher scores are associated with higher quality clusters. This may be due to the reduction of features used in *KAMILA*. The negative silhouette score for *KAMILA* means that the distance between a point and its cluster was larger than the distance to other clusters indicating it was assigned to the wrong cluster.

Mean differences between clusters Despite previous results, *KAMILA* showed greater separateness between clusters than *mclust*. Table 2 displays the mean differences between each cluster for each continuous variable from the KAMILA

clusters. Variables with significant differences are denoted (S) and (NS) for non-significant. mclust had no significant difference between clusters. This is why we chose to move forward with *KAMILA* for clustering.

Cluster 1 had young donors and the youngest recipients and the highest *eGFR1Year* and *GraftSurvivalDays*, indicating better kidney function and longer graft survival. Cluster 2 had the oldest donors and recipients. This cluster also has the lowest graft survival. This demonstrates the strong relationship between age and transplant outcomes. Cluster 3 had no defining characteristics but could be considered the median cluster. Cluster 4 had the youngest donors while having older recipients, we also see that this group of transplants were performed furthest in the past. Recipients had the second-highest survival outcomes suggesting that a young donor age could compensate for older donors. Cluster 5 had older donors and recipients and poor survival outcomes, reinforcing the relationship between age and eGFR at 1 year.

Based on this analysis, it appears that clusters 1 and 2 show the greatest differences between each other. Age is the key variable associated with graft survival both for $eGFR1Year$ and $GraftSurvivalDays$. While several of the PRS were not significant. This is likely due to not being included in the clustering algorithm.

t-SNE Plots The results from t-SNE plots were not interpretable. Despite implementing the algorithm in Python and r, the algorithm would not properly converge and patterns found in the plots were mostly due to randomness. These included spirals and streaking lines. Despite performing fine-tuning on perplexity and step size, this issue could not be overcome.

4.2 LR versus ML Models

The RMSE and R-Squared scores for the four models are as follows, LR (15.38, 0.297), XGBR (15.18, 0.315), RFR (15.39, 0.30) and SVR (15.41, 0.295). Table 3 illustrates these results. While these results suggest that the XGBR has the highest predictive power of the four models, we found that there was no significant difference between LR and the other models (p-values of 0.66, 0.87, and 0.38 respectively with confidence level 0.05). This indicates that LR is comparable to the other three models in terms of its predictive performance and therefore the ML models don't outperform the traditional LR method.

4.3 Cluster-Then-Predict Method

In addition to comparing the performance of four regression models, we also investigated a cluster-then-predict approach to predicting the eGFR at 1 year. The results obtained are as follows.

The best scores for each cluster are as follows. The RFR was most effective for cluster 1 (15.34, 0.18) Table 4, cluster 2 (15.38, 0.07) Table 5, cluster 4 (17.13,

-0.17) Table 7, and cluster 5 (15.32, 0.01) Table 8,. Finally, SVM performed best on cluster 3 (17.72, 0.09) Table 6,.

To compare the performance of the cluster-then-predict approach to the original XGBR, we calculated the average RMSE and R-squared scores for each of the regression models across all five clusters. From this, we obtained an average RMSE score of 16.18 and an average R-squared score of 0.035. We then compared the error values to the scores obtained by the original XGBR using a Student t-test. This obtained the p-values 0.49 (cluster 1), 0.09 (cluster 2), 0.09 (cluster 3), 0.42 (cluster 4) and 0.6 (cluster 5). This told us that there was no statistically significant difference between the two approaches.

5 Conclusion

The first part of our study aimed to compare the performance of a statistical model and several ML models in predicting eGFR one year after kidney transplantation. It was discovered that both the statistical model and the ML models performed equally well with similar RMSE and R-squared scores. Therefore, the distinct value of applying ML approaches to this problem requires further justification. By evaluating multiple models, we were able to gain a comprehensive understanding of the strengths and weaknesses of each model. However, further research is needed to determine whether implementing different ML models or using a larger dataset would impact the results obtained. While the study did not find a clear advantage of applying ML approaches to this specific problem, they could offer benefits in other areas of kidney transplant prediction such as the identification of new risk factors and donor-recipient matching.

6 Limitations

While our study demonstrates the similar performance of ML models with traditional statistical models, it is important to acknowledge some limitations and shortcomings that we identified upon retrospection. Firstly, during the clustering phase of our project, we used *KAMILA* and *mclust* libraries alone. However, previous research has shown multi-modal consensus clustering to produce superior results. When performing clustering analysis, using standardised mean difference would have been preferable to calculating the 95% Confidence Interval as this was the standard method used in previous works. Additionally, when implementing the prediction models on top of the clusters we could have used them to predict the entire test set, producing a naive ensemble, rather than predicting only the within-cluster observations. In conclusion, while these limitations must be acknowledged, they do not diminish the significance of the findings and the overall contribution of this research to the field.

References

[1] I. Ezzine and L. Benhlima, "A Study of Handling Missing Data Methods for Big Data," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, ISSN: 2327-1884, Oct. 2018, pp. 498–501. DOI: 10.1109/CIST.2018.8596389.

[2] A. Tsvetanova, M. Sperrin, N. Peek, I. Buchan, S. Hyland, and G. Martin, "Inconsistencies in handling missing data across stages of prediction modelling: A review of methods used," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, ISSN: 2575-2634, Aug. 2021, pp. 443–444. DOI: 10.1109/ICHI52183.2021.00076.

[3] A. K. Tripathi, G. Rathee, and H. Saini, "Taxonomy of Missing Data along with their handling Methods," in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, ISSN: 2640-074X, Nov. 2019, pp. 463–468. DOI: 10.1109/ICIIP47207.2019.8985715.

[4] MCP, *What is Phenotyping? – The Department of Molecular & Comparative Pathobiology*, en-US. [Online]. Available: https://mcp.bs.jhmi.edu/what-is-phenotyping/ (visited on 01/09/2023).

[5] C. Thongprayoon, S. Mao, C. Jadlowiec, *et al.*, "Machine Learning Consensus Clustering of Morbidly Obese Kidney Transplant Recipients in the United States," *Journal of Clinical Medicine*, vol. 11, p. 3288, Jun. 2022. DOI: 10.3390/jcm11123288.

[6] K. K. Jalakam, "Multi-Modal Consensus Clustering to Identify Phenotypes of Kidney Transplant Donors and Recipients and Their Association With Survival," en, Nov. 2022, Accepted: 2022-11-21T13:19:35Z. [Online]. Available: https://DalSpace.library.dal.ca//handle/10222/82083 (visited on 01/07/2023).

[7] D. Reynolds, "Gaussian Mixture Models," en, in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds., Boston, MA: Springer US, 2009, pp. 659–663, ISBN: 978-0-387-73003-5. DOI: 10.1007/978-0-387-73003-5_196. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_196 (visited on 01/10/2023).

[8] A. Foss, M. Markatou, B. Ray, and A. Heching, "A semiparametric method for clustering mixed data," en, *Mach Learn*, vol. 105, no. 3, pp. 419–458, Dec. 2016, ISSN: 1573-0565. DOI: 10.1007/s10994-016-5575-7. [Online]. Available: https://doi.org/10.1007/s10994-016-5575-7 (visited on 01/09/2023).

[9] N. Bolshakova and F. Azuaje, "Cluster validation techniques for genome expression data," en, *Signal Processing*, Genomic Signal Processing, vol. 83, no. 4, pp. 825–833, Apr. 2003, ISSN: 0165-1684. DOI: 10.1016/S0165-1684(02)00475-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168402004759 (visited on 05/05/2023).

[10] S. Riley, Q. Zhang, W.-Y. Tse, A. Connor, and Y. Wei, "Using Information Available at the Time of Donor Offer to Predict Kidney Transplant Survival Outcomes: A Systematic Review of Prediction Models," *Transplant International*, vol. 35, p. 10 397, Jun. 2022. DOI: 10.3389/ti.2022.10397.

[11] S. Senanayake, N. White, N. Graves, H. Healy, K. Baboolal, and S. Kularatna, "Machine learning in predicting graft failure following kidney transplantation: A systematic review of published predictive models," *International Journal of Medical Informatics*, p. 103 957, Aug. 2019. DOI: 10.1016/j.ijmedinf.2019.103957.

[12] S. Bae, A. B. Massie, B. S. Caffo, K. R. Jackson, and D. L. Segev, "Machine learning to predict transplant outcomes: Helpful or hype? a national cohort study," *Transplant International: Official Journal of the European Society for Organ Transplantation*, vol. 33, no. 11, pp. 1472–1480, Nov. 2020, ISSN: 1432-2277. DOI: 10.1111/tri.13695.

[13] A. Truchot, M. Raynaud, N. Kamar, *et al.*, "Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction," *Kidney International*, vol. 103, no. 5, pp. 936–948, May 1, 2023, Publisher: Elsevier, ISSN: 0085-2538. DOI: 10.1016/j.kint.2022.12.011. [Online]. Available: https://www.kidney-international.org/article/S0085-2538(22)01084-5/fulltext (visited on 04/22/2023).

[14] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, e623, Jul. 5, 2021, Publisher: PeerJ Inc., ISSN: 2376-5992. DOI: 10.7717/peerj-cs.623. [Online]. Available: https://peerj.com/articles/cs-623 (visited on 04/22/2023).

[15] S. Trivedi, Z. Pardos, and N. Heffernan, "The Utility of Clustering in Prediction Tasks," Sep. 2015.

[16] E. Mark, D. Goldsman, B. Gurbaxani, P. Keskinocak, and J. Sokol, "Using machine learning and an ensemble of methods to predict kidney transplant survival," en, *PLOS ONE*, vol. 14, no. 1, e0209068, Jan. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0209068. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209068 (visited on 10/17/2022).

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A Appendix A

Table 1: Variable definitions

Variable	Name Meaning	Data Type
$DonAge$	Age of donor	Numeric
$RecAge$	Age of recipient	Numeric
$DonSex$	Sex of donor	Categorical
$RecSex$	Sex of recipient	Categorical
$SexMismatch$	Sex of recipient	Categorical
$AgeDifference$	Donor age minus recipient age	Categorical
$GraftDate$	Date of transplant	DateTime
$Year$	Year of transplant	Numeric
$ColdIshemiaTime$	Time for kidney spent on ice between removal and transplantation	Numeric
$PrimaryRenalDisease$	The primary disease responsible for recipient kidney failure	Categorical
$RecPC1$	Genetic Principle Component of Recipient	Numeric
$DonHypertensionPRS$	PRS of hypertension for donor	Numeric
$RecHypertensionPRS$	PRS of hypertension for recipient	Numeric
$DonStrokePRS$	PRS of stroke for donor	Numeric
$RecStokePRS$	PRS of stroke for recipient	Numeric
$DoneGFRPRS$	PRS of eGFR for donor	Numeric
$ReceGFRPRS$	PRS of eGFR for recipient	Numeric

Table 2: Cluster Mean Comparison Using 95% CI with (max) and (min)

Cluster No. (N=1613)	**1 (n = 250)**	**2 (n = 362)**	**3 (n = 418)**	**4(n = 197)**	**5 (n = 386)**
DonAge (S)	(23.79, 25.77)	(59.67, 60.85)	(49.63, 51.05)	(20.1, 21.88)	(43.38, 44.47)
RecAge (S)	(29.8, 31.38)	(57.22, 58.8)	(34.44, 35.74)	(50.68, 52.73)	(53.64, 55.17)
Year (S)	(1999.2, 2000.39)	(2001.27, 2002.04)	(2000.51, 2001.31)	(1998.89, 2000.37)	(2000.8, 2001.57)
RecPC1 (S)	(0.0027, 0.0058)	(-0.0022, -0.0001)	(-0.0007, 0.0014)	(0.0021, 0.0056)	(0.0007, 0.0029)
RecHypertensionPRS (NS)	(-0.15, 0.09)	(-0.15, 0.04)	(-0.13, 0.07)	(-0.19, 0.07)	(-0.17, 0.02)
DonHypertensionPRS (NS)	(-0.25, 0.0)	(0.07, 0.29)	(0.12, 0.31)	(-0.35, -0.07)	(0.11, 0.31)
DoneGFRPRS (NS)	(0.02, 0.24)	(-0.15, 0.06)	(-0.05, 0.13)	(-0.12, 0.17)	(-0.03, 0.18)
DonStrokePRS (NS)	(-0.01, 0.25)	(-0.02, 0.18)	(-0.03, 0.16)	(-0.1, 0.2)	(-0.04, 0.16)
RecHAKVPRS (NS)	(-0.12, 0.13)	(-0.13, 0.07)	(-0.06, 0.13)	(-0.04, 0.23)	(-0.14, 0.05)
ColdIschemiaTime (NS)	(1117.23, 1217.81)	(1088.95, 1163.09)	(1117.71, 1184.01)	(1104.04, 1220.67)	(1129.57, 1207.1)
GraftNo (S)	(1.093, 1.179)	(1.041, 1.092)	(1.103, 1.169)	(1.027, 1.095)	(1.082, 1.146)
eGFR1Year (S)	(63.45, 67.95)	(40.2, 43.25)	(47.09, 50.5)	(59.49, 64.11)	(48.33, 51.54)
GraftSurvivalDays (S)	(3188.91, 3601.09)	(2473.0, 2688.73)	(2754.21, 3000.04)	(2982.59, 3416.56)	(2665.48, 2912.85)

Table 3: Regression Models RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	15.3837	0.29697
XGBoost Regression	15.1818	0.315303
Random Forest Regression	15.3867	0.296694
Support Vector Regression	15.41	0.294569

Table 4: Cluster 1 RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	16.6767	0.0265268
XGBoost Regression	15.7266	0.134282
Random Forest Regression	**15.335**	**0.176858**
Support Vector Regression	16.5512	0.0411167

Table 5: Cluster 2 RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	15.8055	0.0129758
XGBoost Regression	15.8753	0.00423524
Random Forest Regression	**15.3793**	**0.0654917**
Support Vector Regression	16.2793	-0.0470841

Table 6: Cluster 3 RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	18.1133	0.0489944
XGBoost Regression	19.0387	-0.0506554
Random Forest Regression	18.2673	0.0327552
Support Vector Regression	**17.7225**	**0.0895876**

Table 7: Cluster 4 RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	17.3739	-0.200554
XGBoost Regression	18.2315	-0.321991
Random Forest Regression	**17.1262**	**-0.166557**
Support Vector Regression	17.1431	-0.168858

Table 8: Cluster 5 RMSE and R-Squared values

Regression Model	RMSE	R-Squared
Linear Regression	16.2202	-0.108311
XGBoost Regression	15.8401	-0.0569823
Random Forest Regression	**15.3228**	**0.010936**
Support Vector Regression	15.4274	-0.00261412