Biomedical Signal Processing and Control 52 (2019) 456–462

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation

Torgyn Shaikhina[a], Dave Lowe[b], Sunil Daga[d,e], David Briggs[c], Robert Higgins[e], Natasha Khovanova[a,*]

[a] School of Engineering, University of Warwick, Coventry, CV47AL, UK
[b] Royal Liverpool and Broadgreen University Hospital NHS Trust, Liverpool, UK
[c] NHS Blood and Transplant Birmingham, UK
[d] Warwick Medical School, UK
[e] University Hospitals Coventry and Warwickshire NHS Trust, UK

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form
15 December 2016
Accepted 24 January 2017
Available online 9 February 2017

Keywords:
Machine Learning
Small data sets
Biomedical systems
Decision Tree
Random Forest
Kidney transplants
Antibody-mediated acute rejection

ABSTRACT

Clinical datasets are commonly limited in size, thus restraining applications of Machine Learning (ML) techniques for predictive modelling in clinical research and organ transplantation. We explored the potential of Decision Tree (DT) and Random Forest (RF) classification models, in the context of small dataset of 80 samples, for outcome prediction in high-risk kidney transplantation. The DT and RF models identified the key risk factors associated with acute rejection: the levels of the donor specific IgG antibodies, the levels of IgG4 subclass and the number of human leucocyte antigen mismatches between the donor and recipient. Furthermore, the DT model determined dangerous levels of donor-specific IgG subclass antibodies, thus demonstrating the potential of discovering new properties in the data when traditional statistical tools are unable to capture them. The DT and RF classifiers developed in this work predicted early transplant rejection with accuracy of 85%, thus offering an accurate decision support tool for doctors tasked with predicting outcomes of kidney transplantation in advance of the clinical intervention.

1. Introduction

Machine Learning (ML) pertains to the ability of data-driven models to "learn" information about a system directly from observed data without predetermining mechanistic relationships that govern the system. ML algorithms are able to adaptively improve their performance with each new data sample and discover hidden patterns in complex heterogeneous and high dimensional data [1–3]. ML has become the core technology for numerous real-world applications: from weather forecasting and DNA sequencing, to Internet search engines and image recognition [1,4–6].

In clinical and biomedical engineering domain ML offers predictive models, such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs), which are able to map highly non-linear heterogeneous input and output patterns even when physiological relationships between model variables could not be determined due to complexity, pathologies, or lack of biological understanding [6,5,7,8]. Nevertheless, ML models are rarely viewed in the context of small data, where insufficient number of training samples can compromise the learning success [9,10]. Among various ML classifiers, DTs are particularly well suited for clinical classification task. DTs are easy to interpret by non-statistician and are intuitive to follow. They cope with missing values and are able to combine heterogeneous data types into a single model, whilst also performing an automatic principal feature selection [8,11].

With advances in immunosuppressive drugs and surgical techniques organ and tissue transplantation is recognised as an effective treatment for many pathologies including end-stage kidney (renal) disease. Organ transplantation can dramatically improve patients' quality of life, often offering the only solution for their survival [12]. This area, however, does not generate large patient datasets due to the severity of accompanying diseases and the complexity of clinical operations, thus preventing wide application of ML techniques for the prediction of transplant outcomes. There are few studies available in the literature demonstrating recent advances in the emerging area of ML applications to kidney transplantation.

[*] Corresponding author.
E-mail address: N.Khovanova@warwick.ac.uk (N. Khovanova).

https://doi.org/10.1016/j.bspc.2017.01.012
1746-8094/

Greco et al. studied long-term kidney allograft survival and came to the conclusion that "decision trees in clinical practice may be a suitable alternative to the traditional statistical methods, since it may allow one to analyse interactions between various risk factors beyond the previous knowledge" [13]. Their DT model, based on 194 patients with 9 known clinical indicators, predicted a 5 year allograft survival with test accuracy of 74%–88%.

Krikov et al. in their large-scale, multi-centre study [14] analysed 92,844 patient records from the US Renal Data System. Their DT model for long-term kidney allograft survival was based on 31 predictors and the accuracy of the model was measured by the Area Under the Receiver Operating Characteristic (ROC) Curve (*AUC*) [15]. The DT models achieved *AUC* of 0.63, 0.64, 0.71, 0.82, and 0.90 (note that an AUC of 1 corresponds to the perfect model accuracy) for the 1, 3, 4, 5, and 10 year predictions, respectively. The trend – the further into the future the forecast scope is, the better its accuracy – appears unintuitive to those working with real-world forecasts. This phenomenon can be explained in part by the way the model accuracies were *measured* and how this was influenced by the reduced follow-up and class imbalance dynamics over the years as more allografts fail.

Decruyenaere et al. compared traditional logistic regression method with 8 different ML algorithms for prediction of delayed graft function (DGF) following kidney transplantation [16]. Their models were developed on 497 single-centre (Belgium) patients from deceased donors and used 24 parameters related to the donor and recipient characteristics, preservation and operation. The authors found that tree-based models achieved low accuracy: *AUC* of 0.53 for DT and 0.74 for RF respectively, which again can be attributed to high class imbalance between DGF + ve (12.5%) and DGF-ve samples. Out of 10 classifiers, a linear SVM performed best with *AUC* of 0.84.

The models in the above studies were developed with a *few hundreds* to a few *tens of thousands* of samples involving national databases. The ratio of the number of observations x to the number of predictor features p was greater than *20* in the four above applications. Such ratio may not, however, be feasible for smaller transplant units wishing to analyse their samples without having to wait for decades until enough operations are conducted. In such cases, when $x/p < 10$, ML modelling faces the challenges of volatility of outcomes among models of the same design due to insufficient data.

For example, Lofaro et al. attempted to predict chronic allograft nephropathy within 5 years post-transplant from 23 clinical indications based on only 80 samples ($x/p = 3.5$) [17]. The authors used DT model and chose to report one of the tree models that has the largest *AUC* (*AUC* = 0.847, 62.5% sensitivity, 7.2% false-positive rate) and another tree with a different structure (*AUC* = 0.824, 81.3% sensitivity and 25% false-positive rate). The volatility among the DT trials were not explicitly disclosed, but the two presented DT models showed significant variation in performance and hierarchy.

Such volatility in performance of ML models is not unique to the study by Lofaro et al. [17]. On the contrary, high variablitity among the models of the same design is characterisic of small-data applications of ML algorithms which have embedded degrees of randomness in their training and initialisation routines. In previous work we have shown that identical ANNs suffer from large discrepancies in their predictions due to randomised initial conditions, training order and the split between the training and validation samples [18]. Large descripances in predictions based on small datasets are common for other ML approaches. To extend the benefits of ML to a wider range of models for clinical applications, it is essential to develop methods that would cope with the limited data size. We have previously developed a framework for application of ANN to *regression* tasks based on small data, which enabled consistent comparisons between various ANN designs and quantification of random effects and led to successful and robust regression models [18–20].

The current study *aims* to adress small-data applications of ML models for *classification* tasks. Specifically, the paper considers Decision Trees (DTs) and Random Forests (RFs) for early prediction of acute antibody-mediated rejection (ABMR) in kidney transplantion based on pre-operative (baseline) clinical indicators.

For a successful transplantation outcome, the recipient and donor should be matched for tissue proteins called human leukocyte antigen (HLA). HLA mismatches between the transplant recipient and their donor may cause the development of antibodies aginst HLA, which can subsequently lead to transplant failure and endanger a future transplant if it has an HLA type reactive with the antibodies. HLA antibodies can also be stimulated by pregnancy and blood transfusion. Patients with *preformed* HLA donor-specific antibodies (DSAs) have longer waiting times for surgery or are unable to receive a renal transplant. Antibody incompatible transplantation (AIT), now well-established [21,22], allows one to decrease DSA levels prior to transplantation and operate on patients with HLA mismatches. However, above 40% of kidneys still experience a rejection episode, because complete elimination of DSAs and immunological memory is not practical. Neither types of *harmful* DSAs nor their *acceptable levels* before transplantation are known. Among isotypes of HLA antibodies, Immunoglobulin IgG is predominant and is considered to be the agent of humoral rejection. Its four subclasses (IgG1-4) exhibit functional differences and associate with differences in clinical outcome [23]. Our earlier work using conventional statistical analysis based on logistic regression revealed that IgG4 subclass presents a significant risk factor for ABMR in AIT [23].

The *primary objectives* of this study is by using ML to independently confirm the key risk factors associated with early (within first 30 days following the transplantation) ABMR and to find baseline *levels* of DSAs for *safe transplantation*, i.e. *how* much of DSAs can be tolerated to make certain that the donor kidney is safely accepted. Note that the solution of the latter task is not feasible by conventional statistical analysis. The *secondary objective* is to produce an accurate patient-specific predictive model using DT and RF-based methods in order to support clinical decision-making.

2. Methods

2.1. The data

80 patients who received HLA incompatible renal allografts between 2003 and 2012 are included in the study: 49 female and 31 male patients with average age 41.8 ± 11.6 years (range = 18–68 years) at time of transplantation. Full description of patients' baseline characteristics can be found in [23]. Antibody levels were measured in serum taken before antibody reduction treatments using a fluorescence immunoassay and are given as Median Fluorescence Intensity (MFI) [24].

The following 14 baseline (measured before transplantation) parameters comprised the input feature set for the model:

- 7 *continuous:* highest IgG DSA MFI level, patient's age, years on dialysis (ESRD duration), and 4 total IgG subclass (1–4) MFI levels
- 4 *categorical*: cytometery cross-match (1 = bead, 2 = flow or 3 = CDC), total number of HLA mismatches between donor and recipient (0–6), the number of class II HLA-DR mismatches (0–2), and the number of previous transplants (0–2),
- 3 *binary:* gender (male/female), the presence of both HLA Class I and Class II DSA (yes/no), and a marker of whether the transplant was from live or deceased donor.

The output, i.e. occurrence or absence of ABMR, was a two-class binary variable, where ABMR + ve corresponded to the early rejection of kidney (class '1′), and ABMR-ve corresponded to rejection-free outcome (class '0′).

Data from 60 patients, sampled at random, were used for model training and the remaining 20 samples were reserved for independent tests. Same test cohort was used for both the final DT and RF models. The data were well balanced (46 ABMR + ve and 34 ABMR-ve samples), but contained 3 samples with partially missing fields. In one of the samples the ESRD duration was lost upon collection; in two other samples IgG1-3 values were not recorded. The 3 missing samples were included in our study to ensure that the models are able to make predictions on incomplete data, which are commonly encountered in clinical setting [11].

2.2. Decision Tree (DT)

As implied in its name, DT is a tree-like structure, where leaves represent outcome labels, i.e. ABMR + ve (1) or ABMR-ve (0), and branches represent conjunctions of input features that resulted in those outcomes.

A binary DT separates the data (parent node) into two subsets (child nodes) by calculating the best feature split determined by a chosen split criterion. The two resulting subsets become the new parent nodes and are subsequently divided further into two child nodes. The binary split continues until all observations are classified. The algorithm is nonparametric, i.e. no assumptions are made regarding the underlying distribution of the predictor variables.

2.2.1. DT design in this study

The DT design in the present study was based on the standard CART algorithm implemented using MATLAB^{TM} [25]. Throughout the training process, the dataset was recursively divided according to the split criterion until the optimal DT hierarchy of nodes was reached. The split optimisation criterion used in this DT model is the Gini's Diversity Index (GDI), which is a measure of node impurity. The node is considered pure when it contains only observations of one class (either ABMR + ve or ABMR-ve); the GDI of a pure node is equal to 0 [26]. The following additional constraints were imposed on the DT size: minimum 10 observations for the node to become a branch node and at least 1 observation per a leaf node. The experiment with DT was repeated 600 times and each time a different model subset was sampled out of the original samples. It was expected to observe high volatility among the performance of those 600 DTs.

2.2.2. Categorical vs continuous predictors

Notably, for a DT classifier, finding an optimal binary split for a continuous predictor is far less computationally intensive than for a categorical predictor with multiple levels. In the former case, DT can split between any two adjacent values of a continuous vector, but for a categorical predictor with i levels, all of the 2^{i-1}–1 splits need to be considered in order to find the optimal one. As an example: to identify the optimal split for the total number of HLA mismatches ($i = 7$) the DT had to consider 63 possibilities.

2.2.3. Pruning

Pruning is a ML technique which can reduce the size of a DT and prevents overtraining. Pruning is achieved by removing the nodes that have least effect on the overall classification performance [25]. In this work pruning was applied in order to penalise complexity of the DT, thus to ensuring only the most significant splits are discovered by the model.

2.3. Random Forest (RF)

RF (or Bagged DTs) is an ensemble method in machine learning which involves construction (growing) of multiple DTs via bootstrap aggregation (bagging) [27–29]. In other words, each time an input is supplied to RF that input is passed down each of the constituent DTs. Each tree predicts a classification independently and "votes" for the corresponding class. The majority of the votes decides the overall RF prediction [8,30]. This aggregate vote of several DTs is inherently less noisy and less susceptible to outliers than a single DT output, which mitigates the volatility due to small data and improves the robustness of predictions [28,29,31].

RF has a built-in feature selection system and thus can handle numerous input parameters without having to delete some parameters for reduced dimensionality. Variable importance scores for RF can be computed by measuring the increase in prediction error if the values of a variable under question are permuted across the out-of-bag observations (this is called permutation test). This score is computed for each constituent tree, averaged across the entire ensemble and divided by the standard deviation.

The RF was comprised of 600 fully grown trees. This number of trees was selected to correspond to the number of individual DTs considered in Section 2.2.1. Although it was expected that the RF model would produce substantially more robust results than 600 DTs, the experiment with RF was repeated 10 times to monitor for the variance due to small data.

A constituent tree of RF is different from a DT in 2.1.1 in following ways:

- Overfitting was controlled by out-of-bag validation at 90% of the samples, as opposed to DT pruning
- Minimum number of samples per leaf node was increased to 3 in order to compensate for otherwise very large trees grown without pruning
- Only a subset of 14 original input parameters, i.e. 9 predictor variables, was used. The reduction in the number of the input parameters was according to the classical model developed in our previous study [23]. In particular, it was demonstrated [23] that recipients age, ESRD duration, the number of class II HLA-DR mismatches, the number of previous transplants, and the marker of whether the donor was a live/diseased were found to be statistically insignificant and their inclusion into consideration reduced the quality of the ABMR model. Out of the nine, six predictor variables were sampled at random for each partial-feature tree in the RF.

2.4. Performance metrics

Predictions made by a DT are continuous real-valued numbers in the range between 0 and 1, which describe the probability of ABMR, however, the expected values recorded for each patient are binary (1 = ABMR + ve, 0 = ABMR-ve). In order to convert the continuous predictions into binary class labels, a 50% (0.5) cut-off point was applied. The difference between predicted and expected binary outcomes was then described by the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), where the sum of $TP + TN + FP + FN = n$ is the total number of observations. The following standard performance metrics were used to assess the accuracy of both RF and DT:

- **Correct classification rate**, C, *measures the proportion of correctly identified observations of both classes:* $C = (TP + TN)/n$
- **Positive predictive value** (or precision), $PPV = TP/(TP + FP)$
- **Negative predictive value**, $NPV = TN/(TN + FN)$

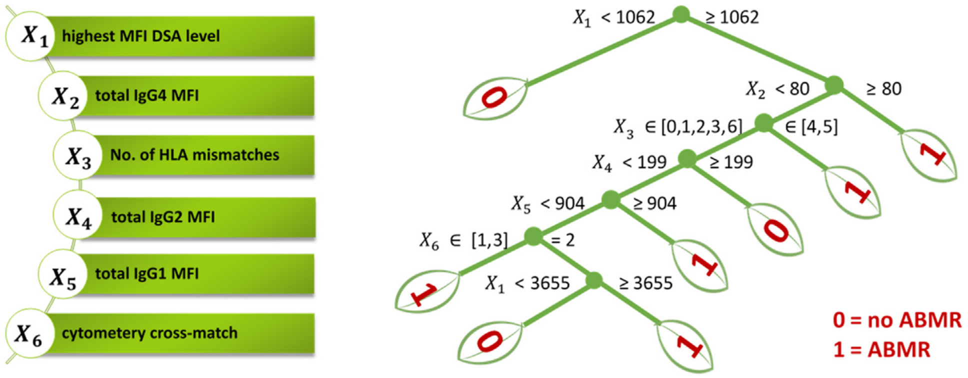

Fig. 1. DT model schematic showing the split hierarchy with 7 branch nodes and 8 leaf nodes based on 6 variables X_1 to X_6.

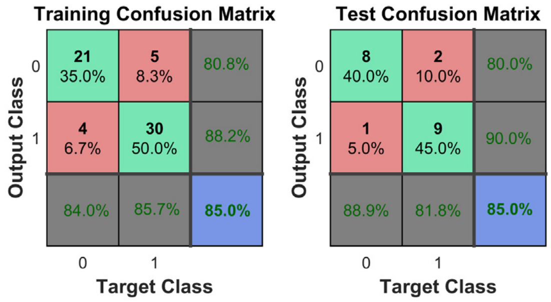

Fig. 2. Confusion matrices for the training dataset (left) and the test samples (right) for DT model, where the squares provide the performance metrics described in Section 2.4. In each confusion matrix, the green squares correspond to *TP* and *TN* values, and the red squares represent *FP* and *FN* values. The four grey squares (from the top right, clockwise) represent the *NPV*, *PPV* (or precision), Sensitivity and Specificity. The blue square provides the overall Correct Classification Rate, *C*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

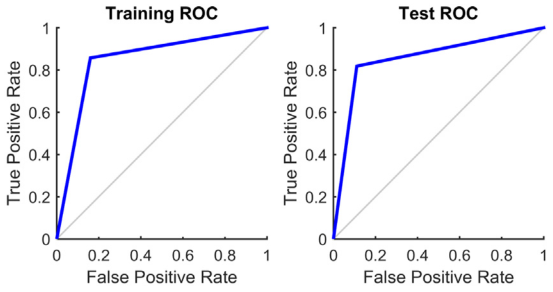

Fig. 3. ROC curves for DT classification accuracy on the training dataset (left) and on the test samples (right).

- **Sensitivity** (or recall, or *TP* rate), Sn, measures the proportion of positives that are correctly identified as such: $Sn = TP/(TP + FN)$
- **Specificity** (or *TN* rate), *Sp*, measures the proportion of negatives that are correctly identified as such: $Sp = TN/(TN + FP)$
- **Area under the ROC curve** [15], *AUC*. ROC curve depicts TP rate versus FP rate at various discrimination thresholds and is commonly used in medical statistics. On the unit ROC space, a perfect prediction would yield an *AUC* of 1.0. A random coin flipping would result in points along the diagonal and the corresponding *AUC* of 0.5. *AUC* is also known as c-statistic.

3. Results

3.1. Decision Tree (DT) model

The DT model in Fig. 1 was developed after considering 600 DTs built on different subsets of the data by permuting the test and model datasets with each other. The DT was able to correctly predict incidence of ABMR in $C = 85\%$ cases on both training and test datasets (Fig. 2). When evaluated on the test cohort, the DT identified ABMR + ve patients with 81.8% sensitivity and ABMR-ve cases with 88.90% specificity (Fig. 2). Fig. 3 shows the classifier ROC curves with $AUC_{train} = 0.849$ on training samples and $AUC_{\mathrm{test}} = 0.854$ for the DT predictions on tests (Fig. 3).

Out of 14 possible predictors, the DT identified the following 6 variables as key to ABMR prediction (Fig. 1): the highest IgG MFI level, total IgG4 MFI level, number of HLA mismatches, total IgG2 MFI level, the total IgG1 MFI level and cytometery cross-match. This was consistent with the results obtained using classical statistical methods on the same dataset [23]. None of the remaining 8 variables were used by the DT to predict ABMR.

Additionally, the node splits in the DT model (Fig. 1) provide an indication as to what specific *levels* of the HLA DSA antibodies were statistically associated with each of the ABMR + ve/ABMR-ve classes. For instance, the DT identified that all patients with the highest IgG levels below MFI 1062 belonged to the ABMR-ve group (no rejection), while those with the highest IgG level ≥1062 and the IgG4 MFI level ≥80 had a high (85%) likelihood of early transplant rejection. Similarly, 85% of patients with 4 or 5 HLA mismatches, the highest IgG level ≥1062, and IgG4 MFI level <80 belonged to the ABMR + ve group.

As expected, considerable volatility in performance and structure could be observed among the 600 DTs. However, a persistent pattern was noticed: 14 out of 600 DTs used the same 6 variables for classification as the model in Fig. 1. Further comparison of the performances of these 14 instances and the remaining 586 DTs was carried out. Fig. 4 shows that despite the large variance ($\sigma = 0.013$) in performance of DTs, those 14 DTs based on the 6 variables identified in our DT model have a significantly higher predictive power ($p < 0.002$).

3.2. Random Forest (RF) model

A RF of 600 trees achieved $C = 91.7\%$ during the training phase and correctly classified 85% of test cases, which is analogous to the DT model performance on the same test cohort presented in 3.1 (Table 1). When further evaluated on the test cohort, RF performed with 92.3% sensitivity, 71.4% specificity (Fig. 5) and the AUC_{test} of 0.819 (Fig. 6).

Experiment with RF was repeated 10 times in order to determine whether the consistency improved compared to the DT model. The results showed significantly reduced variance ($\sigma = 0.002$), and overall consistently high performance (Fig. 7).

Table 1
Predictive performance of the two ML models.

Performance measures as defined in Section 2.4	DT		RF	
	training	test	training	test
Correct classification rate, *C* (%)	85.0	85.0	91.7	85.0
Sensitivity, *Sn* (%)	85.7	81.8	93.9	92.3
Specificity, *Sp* (%)	84.0	88.9	88.9	71.4
Positive Predictive Value, *PPV* (%)	88.2	90.0	91.2	85.7
Negative Predictive Value, *NPV* (%)	80.8	80.0	92.3	83.3
Area under the ROC curve, *AUC*	0.849	0.854	0.914	0.819

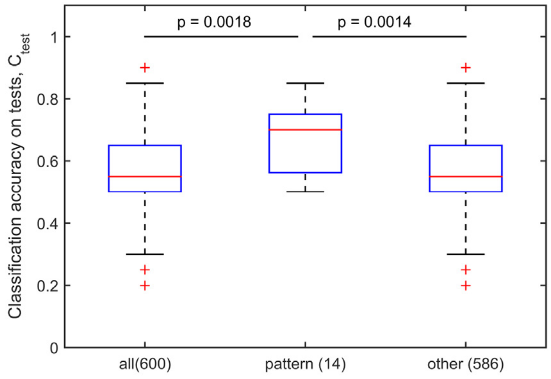

Fig. 4. Wilcoxon rank sum test [32] for median *C* based on 600 DTs and on the subset of DTs with repeating pattern. The DTs are of identical training parameters and design and only vary in the split between training and testing subsets.

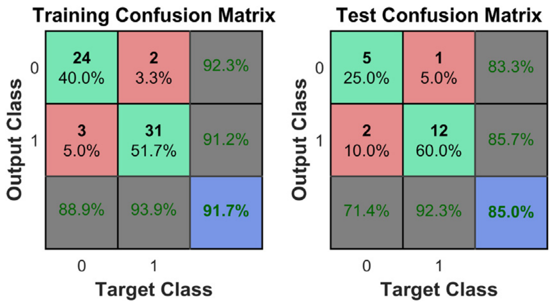

Fig. 5. Confusion matrix for the training dataset (left) and the test samples (right) for RF model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

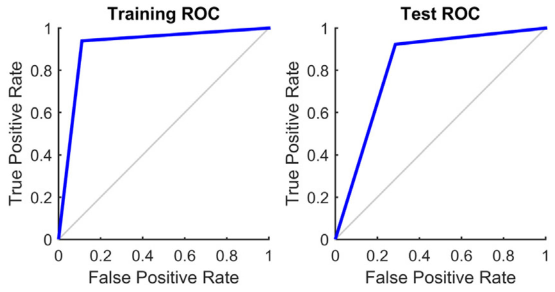

Fig. 6. ROC curves for RF classification accuracy on the training (left) and test (right) samples.

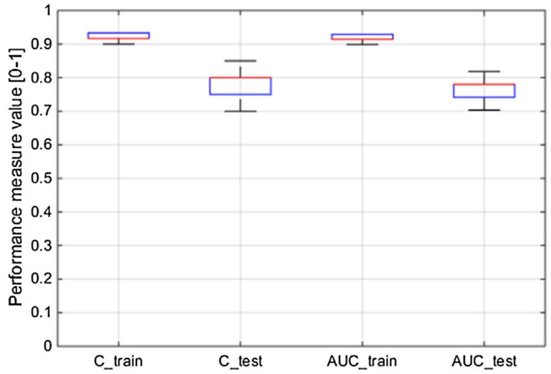

Fig. 7. Distributions of performance measures C_{train}, C_{test}, AUC_{train}, AUC_{test} for 10 RFs.

The variable importance scores were computed in order to identify the key important parameters used by the RF classifier. As

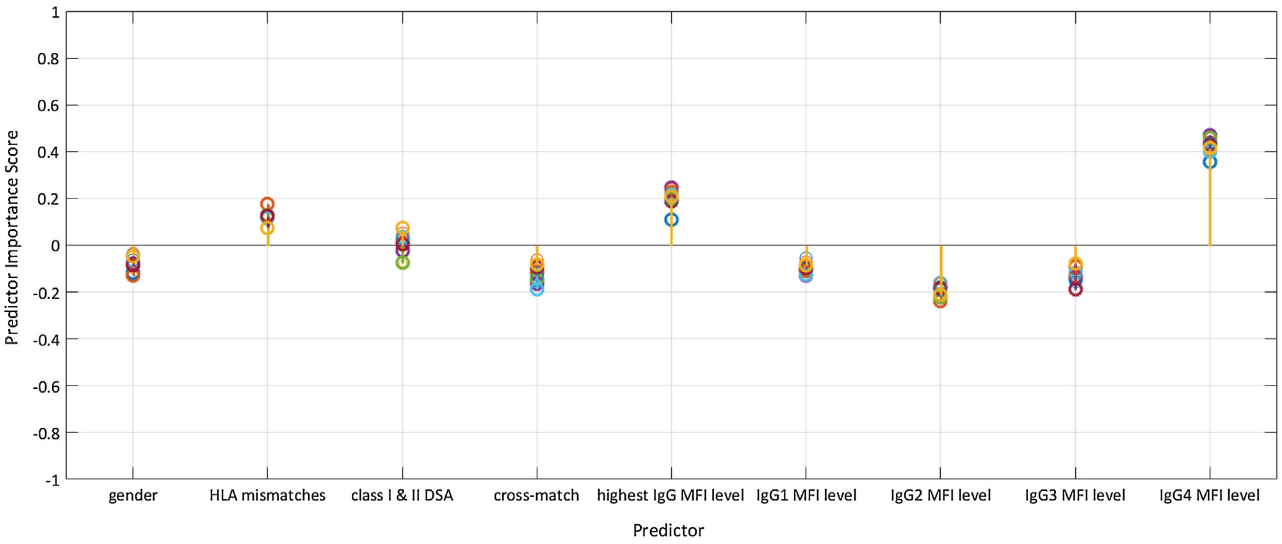

Fig. 8. Variable importance scores evaluated by a permutation test across 10 RFs.

shown in Fig. 8, IgG4 level is the single most important factor, followed by the highest MFI IgG level, and the number of HLA mismatches. This result further confirms our hypothesis that IgG4 is a key risk factor in kidney rejection in early post-transplant period [23].

4. Discussion

The test classification accuracy of 85% achieved by the models demonstrates that the ML approach can be effectively applied to predictive modelling in renal transplantation despite the small number of observations and heterogeneous input parameters. Based on only 80 cases, our DT model achieved a similarly high level of performance for acute ABMR as the model of Krikov et al. [9] for kidney allograft survival, which was built on a national database of 92,844 patient records. The proposed models outperform in their accuracy (*AUC* of 0.819 for RF and 0.854 for DT) some of the highest-performing models in the area of kidney transplantation discussed in Section 1 [13,14,16,17].

Our DT model, chosen from the group of DT models with the highest prediction power, was able to successfully determine the optimal set of parameters associated with early rejection. The 6 key predictors identified by the DT are confirmed by previously developed logistic regression likelihood multivariate model [23], which is a tool of choice in medical statistics for binary classification [24,33,34]. The superiority of the DT model is that it was also able to determine the level of antibodies associated with ABMR, which conventional statistical methods were unable to provide. It is important to note that it has been intuitively known by transplant doctors that harmful highest IgG antibody levels were at around 1000 MFI [35–37] which our model confirmed to be at 1062 MFI (Fig. 1). Additionally, the harmful levels of IgG4 were identified to be at 80 MFI (Fig. 1) addressing the clinical aim as set in the Introduction.

The RF model provides an extension to the DT model with the purpose of improving the robustness of the classification tool. A RF is so-called black-box model and less interpretable than DT, but allows for better consistency of results and robustness of predictions. Our DT and RF are equally well-equipped to handle partially missing data and managed to classify correctly the 3 cases with incomplete data.

Tree-based models can be implemented in the electronic decision support system by means of standard computational resources. When used for clinical decision support, our models can provide a simulation tool to explore various clinical scenarios and identify patients at risk of ABMR prior to the operation, and thus leaving more time to make essential adjustment to treatment.

It is important to state that outcomes from this single-centre study may not generalise on a larger population in and outside of the United Kingdom. They may be affected by institutional bias, and therefore a further work comparing the results on extended datasets from other centres would be beneficial. Despite this limitation, the achieved outcomes remain significant to the area of kidney transplantation.

To our best knowledge, this study is the only work aimed at developing ML models for prediction of acute ABMR based on specific MFI levels of IgG subclasses. Predictions made by our DT and RF models are patient-specific yielding an accurate and robust tool for ABMR risk stratification preceding transplantation.

5. Conclusions

1) The incidence of acute antibody mediated rejection was successfully modelled from 14 clinical baseline characteristics, including pre-transplant DSA levels, using a DT. Despite a small dataset of 80 samples, this graphical and easily-interpretable DT model revealed that the highest MFI DSA levels, the total IgG4 subclass MFI and the number of HLA mismatches are the highest discriminating factors between ABMR + ve and ABMR-ve patients.
2) This research identified that patients with (a) the highest MFI DSA levels below 1062 belong to the ABMR-ve group (rejection rate of 0%), while those with (b) the highest MFI DSA levels ≥1062 AND the total IgG4 subclass MFI level ≥80 have a high likelihood of early ABMR (rejection rate of 85%). Similarly, patients in (b) with 4 or 5 HLA mismatches are likely to develop ABMR (rejection rate of 85%).
3) We developed a RF model of 600 bagged DTs, which provided a robust classification with 85% accurate predictions in determining the acute ABMR of kidney transplants for each individual patient characterised by the presence of DSAs with high MFI levels, high IgG4 subclass MFI levels and a large number of HLA mismatches.

Acknowledgment

This work was supported by the EPSRC UK Grant EP/K02504X/1.

Details on the data used for analysis are available from the University of Warwick institutional repository at http://wrap.warwick.ac.uk/85691.

References

[1] S. Russell, P. Norvig, Artificial Intelligence: A Modern Appproach, 3rd ed., Pearson, Harlow, 2013.
[2] E. Alpaydın, Introduction to Machine Learning, 3rd ed., The MIT Press, Cambridge, MA, 2014.
[3] T.M. Mitchell, The discipline of machine learning, Mach. Learn. 17 (2006) 1–7, http://dx.doi.org/10.1080/026404199365326.
[4] V. Lakshmanan, E. Gilleland, A. McGovern, M. Tingley (Eds.), Machine Learning and Data Mining Approaches to Climate Science, Springer International Publishing, Cham, 2015, 10.1007/978-3-319-17220-0.
[5] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, J. Lozano, Machine learning: an indispensable tool in bioinformatics, in: R. Matthiesen (Ed.), Bioinforma. Methods Clin. Res., Humana Press, 2010, pp. 25–48, http://dx.doi.org/10.1007/978-1-60327-194-3_2.
[6] D.L. Hudson, M.E. Cohen, Neural Networks and Artificial Intelligence for Biomedical Engineering, IEEE, New York, 2000.
[7] Artificial Intelligence in Medicine, in: N. Peek, M. June, R. Goebel (Eds.), Springer, Berlin Heidelberg, 2013, http://dx.doi.org/10.1007/978-3-642-38326-7.
[8] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, Decision trees: an overview and their use in medicine, J. Med. Syst. 26 (2002) 445–463, http://dx.doi.org/10.1023/A:1016409317640.
[9] G. Forman, I. Cohen, Learning from little: comparison of classifiers given little training, Proc. PKDD 19 (2004) 161–172, http://dx.doi.org/10.1007/978-3-540-30116-5_17.
[10] R. Lanouette, J. Thibault, J.L. Valade, Process modeling with neural networks using small experimental datasets, Comput. Chem. Eng. 23 (1999) 1167–1176, http://dx.doi.org/10.1016/S0098-1354(99)00282-3.
[11] A. Azar, S. El-Metwally, Decision tree classifiers for automated medical diagnosis, Neural Comput. Appl. 23 (2013) 2387–2403, http://dx.doi.org/10.1007/s00521-012-1196-7.
[12] T.S. Purnell, P. Auguste, D.C. Crews, J. Lamprea-Montealegre, T. Olufade, R. Greer, P. Ephraim, J. Sheu, D. Kostecki, N.R. Powe, H. Rabb, B. Jaar, L.E. Boulware, Comparison of life participation activities among adults treated by hemodialysis, peritoneal dialysis, and kidney transplantation: a systematic review, Am. J. Kidney Dis. 62 (2013) 953–973, http://dx.doi.org/10.1053/j.ajkd.2013.03.022.
[13] R. Greco, T. Papalia, D. Lofaro, S. Maestripieri, D. Mancuso, R. Bonofiglio, Decisional trees in renal transplant follow-up, Transplant. Proc. 42 (2010) 1134–1136, http://dx.doi.org/10.1016/j.transproceed.2010.03.061.
[14] S. Krikov, A. Khan, B.C. Baird, L.L. Barenbaum, A. Leviatov, J.K. Koford, A.S. Goldfarb-Rumyantzev, Predicting kidney transplant survival using tree-based modeling, ASAIO J. 53 (2007) 592–600, http://dx.doi.org/10.1097/MAT.0b013e318145b9f7.
[15] R.H. Riffenburgh, Statistics in Medicine, 3rd ed., Academic Press, Burlington, 2012.
[16] A. Decruyenaere, P. Decruyenaere, P. Peeters, F. Vermassen, T. Dhaene, I. Couckuyt, Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods, BMC

Med. Inform. Decis. Mak. 15 (2015), http://dx.doi.org/10.1186/s12911-015-0206-y.

[17] D. Lofaro, S. Maestripieri, R. Greco, T. Papalia, D. Mancuso, D. Conforti, R. Bonofiglio, Prediction of chronic allograft nephropathy using classification trees, Transplant. Proc. 42 (2010) 1130–1133, http://dx.doi.org/10.1016/j.transproceed.2010.03.062.

[18] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, N. Khovanova, Machine learning for predictive modelling based on small data in biomedical engineering, IFAC-Papers Online 48 (2015) 469–474, http://dx.doi.org/10.1016/j.ifacol.2015.10.185.

[19] T. Shaikhina, N. Khovanova, Handling limited datasets with neural networks in medical applications: a small data approach, Artif. Intell. Med. 75 (2017) 51–63, http://dx.doi.org/10.1016/j.artmed.2016.12.003.

[20] N.A. Khovanova, K.K. Mallick, T. Shaikhina, Neural networks for analysis of trabecular bone in osteoarthritis, Bioinspired Biomim. Nanobiomater. 4 (2015) 90–100, http://dx.doi.org/10.1680/bbn.14.00006.

[21] R. Higgins, M. Hathaway, D. Lowe, F. Lam, H. Kashi, L.C. Tan, C. Imray, S. Fletcher, D. Zehnder, K. Chen, N. Krishnan, R. Hamer, D. Briggs, Blood levels of donor-specific human leukocyte antigen antibodies after renal transplantation: resolution of rejection in the presence of circulating donor-specific antibody, Transplantation 84 (2007) 876–884, http://dx.doi.org/10.1097/01.tp.0000284729.39137.6e.

[22] J.M. Gloor, M.D. Stegall, ABO incompatible kidney transplantation, Curr. Opin. Nephrol. Hypertens. 16 (2007) 529–534, http://dx.doi.org/10.1097/MNH.0b013e3282f02218.

[23] N. Khovanova, S. Daga, T. Shaikhina, N. Krishnan, J. Jones, D. Zehnder, D. Mitchell, R. Higgins, D. Briggs, D. Lowe, Subclass analysis of donor HLA-specific IgG in antibody-incompatible renal transplantation reveals a significant association of IgG4 with rejection and graft failure, Transpl. Int. 28 (2015) 1405–1415, http://dx.doi.org/10.1111/tri.12648.

[24] D. Lowe, R. Higgins, D. Zehnder, D.C. Briggs, Significant IgG subclass heterogeneity in HLA-specific antibodies: implications for pathogenicity, prognosis, and the rejection response, Hum. Immunol. 74 (2013) 666–672, http://dx.doi.org/10.1016/j.humimm.2013.01.008.

[25] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[26] D. Coppersmith, S.J. Hong, J.R. Hosking, Partitioning nominal attributes in decision trees, J. Data Min. Knowl. Discov. 3 (1999) 197–217, http://dx.doi.org/10.1023/A:1009869804967.

[27] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, Refrig. Air Cond. 57 (1993) 436, http://dx.doi.org/10.1111/1467-9639.00050.

[28] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, http://dx.doi.org/10.1007/BF00058655.

[29] L. Breiman, Random forest, Mach. Learn. 45 (1999) 1–35, http://dx.doi.org/10.1023/A:1010933404324.

[30] R.J. Marshall, The use of classification and regression trees in clinical epidemiology, J. Clin. Epidemiol. 54 (2001) 603–609, http://dx.doi.org/10.1016/S0895-4356(00)00344-9.

[31] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (1999) 169–198, http://dx.doi.org/10.1613/jair.614.

[32] C.J. Wild, G.A. Seber, The wilcoxon rank-Sum test, in: Chance Encount. A First Course Data Anal. Inference, John Wiley & Sons, New York, 2000, pp. 611.

[33] M. Behera, E. Fowler, T. Owonikoko, W. Land, W. Mayfield, Z. Chen, F. Khuri, S. Ramalingam, J. Heine, Statistical learning methods as a preprocessing step for survival analysis: evaluation of concept using lung cancer data, Biomed. Eng. Online 10 (2011) 10–97, http://dx.doi.org/10.1186/1475-925X-10-97.

[34] W. Bouwmeester, N.P.A. Zuithoff, S. Mallett, M.I. Geerlings, Y. Vergouwe, E.W. Steyerberg, D.G. Altman, K.G.M. Moons, Reporting and methods in clinical prediction research: a systematic review, PLoS Med. 9 (2012) e1001221, http://dx.doi.org/10.1371/journal.pmed.1001221.

[35] G. Dieplinger, V. Ditt, W. Arns, A. Huppertz, T. Kisner, M. Hellmich, U. Bauerfeind, D.L. Stippel, Impact of de novo donor-specific HLA antibodies detected by Luminex solid-phase assay after transplantation in a group of 88 consecutive living-donor renal transplantations, Transpl. Int. 27 (2014) 60–68, http://dx.doi.org/10.1111/tri.12207.

[36] E.F. Reed, P. Rao, Z. Zhang, H. Gebel, R.A. Bray, I. Guleria, J. Lunz, T. Mohanakumar, P. Nickerson, A.R. Tambur, A. Zeevi, P.S. Heeger, D. Gjertson, Comprehensive assessment and standardization of solid phase multiplex-bead arrays for the detection of antibodies to HLA, Am. J. Transplant. 13 (2013) 1859–1870, http://dx.doi.org/10.1111/ajt.12287.

[37] M.L. Arnold, I.S. Ntokou, I.I.N. Doxiadis, B.M. Spriewald, J.N. Boletis, A.G. Iniotaki, Donor-specific HLA antibodies: evaluating the risk for graft loss in renal transplant recipients with isotype switch from complement fixing IgG1/IgG3 to noncomplement fixing IgG2/IgG4 anti-HLA alloantibodies, Transpl. Int. 27 (2014) 253–261, http://dx.doi.org/10.1111/tri.12206.