Predicting the Progression of IgA Nephropathy using Machine Learning Methods

Junhyug Noh[1], Dharani Punithan[2], Hajeong Lee[3]
jhroh86@gmail.com, punithan.dharani@gmail.com, mdhjlee@gmail.com
[1]Computer Science and Engineering, College of Engineering, Seoul National University, South Korea
[2]Institute of Computer Technology, Seoul National University, South Korea
[3]Internal Medicine, College of Medicine, Seoul National University, South Korea

ABSTRACT

We predict the progression of Immunoglobulin A Nephropathy using three of the most widely used supervised classification machine learning algorithms : Classification and Regression Trees, Logistic Regression (in two different forms), and Feed-Forward Neural Networks. The problem is treated as a classification problem, of predicting progression to end-stage renal disease in the ten years following initial diagnosis. All four methods yielded good classifiers, with AUC performance between 0.85 (decision tree) and 0.89 (neural network). The results were generally in-line with expectations, with poor kidney performance on presentation, and evident macroscopic and microscopic damage, all associated with poorer prognosis. However, the association between normal systolic blood pressure status and poor prognosis, for some patients under specific conditions, was entirely unanticipated, and warrants further investigation.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Applications and Expert Systems, Learning

General Terms

Theory

Keywords

Immunoglobulin A Nephropathy (IgAN), End-Stage Renal Disease (ESRD), Classification and Regression Trees (CART), Logistic Regression, Neural Networks, Receiver Operating Characteristic (ROC), Area Under Curve (AUC), Missing Completely At Random (MCAR)

1. INTRODUCTION

Immunoglobulin A Nephropathy (IgAN) is the most common glomerulonephritis worldwide and the key cause of End-Stage Renal Disease (ESRD). Its clinical course is highly variable, with a 10-year renal survival rate in the range 70–80% [6]. Because patients are usually diagnosed at fairly young age, 20-30% of IgAN patients experience ESRD during their life.

Renal (kidney) function is measured by glomerular filtration rate (GFR), the volume of blood filtered from the renal glomerular capillaries per unit time. The severity of IgAN can be classified into five stages. The end of the progression is ESRD, a severe illness requiring either regular dialysis or kidney transplantation, and with poor life expectancy. The fifth stage, although it retains some kidney function, is nevertheless a severe illness, and generally progresses to ESRD. Another important stage in defining the progression is the CR2 stage, at which the serum creatinine level (a measure of the elimination effectiveness of the kidneys) has doubled.

1.1 Problem Definition

The insidious disease course and its high variability make it difficult for physicians to predict renal outcome at the time of diagnosis. This has both medical and social consequences. It is difficult for physicians to determine how aggressively to treat each individual case (as is common in medicine, aggressive treatments have more severe consequences). And it is difficult for patients to make long-term plans because of this uncertainty. Previous studies have determined some factors associated with poor renal prognosis, including initial renal function, blood pressure, and the amount of proteinuria [1]. However, they have not been able to demonstrate reliable outcome prediction. Our aim in this work is to provide more robust predictors using machine learning techniques. Specifically, we assume that we have the initial presentation and biopsy data for a patient, and aim to predict the progression to ESRD within a specific period (10 years). The outcome of IgAN progression is dichotomous (ESRD or not), and hence we have a binary prediction (classification) problem.

1.2 Motivation

By far the major challenge in the field is the identification, at an early stage, of the patients at highest risk of progression to ESRD. The tools and methods for predicting renal prognosis are limited. There is some evidence that genetic and social factors influence IgAN progression, hence it is specifically of interest to investigate progression in the relatively homogeneous Korean population.

1.3 Importance

The prevalence of glomerular diseases varies based on geographic area, race, age and other factors. Race/ethnicity

.

BICT 2014, December 01-03, Boston, United States

DOI 10.4108/icst.bict.2014.257893

is one of the risk factors for IgAN. Studies show that IgAN is particularly prevalent and its course more severe in patients of Asian ancestry. Hence investigation of Asian (Korean) populations can be especially effective in identifying risk factors for progression.

1.4 Contribution

Though IgAN has been widely studied in Asian countries including South Korea [13, 5], Singapore [12], China [16] and Japan [11], their research methodologies were based on traditional descriptive and exploratory statistical analysis. Hence, our proposed use of machine learning algorithms provides a useful complement, potentially useful for clinical investigations and medical and patient decision-making.

1.5 Outline of the paper

In section 2, we describe the background of IgA Nephropathy. Section 3 details our methodologies. The results are presented in section 4 and further analysed in section 5. We summarise our results in section 6.

2. BACKGROUND

2.1 Immunoglobulin A Nephropathy

Immunoglogulin A nephropathy (IgAN), first described by Berger and Hinglais [2], is the most common immune-complex-mediated glomerulonephritis (GN) – inflammation of the glomeruli of the kidney – worldwide [15, 9]. IgAN (or Berger's disease) is a chronic kidney disease in which an antibody, Immunoglobulin A (IgA), forms granular deposits in the glomeruli – blood vessels in the kidney. It is unknown why IgA is trapped in the glomuleri, but its presence causes inflammation. These mesangial IgA deposits affect the ability of the kidneys to perform their normal function of filtering waste, excess water and electrolytes from the blood.

A few IgAN patients experience complete remission, but many eventually progress to ESRD, requiring hemodialysis (for acute kidney failure) or a kidney transplant (for chronic kidney failure) for their survival. IgAN can progress slowly, over many years, through the five stages from worsening renal dysfunction to ESRD. The length of this progression varies from patient to patient, but can be from 10 to 20 years. Furthermore even transplantation is not a complete cure – in many cases, substantial mesangial IgA deposits have recurred in kidneys transplanted into patients who had developed end-stage renal disease due to IgAN [3].

3. METHODS

3.1 Dataset

The dataset has been built up by the Division of Nephrology, Seoul National University Hospital (SNUH) – one of the best-reputed hospitals in South Korea. It details 1623 Korean biopsy-confirmed IgAN patients who were identified between the years 1979 and 2014. The dataset was last updated on May 29th, 2014. Most patients' biopsy tests were analysed by the same laboratory; in the exceptional cases, appropriate corrections were made to retain consistency.

The dataset consists of data about the patients' initial presentation and biopsy, and their GFR information from subsequent follow-up sessions. The dataset includes 68 attributes, grouped into four categories : demographic, laboratory, clinical and histological. The input attributes in our binary classification predictive modelling, come from the initial presentation data; the GFR values measured during the follow-up sessions are not used for the modelling. However, the value of the target attribute, ESRD, also depends on the follow-up GFR data, and we plan to investigate its use for updated predictions in subsequent work.

Table 1: Attributes Used for Modelling

Name	Description	Category
AGE	age of patient	Demographic
SEX	sex of patient	Demographic
GLOM	no. of glomeruli	Histological
CRES%	% of crescent	Histological
GS%	% of global glomerulosclerosis	Histological
SS%	% of segmental glomerulosclerosis	Histological
IF	renal tubule fibrosis	Histological
TA	renal tubule atrophy	Histological
II	renal tubule infiltrate (inflammatory)	Histological
SBP	systolic blood pressure	Clinical
BMI	body mass index	Clinical
SMHX	smoking history	Clinical
HB	hemoglobin	Laboratory
ALB	serum albumin	Laboratory
CHOL	cholesterol	Laboratory
GFR	glomerular filtration rate	Laboratory
PU	24 hours proteinuria	Laboratory

3.2 Attributes Used for Modelling

Among the 68 attributes, we relied on the domain knowledge of the nephrologists to choose 17 independent attributes (refer Table 1) to build machine learning models. They are AGE, SEX, GLOM, CRES%, GS%, SS%, IF, TA, II, SBP, BMI, SMHX, HB, ALB, CHOL, GFR and PU.

GFR is computed with the standard Modification of Diet in Renal Disease (MDRD) equation [14], adapted for Koreans:

$$\begin{aligned} \text{GFR} = 175 \quad & * \quad \text{AGE}^{-0.203} * \text{CR}^{-1.154} \\ & * \quad 1.0 \text{ (if male)} \\ & * \quad 0.742 \text{ (if female)} \end{aligned} \tag{1}$$

where GFR is measured in ml/min/$1.73m^2$. The normal GFR value is above 90ml/min/$1.73m^2$ with no proteinuria. If the GFR is very low ($<$ 15ml/min/$1.73m^2$), the patient is more likely to progress to ESRD. In the equation 1, CR is the creatine level.

3.3 Target Attribute

The target attribute, ESRD, is a binary variable taking values 0 (negative class indicating the absence of ESRD – non-ESRD) and 1 (positive, the presence of ESRD). From the original dataset, we remove any records missing ESRD status (labels), leaving 1606 medical records.

3.4 Setting Prediction Period

In medical practice, 5- and 10-year survival rates are generally used for estimating the prognosis of a disease. 5-year survival is more useful in aggressive diseases with a shorter

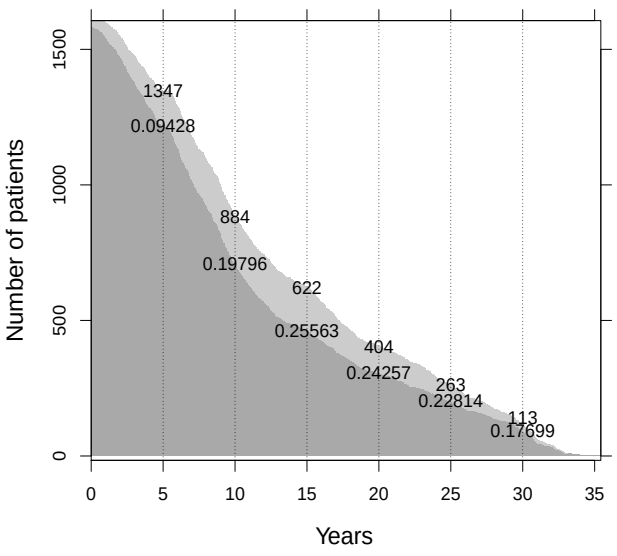

Figure 1: Total Cases and Positive Ratio vs Years since Biopsy (Dark Grey: Negative Cases; Light Grey: Positive Cases)

life expectancy following diagnosis, whereas 10-year is more practical in less invasive diseases with a long life expectancy. Following this model, ESRD progression is also generally expressed by 5- and 10-year renal survival. Using standardised periods is important for understanding disease severity and comparing treatment effectiveness.

Exploratory data analysis was used to identify the most suitable target period for prediction. If we chose too short a period, positive cases would be too rare, and predictions would be of limited value. On the other hand, because the data is still being accumulated, a long period would have too few overall cases. Figure 1 shows this graphically. Against the prediction period, we plotted the number of patients whose records cover that period (a patient's records are considered to cover a period of N years if the interval between the patient's initial biopsy date and the last database update is at least N years). We divided the patients into those who had not reached ESRD after N years (dark grey) and those who had (light grey). The whole numbers on the plot are the total number of patients in the sample, while the real numbers are the proportion of positive cases. From the figure, we concluded that 5 years was too short to be useful (too few positive cases), while 15 was too long (too few overall cases). Thus data properties confirmed our choice of 10 years.[1] This left 884 cases for modelling.

3.5 Missing Values

Initially, the medical records were maintained manually, and there are missing values in those older records. Thus the missing values mainly depend on the patient's first-visit date – the records were computerised in 1999, after which missing values are rare. However, the first-visit date is not used for modelling. Thus, the nature of the missing data relative to our learning task is MCAR (Missing Completely At Random) and we can use Complete Case Analysis without incurring bias. Taking this into account, we finally arrived at a cohort of 655 patients' data for our modelling.

3.6 Data Partitions

The 655 records divide into 510 negative (ESRD = 0) and 145 positive (ESRD = 1). We split the 655 records into two disjoint datasets: a training dataset (80% (= 524 records)) and a test dataset (20% (= 131 records)) by stratified random sampling. We also formed a third "mixed" dataset of records covering less than 10 years. The mixed dataset includes 448 patients: 402 cases without ESRD and 46 with. ESRD is irreversible, so cases with ESRD (= 1) before 10 years also be positive after 10 years. For the 46 positive cases in the mixed dataset, we can validate the prediction of the models (true positives). But cases without ESRD (= 0) now may progress to ESRD (= 1) within 10 years. The 402 negative cases can be predicted by the model (illustrating usefulness), but cannot validate it.

We built the binary classifiers using the training dataset. We applied the prediction models to the held-out test dataset to analyse the performance of the classifiers. Finally, we both validated (for ESRD = 1 cases) and predicted (for ESRD = 0 cases) the ESRD stage after 10 years for the observations in the mixed dataset, using the classifiers.

3.7 Building Models and Parameter Tuning

We built classifiers using R's statistical modelling tools and libraries [19] : RPART [18] (classification tree), GLM [7] (logistic regression) and NNET [17] (neural network). We used cross validation to avoid overfitting and tune model parameters. The parameters and values used for learning are described in section 4.

For each method, we first defined the sets of learning parameter values. For each set of parameters, we performed 5-fold cross validation, splitting the training set (80% = 524 records) into 5 cross validation folds [10] by random sampling (the same 5 folds were used throughout). One among the 5 folds was held out, and the model was fitted to the remainder. It was used to predict the held-out fold. This was repeated for each fold. We then computed the average prediction performance across folds.

We used Receiver Operating Characteristic (ROC) analysis to choose the best parameter set, computing the average Area Under the ROC Curve (AUC) across the 5 cross validation sets. AUC is an effective measure to find the best candidate model: larger AUC is better. Then we re-trained using these parameters on the full training set.

3.8 Performance Evaluation of Models

We assessed the performance of the trained models by applying them to the held-out test dataset to predict ESRD. We again used the ROC curve analysis to evaluate the performance (discriminatory ability) of the different learning models. ROC is a plot of Sensitivity (true positive rate) against (1− Specificity) (false positive rate). We used the ROC plot to detect the best cut-off point maximising both sensitivity and specificity. The best cut-off point is that on

[1]We assume that the ESRD value is 0 until the ESRD date (the date on which ESRD is confirmed as 1). Specifically, this means that if the difference between the first biopsy date and the ESRD date is greater than 10 years, we consider those cases as non-ESRD (ESRD = 0).

the ROC curve closest to the coordinate (0, 1) – i.e. nearest to the upper left corner of the ROC plot. We compared the performance of the models by estimating AUC.

3.9 Validation and Prediction of Models

46 cases in the mixed dataset have known disease status (ESRD = 1). They can assess the model's ability to correctly identify positives cases (ESRD = 1). Our priority of analysis, on the mixed dataset, is on sensitivity.

4. RESULTS

We discuss the values chosen for model parameters, training results, performance evaluation of the classifiers on the test set, and validation and prediction on the mixed dataset. The classifier performance is illustrated visually with ROC plots and analysed with statistical measures: sensitivity, specificity, accuracy and AUC.

4.1 Classification & Regression Trees (CART)

We created the classification trees using the RPART [18, 19] (Recursive PARTitioning) routines of R, which implement CART [4]. RPART explores the attributes and threshold values for splitting, choosing the decision rules which minimise classification impurity using the Gini index.

4.1.1 Choosing the CART Parameters:

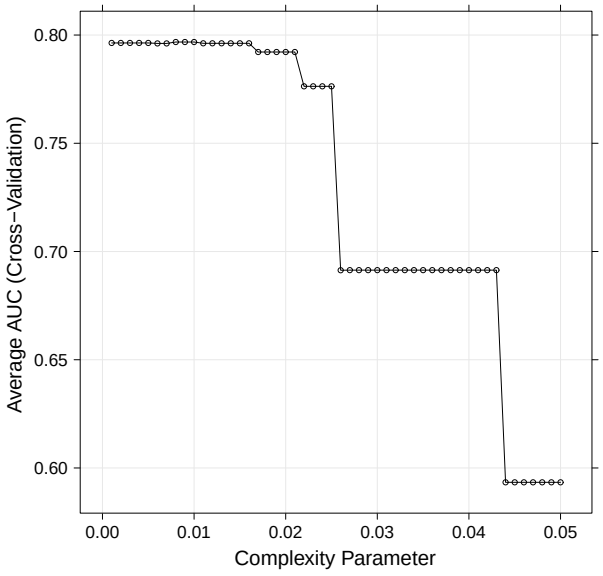

Figure 2: Average AUC Vs. Complexity Parameter with Max. Tree Depth = 30

RPART has two main control parameters: complexity parameter (**cp** – controls pruning of splits) and maximum tree depth (**maxdepth** – sets the maximum depth of the tree, with the root node as depth 0). We chose model parameters using cross-validation as described in sub-section 3.7.

The average AUC across the 5-fold cross validation sets for complexity paramaters (cp) in the range $[0.00, 0.05]$ is plotted in Figure 2. Figure 3 shows the average AUC against maximum tree depth from 2 to 30. The settings cp = 0.01 and maxdepth = 30 gave highest average AUC (= 0.797) (figures 2 and 3), so they were used in the rest of the analysis.

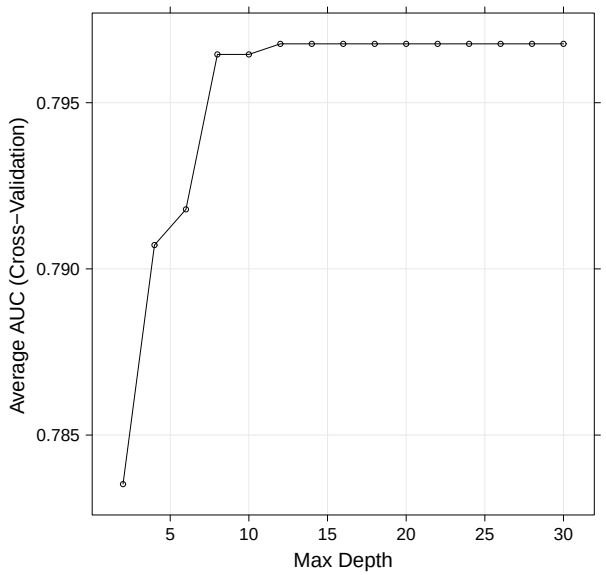

Figure 3: Average AUC Vs. Max. Tree Depth with Complexity Parameter = 0.01

4.1.2 Training Dataset Results:

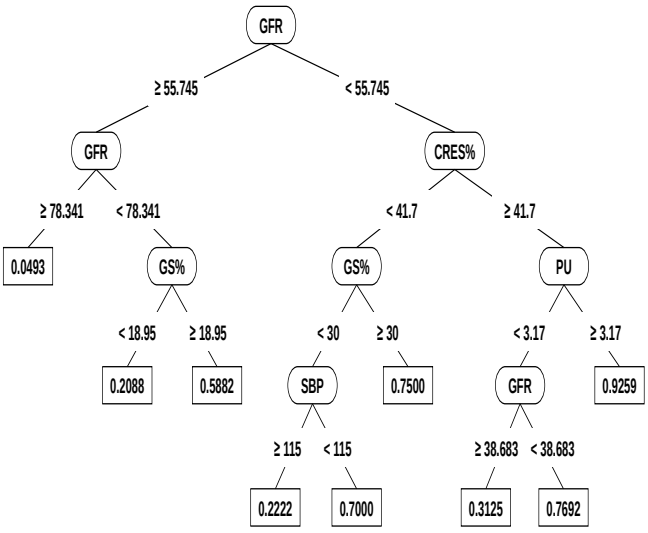

Figure 4: Classification Tree from Chosen Parameter Settings: Complexity Parameter= 0.01 **and Max. Tree Depth**= 30

Figure 4 shows the binary classification tree derived from the training set ($cp = 0.01$, $maxdepth = 30$). Internal node labels are attributes, edge labels are attribute threshold values for the split, and leaf (terminal) nodes show the sample relative frequency of ESRD = 1, given the decision rule. The AUC for this tree was 0.797.

4.1.3 Evaluation on Test Dataset:

In Figure 5, the ROC curve rises well above the diagonal, indicating good model performance. The best cut-off for predicting ESRD stage was determined as 0.2155. Prob-

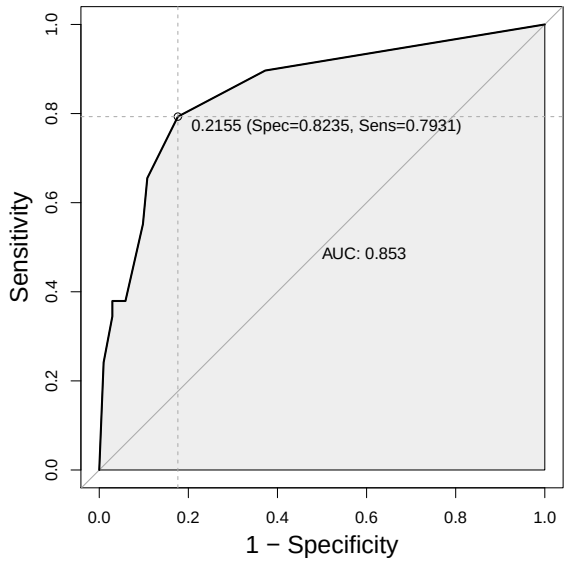

Figure 5: ROC plot for Decision Tree (CART)

abilities (leaf node values) below 0.2155 are classified negative (ESRD = 0), and the rest positive (ESRD = 1). At this cut-off, sensitivity was 0.7931 and specificity 0.8235 (1− specificity = 0.1765). The largest AUC (shaded in grey) had area 0.853. The model accuracy was 0.8167.

4.1.4 Validation and Prediction on Mixed Dataset:

We used the model to predict ESRD values for the mixed dataset. The decision tree validated 39 cases as true positive (ESRD = 1) with sensitivity = 0.8478. It predicted 94 among the 402 unknown cases to progress to ESRD = 1 within 10 years.

4.2 Logistic Regression

We used the GLM [7, 19] library of R to fit the logistic regression model. It is a generalised linear model (GLM) using a binomial distribution for the response with a logit link function. We used p-values < 0.05 to identify significant attributes.

4.2.1 Without Stepwise Variable Selection:

The significant attributes (p-values < 0.05) selected by GLM were GFR, CRES%, CHOL, HB, AGE, GS%, GLOM and SMHX (refer Table 2), with AUC = 0.831.

4.2.2 Stepwise Variable Selection:

We used stepwise variable selection to choose the most important variables. This adds or removes variables repeatedly, improving the model at each step. If there is no available improvement by adding or subtracting variables, the algorithm stops and returns the new model.

We used AIC (Akaike's information criterion) [19], often used as the model selection criteria for GLM, to fit the model. The procedure for deletion or inclusion is based on AIC, defined as (-2 maximised log-likelihood + 2 number of attributes). It stops when the AIC cannot be improved.

The stepwise logistic regression model selected the significant attributes (p-values < 0.05) as GFR, CRES%, CHOL, HB, GS%, AGE, SMHX, GLOM and SEX (refer Table 3). With this variable selection, the AUC was 0.831.

Table 2: Logistic Regression with All Variables

Name	Coefficient	p-value
Intercept	3.4183	0.0705
GFR	−0.0292	$5.96e-07$
CRES%	0.0236	0.0007
CHOL	0.0080	0.0143
HB	−0.1826	0.0186
AGE	−0.0284	0.0194
GS%	0.0282	0.0256
GLOM	−0.0160	0.0315
SMHX	0.3765	0.0469
SEX	0.5940	0.0735
SBP	−0.0079	0.2138
IF	0.5143	0.5138
TA	0.1267	0.6557
PU	−0.0289	0.6836
II	−0.0878	0.7115
SS%	−0.0087	0.7135
ALB	−0.0562	0.8403
BMI	−0.0082	0.8431

Table 3: Logistic Regression with Variable Selection

Name	Coefficient	p-value
Intercept	2.1967	0.0449
GFR	−0.0271	$7.75e-07$
CRES%	0.0245	$7.12e-05$
CHOL	0.0074	0.0105
HB	−0.1747	0.0140
GS%	0.0286	0.0167
AGE	−0.0274	0.0168
SMHX	0.4010	0.0302
GLOM	−0.0152	0.0358
SEX	0.5366	0.0941

4.2.3 Evaluation on Test Dataset:

For logistic regression, the best cut-off found was 0.1902 on the ROC, with AUC = 0.878, sensitivity = 0.8276 and specificity = 0.7745 (1− specificity = 0.2255). The classification accuracy of the model was 0.7862 (Figure 6).

Stepwise logistic regression model gave a cut-off of 0.1559, with the corresponding sensitivity = 0.8966 and specificity = 0.7255 (1− specificity = 0.2745) (refer Figure 7).The largest AUC was 0.876, and the model accuracy 0.7633.

4.2.4 Validation and Prediction on Mixed Dataset:

Logistic regression with all variables correctly identified 45 instances from the mixed dataset as positive cases, with sensitivity = 0.9783. It predicted that 154 cases would progress to ESRD = 1 within 10 years.

The stepwise logistic regression model predicted 171 cases to progress to ESRD = 1, and validated 45 cases as true positive cases (sensitivity = 0.9783).

4.3 Neural Networks

We used the NNET [17, 19] package in R, which builds feed-forward neural networks with a single hidden layer.

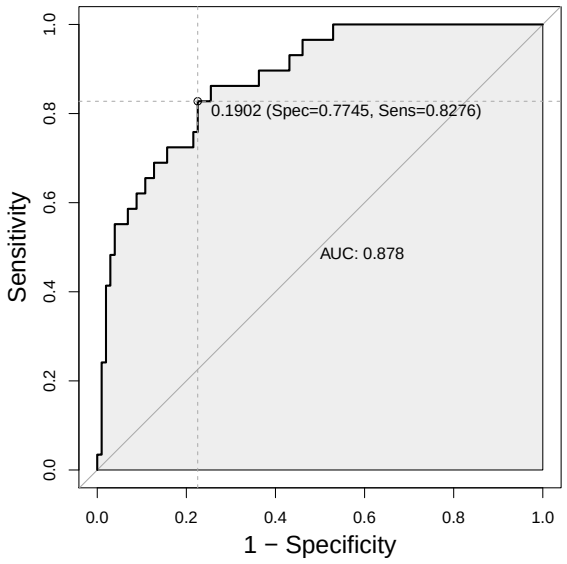

Figure 6: ROC : Logistic Regression with All Variables

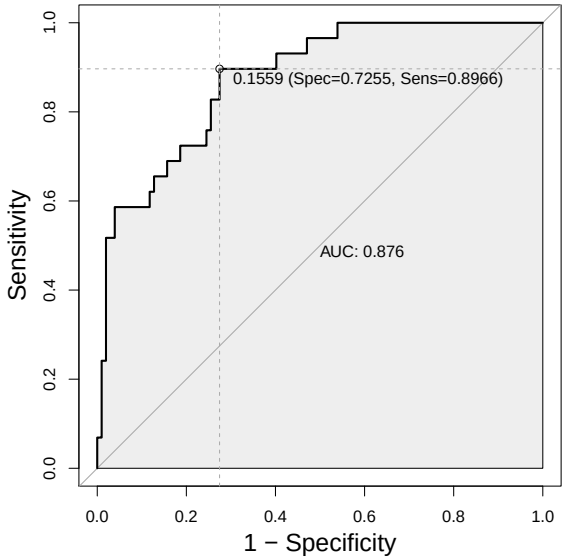

Figure 7: ROC : Logistic Regression with Variable Selection

4.3.1 *Choosing the Model Parameters:*

The **decay** parameter ensures that the model does not overtrain, and the **size** parameter specifies the number of nodes in the hidden layer. From Figure 8, we observe that the best model (AUC = 0.840) has 8 hidden layer nodes and a decay parameter of 0.26.

4.3.2 *Training Dataset Results:*

We used all 17 attributes in the input layer, and a hidden layer of 8 nodes. The relative importance of the 17 input variables are listed in Table 4. The relative importance was

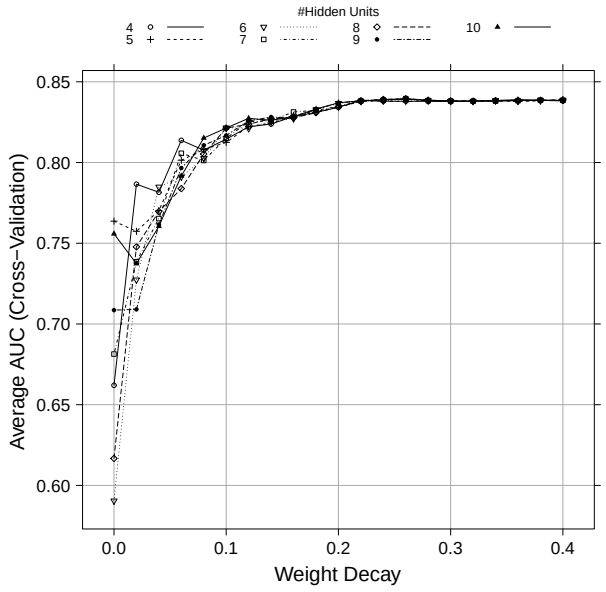

Figure 8: Average AUC Vs. Decay

computed with Garson's algorithm [8], which determines the overall influence of each predictor variable. The most important attributes were GFR, CRES%, HB and GS%.

4.3.3 *Evaluation on Test Dataset:*

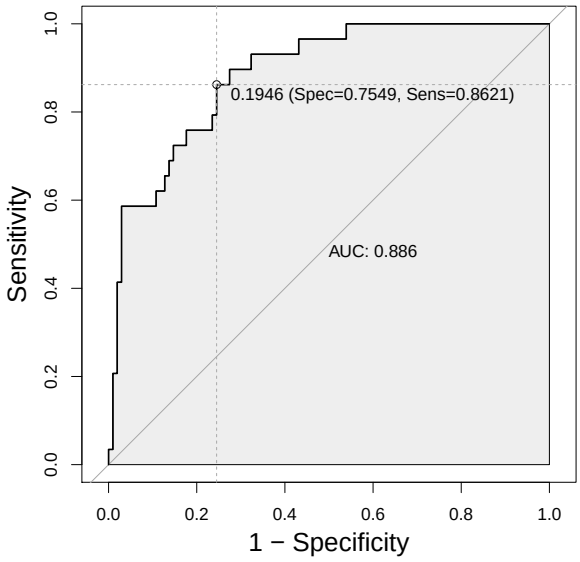

Figure 9: ROC : Neural Network

From Figure 9, we determined the best cut-off as 0.1946. At this cut-off, the sensitivity was 0.8621 and specificity 0.7549 (1− specificity = 0.2451), the classification accuracy was 0.7786 and the AUC 0.886.

4.3.4 *Validation and Prediction on Mixed Dataset:*

The neural network predicted 142 cases to progress to ESRD = 1 within 10 years. It validated 43 cases as true

Table 4: Neural Network: Attribute Importance

Name	Attribute Importance
GFR	24.5698
CRES%	12.5496
HB	11.1082
GS%	10.1918
GLOM	8.9621
CHOL	7.2360
AGE	6.4130
SMHX	4.1041
SBP	3.0350
IF	2.5258
SEX	2.3493
TA	2.3475
ALB	1.6005
BMI	1.2089
SS%	0.9182
II	0.6757
PU	0.2045

positives, with sensitivity = 0.9348.

4.4 Model Assessment: Comparative Study

Table 5: Performance Comparison of Classifiers on Test Set

Classifier	Cut-off	Accu.	Sens.	Spec.	AUC
Decision Tree	0.2155	0.8167	0.7931	0.8235	0.853
Logistic Regression	0.1902	0.7862	0.8276	0.7745	0.878
Logistic Regression (stepwise)	0.1559	0.7633	0.8966	0.7255	0.876
Neural Network	0.1946	0.7786	0.8621	0.7549	0.886

Table 6: Performance Comparison of Classifiers on Mixed Dataset

Classifier	Sensitivity
Decision Tree	0.8478
Logistic Regression	0.9783
Logistic Regression (stepwise)	0.9783
Neural Network	0.9348

4.4.1 Test Dataset

We compared the performance (discriminatory ability) of the models generated from the learning algorithms on the test set. GFR (computed based on the attributes AGE and CR), GS% and CRES% were the three common attributes chosen by all models. The measures such as cut-off, accuracy (Accu.), sensitivity (Sens.), specificity (Spec.) and AUC are shown in Table 5. All AUC estimates lay between 0.8 and 0.9, meaning that all models were good classifiers.

4.4.2 Mixed Dataset

We compared validation and prediction of the models on the mixed dataset. Table 6 shows that both the logistic regression models, whether with all attributes or with selected attributes (stepwiseAIC) validated the true positive cases with high sensitivity = 0.9783. The sensitivity measures of the other models were also good.

Among the 402 (ESRD = 0) cases in the mixed dataset, the decision tree predicted 94 cases, logistic regression 154 cases, stepwise logistic regression 171 cases, and the neural network 142 cases to progress to ESRD = 1 within 10 years.

5. ANALYSIS

Decision trees are more comprehensible than other learning models. As they are commonly used by medical practitioners for diagnosis, we further analyse the results.

The decision tree of Figure 4 clearly indicates that the initial disease stage at first presentation is critical to the probability of progression to ESRD within 10 years. Low GFR and high 24-hour proteinuria (i.e. ineffective processing by the kidneys), high crescent cell percentage (i.e. visible damage to cells on microscopic examination) and high percentage of global glomerulosclerosis (i.e. macroscopic damage) are all indicative of poor prognosis, as would be anticipated. The specific cutoff values can be useful to clinical practitioners.

The only surprising outcome of the tree is the relationship with systolic blood pressure – almost inverse to the findings of other researchers [1]. Individuals presenting with low GFR but a relatively lower percentage of crescent cells and global glomerulosclerosis (i.e. worse processing by the kidneys, but less obvious damage) have better progression if they *are* hypertensive or prehypertensive. One tentative explanation is that these patients need higher renal perfusion to preserve their remaining kidney function, which higher blood pressure promotes. This unanticipated interaction between the variables certainly warrants further investigation.

6. SUMMARY AND CONCLUSIONS

We built classifiers for predicting the probability of ESRD in IgAN patients within 10 years using decision trees, logistic regression (all variables and stepwise selection) and neural networks. All four classifiers had good AUC performance, and provide useful information to the practitioner.

At the best cut-off value of each model, the classifier emphasised specificity in the case of classification/regression trees, and sensitivity in the other three cases. Thus, for both clinical application (e.g. in determining treatment) and in informing patients, it would be important for decision making to be informed by all these results, depending on the relative weighting to be given to type 1 and type 2 errors.

In all four models, the presenting glomerular filtration rate, the extent of global glomerulosclerosis (macroscopic appearance) and the percentage of crescent cells (microscopic appearance) are prognostically important. Patients who already have significantly impaired kidney performance generally have poorer outcomes; even when performance is not yet badly impaired, high levels of visible damage, either microscopically or macroscopically, indicate poorer prognoses.

Of the learning methods we tested, the logistic models showed the greatest sensitivity in validating the true positive cases in the mixed dataset, although the neural networks generated the largest AUC. Differences between the

classifiers were relatively small, but may be significant in individual cases.

Overall, the model based machine learning approach for predicting ESRD status of IgAN patients after a specific period can be useful for making medical and lifetime decisions.

6.1 Future Work

Missing values in this dataset reduced the available data for the analysis. Hence, we will explore the imputation methods to potentially increase the data size. Theoretically, ensemble learning methods improve the predictor preformance. Hence, we will explore the use of ensemble methods including bagging, boosting and random forest for this application.

The relatively small sample sizes mean that it is not feasible, using these methods, to explore prediction over substantially shorter or longer periods than 10 years. They also mean that, because patients initially present at very different stages of the disease, the training data is highly heterogeneous, leading to higher prediction errors. Finally, treating this problem as a classification problem from initial data means that subsequently accumulated data is not used. Thus, for a patient eight years out from initial diagnosis, all we can offer is the same prediction that was given at the start – the highly informative subsequent progression of the GFR measurements cannot be used by the classifier.

One alternative approach, instead of modelling progression over a specific period using classification methods, is to probabilistically model the progression process itself. We are currently building a genetic programming system that learns probabilistic models describing the progression of the disease. If successful, this system should yield incremental probabilistic predictions, taking into account the progressive data for the patient, over a range of time periods.

7. ACKNOWLEDGMENTS

This work was supported by the Interdisciplinary Research Initiatives Program by College of Engineering and College of Medicine, Seoul National University(2014). Our special thanks to Division of Nephrology, Seoul National University Hospital (SNUH) for providing the data. The Institute of Computer Technology (ICT) at Seoul National University provided research facilities for the study.

8. ADDITIONAL AUTHORS

Jung Pyo Lee (Department of Internal Medicine, Seoul National University Boramae Medical Center, South Korea, email: `nephrolee@gmail.com`), Yon Su Kim and Dong Ki Kim (Internal Medicine, College of Medicine, Seoul National University, South Korea, `yonsukim@snu.ac.kr and kkim73@gmail.com`), and RI (Bob) McKay (Computer Science and Engineering, College of Engineering, Seoul National University, South Korea, `rimsnucse@gmail.com`).

9. REFERENCES

[1] L. P. Bartosik, G. Lajoie, L. Sugar, and D. C. Cattran. Predicting progression in IgA nephropathy. *American Journal of Kidney Diseases*, 38(4):728 – 735, 2001.

[2] J. Berger and N. Hinglais. Intercapillary deposits of IgA-IgG. *Journal d'Urologie et de Nephrologie*, 74(9):694, 1968.

[3] J. Berger, H. Yaneva, B. Nabarra, and C. Barbanel. Recurrence of mesangial deposition of IgA after renal transplantation. *Kidney Int.*, 7(4):232–241, 1975.

[4] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

[5] I. J. Choi, H. J. Jeong, D. S. Han, J. S. Lee, K. H. Choi, S. W. Kang, S. K. Ha, H. Y. Lee, and P. K. Kim. An analysis of 4,514 cases of renal biopsy in Korea. *Yonsei Medical Journal*, 42(2):247–254, 2001.

[6] G. D'amico. The commonest glomerulonephritis in the world: IgA nephropathy. *Quarterly Journal of Medicine*, 64(3):709–727, 1987.

[7] A. J. Dobson. *An introduction to generalized linear models*. CRC press, 2001.

[8] G. D. Garson. Interpreting neural network connection weights. *Artificial Intelligence Expert*, 6(4):46–51, 1991.

[9] B. A. Julian, F. B. Waldo, A. Rifai, and J. Mestecky. IgA nephropathy, the most common glomerulonephritis worldwide: a neglected disease in the United States? *The American Journal of Medicine*, 84(1):129–132, 1988.

[10] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1145, 1995.

[11] A. Koyama, M. Igarashi, and M. Kobayashi. Natural history and risk factors for immunoglobulin A nephropathy in Japan. *American Journal of Kidney Diseases*, 29(4):526–532, 1997.

[12] Y. Lau, K. Woo, H. Choong, Y. Zhao, H. Tan, W. Cheung, and H. Yap. ACE gene polymorphism and disease progression of IgA nephropathy in Asians in Singapore. *Nephron*, 91(3):499–503, 2002.

[13] H. Lee, D. K. Kim, K.-H. Oh, K. W. Joo, Y. S. Kim, D.-W. Chae, S. Kim, and H. J. Chin. Mortality of IgA nephropathy patients: a single center experience over 30 years. *PloS one*, 7(12):e51225, 2012.

[14] A. Levey, J. Bosch, J. Lewis, T. Greene, N. Rogers, D. Roth, et al. Modification of diet in renal disease study group: A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Ann Intern Med*, 130(6):461–470, 1999.

[15] M. Levy and J. Berger. Worldwide perspective of IgA nephropathy. *American Journal of Kidney Diseases*, 12(5):340–347, 1988.

[16] L.-S. Li and Z.-H. Liu. Epidemiologic data of renal diseases from a single unit in China: analysis based on 13,519 renal biopsies. *Kidney International*, 66(3):920–923, 2004.

[17] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 1996.

[18] T. M. Therneau and E. J. Atkinson. An introduction to recursive partitioning using the RPART routines. Technical report, Stanford University, 1997.

[19] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 2002.