Using Machine Learning to Refer Patients with Chronic Kidney Disease to Secondary Care

Lee Au-Yeung
Swansea University Medical School
Swansea, United Kingdom SA2 8PP

Xianghua Xie
Department of Computer Science
Swansea University
Swansea, United Kingdom SA1 8EN

James Chess
and Timothy Scale
Wales Kidney Research Unit
and Morriston Hospital
Swansea, United Kingdom SA6 6NL

***Abstract*—There has been growing interest recently in using machine learning techniques as an aid in clinical medicine. Machine learning offers a range of classification algorithms which can be applied to medical data to aid in making clinical predictions. Recent studies have demonstrated the high predictive accuracy of various classification algorithms applied to clinical data. Several studies have already been conducted in diagnosing or predicting chronic kidney disease at various stages using different sets of variables. In this study we are investigating the use of machine learning techniques with blood test data. Such a system could aid renal teams in making recommendations to primary care general practitioners to refer patients to secondary care where patients may benefit from earlier specialist assessment and medical intervention. We are able to achieve an overall accuracy of 88.48% using logistic regression, 87.12% using ANN and 85.29% using SVM. ANNs performed with the highest sensitivity at 89.74% compared to 86.67% for logistic regression and 85.51% for SVM.**

I. Introduction

A study by K Jameson et. al. in 2014 showed that in 2010, the prevalence of stage 3-5 chronic kidney disease (CKD) affects 5.9% of the population of the UK [1]. In the early stages of CKD, patients are manged in primary care in the UK. General Practitioners (GPs) will decide when patients should be referred to the specialist renal team. An Application named Assist-CKD is used in several hospitals to aid with the management of patients with CKD to try and improve outcomes for patients. Assist-CKD is fed with patient blood test data. Blood test readings are presented to the Assist-CKD operators in the form of graphs of temporal blood test readings. Using their clinical expertise, they will click on an alert button if they judge from the chart that a patient would benefit from referral by a GP to the renal team in the near future.

Machine learning can aid in the process of analysing the Assist-CKD graphs. Depending on workload, one or two Assist-CKD operators spend about 1 hour each week reviewing patients' estimated glomerular filtration rate (eGFR) charts. Taken over a year this time adds up to between 52 and 104 hours a year, per site. If this process of assessing the set of blood test results could be automated, it would lead to savings of NHS staff time. Computerising the process with machine learning makes the process systematic and more consistent. Previous performance can be reviewed objectively and the system can be continuously improved in the future.

II. Background and Related Work

In recent years, several studies have been carried out using machine learning techniques to aid in clinical decision making. Some studies have looked into developing applications to diagnose the current health of kidneys. Other studies investigated methods of predicting the future progression of kidney disease. At the time of writing, most of the papers we found which used machine learning in kidney medicine focused on the early diagnosis of CKD. This is where the greatest benefits to patients are.

In 2017, H. Polat et. al. studied the use of 24 variables combined with K-nearest neighbour, support vector machine (SVM) and soft independent modelling of class analogy (SIMCA) classifiers for the early diagnosis of CKD. Use of both SVM and K-nearest neighbour achieved an accuracy of 99.7% with SIMCA achieving 93.5% [2]. In 2015, M. Diciolla et. al. studied the use of machine learning for the diagnosis of IgA Nephropathy (a cause of CKD). The study explored the use of artificial neural networks (ANN), neuro-fuzzy systems (NFS), support vector machines (SVM) and decision tree (DT) classifier algorithms [3]. In this study, artificial neural networks performed the best with an accuracy of 90.1%. Jamshid Norouzi et. al. used an Integrated Intelligent Fuzzy Expert System to predict GFR variations at 6, 12 and 18 month intervals. 465 CKD patients where used in the study. The model could predict the GFR to $> 95\%$ accuracy. They found that the variables with the best correlation to the eGFR at 6 months are underlying disease, weight, GFR and diastolic blood pressure [4]. In 2017, K. Jeberson et. al. studied the use of 11 decision tree classifiers against 11 variables for use in a screening process for diagnosing CKD. they could get a predictive accuracy of 99.75% using the C4.5 decision tree classification algorithm. [5]. All these studies used a range of variables, for example, age gender, serum creatinine, hypertension, blood pressure &c. In Jeberson et. al., 2017, the finding reinforces what doctors already know – that eGFR is the most reliable indicator of the progression of kidney disease.

The data set used in the studies by Jeberson et. al., 2017 and H. Polat et. al., 2017 consisted of data from 400 individuals from the UCI repository, with 250 patients classified as having CKD and 150 patients classified as not having CKD. We believe that the proportions of participants with or without

CKD in this dataset and the low number of individuals may cause bias in the results. However, these studies do prove that machine learning is worth investigating further in this area. The data from the Assist-CKD application that we will be using will typically have data for approximately 12,000 patients from the Swansea renal unit alone. Since Assist-CKD is a production system in current clinical use we can make a more realistic evaluation of our machine learning techniques using its data. In our paper, the primary variable used in our prediction modelling is the set of eGFR readings for each patient. Other papers used the most recent test result for eGFR and/or other recent test results of bodily function and sometimes other medical diagnoses in their classification models. Our paper differs from other studies in the use of machine learning in kidney medicine in that it is using a chronological set of eGFR readings for each patient and no other tests of bodily function. Other additional variables in electronic health records can be used. Given the difficulty and expense in obtaining a variety of data for patients in the clinical setting where Assist-CKD is used and the preference for machine learning techniques to have minimal dimensionality (to save computational time), it is preferable to use as few variables as possible. It may be expensive to obtain various data and could mean more time taken up by NHS staff and more time required by patients attending clinics to obtain test data on them.

III. Problem Analysis

Figures 1 shows charts presented by the Assist-CKD application. Each chart represents a plot of temporal blood test readings for an individual patient. Each point on the chart is a computed eGFR value based on a patient blood test. For ethnicity, we only used white British readings in our study since the plots are the same shape.

From the charts shown in Figure 1, the data characteristics and the problems they pose in machine learning area that:

- Each point on the graph represents a blood test sample. Blood tests are taken at very irregular time intervals for each patient.
- The sampling frequency if very low, hence, the data is very sparse.
- The data is temporal data, which is a form of sequential data. We need to take sequence into account when choosing what feature extraction techniques to use.

IV. Experimental Data

Sample data from the Assist-CKD system was provided by kidney doctors at Morriston Hospital. Use of this blood test data is deemed ethical for this project because the data was provided by kidney doctors for the sole purpose of research into methods of assessing the progression of CKD. Thus, this research will potentially help existing patients as well as future patients. We did not need to interact with patients directly and take blood samples for them specifically for this project. The blood test data is data that has already been collected by the kidney doctors for the purpose of treating patients with

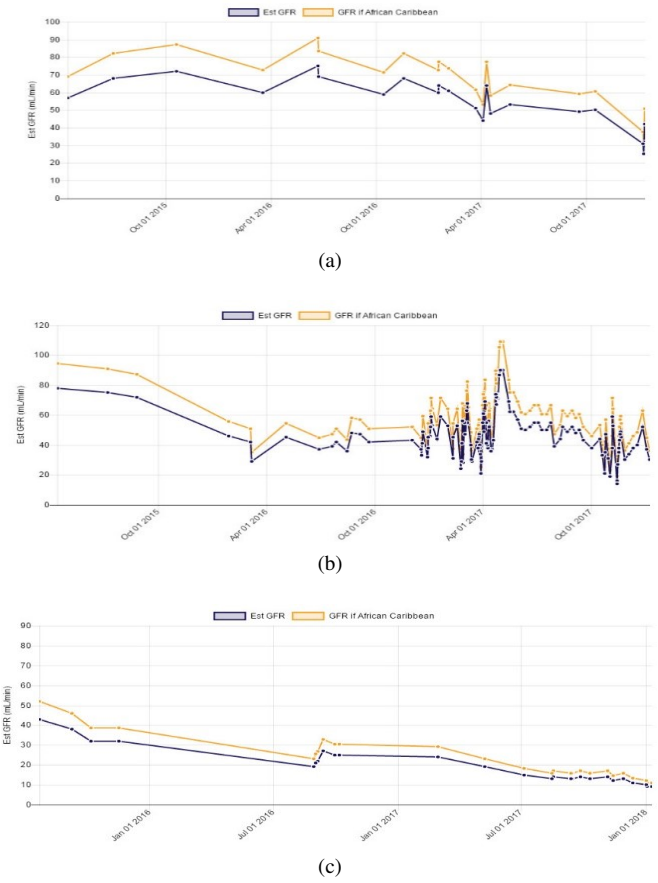


Fig. 1. Examples of eGFR graphs from Assist-CKD. The blue line represents the reading for white British patients and the green line represents the reading for African Caribbean patients. Several characteristics of the data can be observed from these example charts. These characteristics will inform the techniques to be used for the machine learning system.

potential kidney disorders. This project involves research into methods which aid the improvement of the management of kidney disease.

The data is stored in two tables. One table holds blood test results for each patient with variables including year of birth, sex, test location, test date and test result. The other table holds list of each instance of a graph being checked with the test date and a flag indicating whether an alert was raised.

A. Data Preparation

For our model we are interested in assessing patients who are yet to progress to End Stage Renal Disease (ESRD).

Figure 1b is an example of a chart where a patient is on dialysis treatment. Such patients will typically have more than 1 reading the same day which are different. Data for patients on dialysis treatment was removed from the dataset.

In our model we are looking at analysing the graph data up to the date the alert is raised. Any readings taken after an alert was raised were deleted because such readings can skew our models.

Having removed the patients and data which could not be used in our experimentation, we were left with blood test

records for 3,729 patients to be used for testing and training data.

V. Proposed Methods

A. Classification Algorithms

We experimented with 3 classification algorithms: Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Networks. These classifications algorithms require the data to be in a consistent tabular format for the data to be comparable.

Decision Trees - We considered the use of decision tree algorithms. Decision trees are fast to train and are very fast in making predictions. They are simple for humans to analyse their decision process [5]. There is a caveat to be considered when using decision tree based algorithms. One of the most attractive properties of decision tree-based algorithms is their ease of interpretation. However, a study conducted by Hastie et. al. in 2001 shows that the tree structure that is learned is very sensitive to the data being used. A slight change in the data can lead to a significant change in the structure of the tree that is generated. [6, p. 666]. Decision trees can be used in an ensemble to reduce variance. However, this would reduce the ease of interpretability of the classifier model.

Logistic Regression - The logistic regression-based classifier is a simple and robust algorithm from statistics. Logistic regression uses maximum likelihood to determine the model parameters. The workings of a logistic regression classifier can be visualised using a nomogram. Thus, the classification decisions made from it are easy to explain [7]. Logistic regression models are fast to train and fast at making predictions.

Artificial Neural Networks - A lot of research has been done on building diagnosis and predictive models in medicine using Artificial Neural Networks [7, p. 91]. These models have been shown to perform well with very good predictive accuracy [7] [8]. Since neural networks are based on a set of weightings, it is not a straightforward process to extract the reasoning behind the predictions a neural network would make. Algorithms which don't offer a simple method of extracting reasoning are termed "black-box" models. Conversely, algorithms which enable a simple way of displaying the reasoning behind their predictions is are termed "white box" models. Neural networks are classed as a black box model. It is a non-trivial problem to explain their inner workings. This makes harder to build up confidence in their use [9].

ANNs are fast at making predictions. However, a large volume of training data is preferred for ANNs to achieve good results. The larger, the better. ANNs can take a long time to train. Since ANN training algorithms lend themselves to parallel computation, the training time for ANNs can be greatly reduced by using modern highly parallel Graphical Processor Units (GPU).

Support Vector Machines - Support vector machines (SVM) are very effective classifiers and are popular in image classification. SVMs work by constructing a hyperplane in high dimensional space. The best separation between classes is achieved by having the maximum distance between points of different classes. SVMs can be constructed with different types of kernels. Examples of kernel functions are: linear, polynomial and radial. Apart from the linear kernel, SVMs are considered a "black-box" algorithm where it is difficult to extract an explanation of how they make their predictions. With a linear kernel, the structure of the model can be visualised easily by extracting the support vector coefficients that define the hyperplane separating the classes. If the extracted features are linearly separable and a linear kernel can be applied, then SVMs would be a good candidate algorithm for use in a clinical decision system [7]. The time complexity of the SVM algorithm is quadratic. [AW08] This means that with a large dataset, an SVM classifier can take a very long time to train. The Support Vector Machine (SVM) is one of the best performing algorithms for predictive accuracy [7, p. 86].

B. Feature Extraction

1) Making Graphs Comparable: Since blood tests are taken at very irregular time intervals for each patient, the graphs are not systematically comparable to each other directly. To make the graphs directly comparable and suitable for a preparing a machine learning feature matrix, our method of feature extraction is to take imputed values at regular time intervals. Values are imputed by linear interpolation. The value at each time interval is the interpolated value between the two nearest actual readings that the time point falls between. This is illustrated in figure 2.

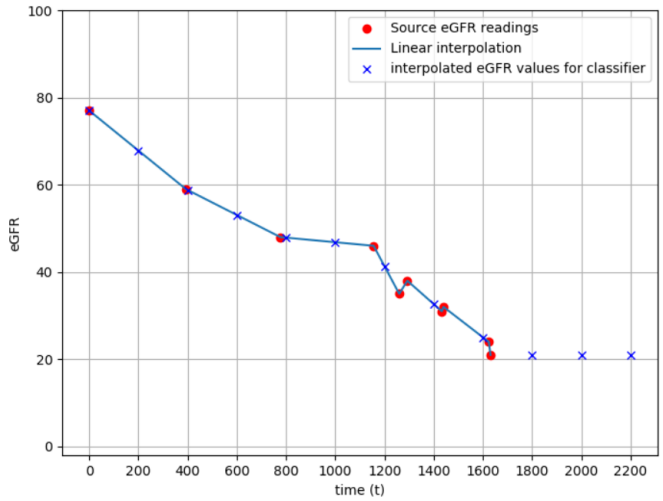
2) Dealing with Different Timescales: Patients have readings over different timescales. In our data, some patients only had one reading. On the other extreme, there were multiple readings over a timescale of 2023 days (5 years, 198 days). To make the charts comparable, we experimented with aligning the data either to the first reading or the last reading. Our hypothesis is that: aligning to the last reading, shown in 2b, will give the best results. This is because decisions are made based on how the most recent readings affect the charts.

3) Age and Sex variables: Age is a feature considered for use because clinicians take age into account in making their decisions. If a patient is very old, dialysis treatment will not be effective and it will reduce the patients quality of life. Age could be factor in how CKD progresses for a patient. In addition, the likelihood of finding a cause of CKD where the natural history could be changed with intervention or a patient benefit from planning for a transplant or dialysis is dependent on age. For that reason the threshold to "alert" on a patient varies with age. Since only the year of birth was provided, the age is taken as the difference between the year of the last test for each patient less the year of birth.

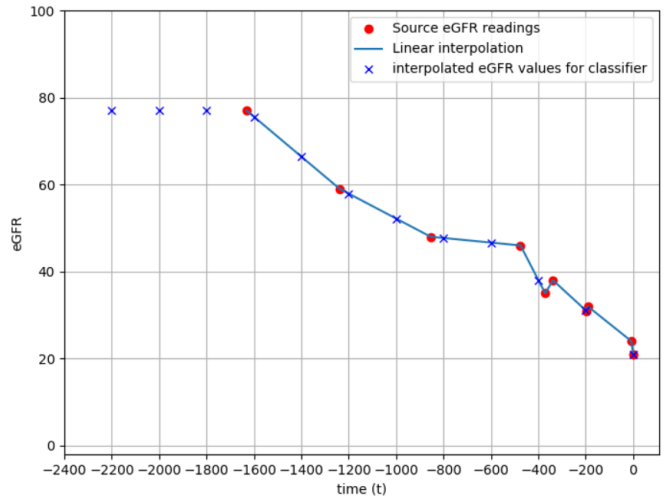
Sex is a considered feature because the gender data is supplied with Assist-CKD and research has shown that gender is a factor in the progression of Kidney disease [13]. The gender was converted to a simple numerical code: 0 = male and 1 = female for passing to a classifier.

C. Feature Matrix Construction

Fig. 3 is an illustration of the feature matrix that is prepared for passing to a classification model. Each row represents a



(a) Aligned to earliest reading



(b) Aligned to latest reading

Fig. 2. Imputation by linear interpolation between each point.

patient record. Each column represents a patient variable.

To impute the eGFR readings, we need to compute a timespan that can accommodate the longest timespan found for a patient in our dataset. For our sample data, 2040 days covered all patients. The imputation interval chosen for testing is 10 days. Hence, each patient would have 205 (including day 0 reading) imputed eGFR readings. Where a patient's timespan is shorter than the longest timespan, the last reading was used as the imputation value for all values beyond the last reading when patient readings were aligned to the earliest reading. Similarly when patient readings were aligned to the latest reading, all imputed readings before the earliest reading used the value of the earliest reading.

Different combinations using either age or sex, or both or neither were tested. The age and sex variables, when used, were concatenated as an extra dimension on the end of the feature matrix.

Once the feature matrix is constructed, it is suitable for passing to our chosen classification algorithms for use.

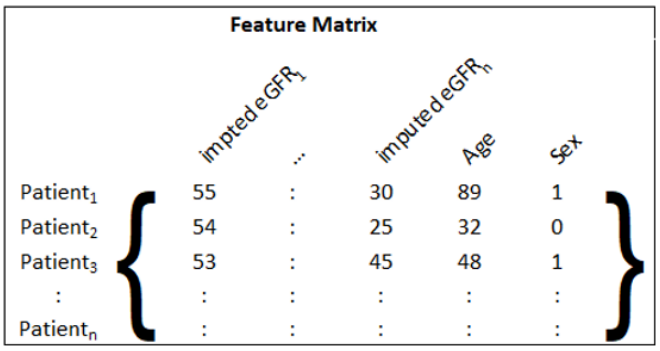


Fig. 3. Visualisation of feature matrix.

#	Type	Parameter
0	Input	max no of days for all charts/max no of imputed values, scaled to 0 .. 1
1	ReLU	Rectified Linear Unit
2	F.C.	Fully Connected with 1024 outputs
3	ReLU	Rectified Linear Unit
4	F.C.	Fully Connected with 256 outputs
5	Softmax	Softmax probability for 2 classes

TABLE I
CONFIGURATION DETAILS FOR (1024,256,2) ANN

D. Neural Network Configuration

Tables I and II shows the configuration of the neural networks we tested.

For both our neural networks the Adam optimisation algorithm was used with a learning rate of 0.001. The neural network was trained with 20 epochs with a batch size of 32.

Deciding on the number of layers and number of neurons in each layer is a non-trivial task and is currently an area of active research. We used the guidelines from "An Introduction to Computing with Neural Nets" [10]. We used two hidden layers because this keeps our neural network simple and two hidden layers is sufficient for creating classification regions for any required shape. We decided on the number of nodes heuristically. In the paper "An end stage kidney disease predictor based on an artificial neural networks ensemble" an systematic method was used to try and optimise the number of nodes in the hidden layer [8]. The process involved iteratively testing numbers of neurons in the hidden layer from 3 to 18, then selecting the number of neurons from the best performing model. However, we can observe that the results show random behaviour as the number of neurons are increased from 3 to 18 and there is no observable trend or convergence.

VI. EXPERIMENTATION AND RESULTS

Two methods were used to evaluate our results. The first method used was k-fold cross validation. k-fold cross validation will provide estimates of test error. This is useful to show if the feature-classification set performs consistently. The second method used was bootstrapping which provides

#	Type	Parameter
0	Input	max no of days for all charts/max no of imputed values, scaled to 0 .. 1
1	ReLU	Rectified Linear Unit
2	F.C.	Fully Connected with 512 outputs
3	ReLU	Rectified Linear Unit
4	F.C.	Fully Connected with 64 outputs
5	Softmax	Softmax probability for 2 classes

TABLE II
CONFIGURATION DETAILS FOR (512,64,2) ANN

a standard error of estimates. The testing pipeline function incorporated either K-fold cross validation or bootstrap testing.

A. Validation using K-fold Cross Validation

K-fold cross validation is a validation technique whereby repeated tests are performed on the same dataset using different cuts of the data. The results from each test are then averaged. This is useful to see if the feature-classification set performs consistently.

For our experimentation, 5 (i.e. k = 5) iterations were performed. During each cycle the testing and training data were split 80% for the training data and 20% for the testing data. For each cycle, a different 20% of testing data to every other test run was used.

B. Bootstrapping

For the top 10 best performing models, the bootstrap method was used to further evaluate the effectiveness of the machine learning model. The bootstrap method is a re-sampling technique whereby a new dataset is created from a base dataset. Data is selected at random from the base dataset with replacement, i.e. each data sample may be selected again. The new dataset is then tested against the model. We performed 100 iterations. An average of the results from all the iterations is then taken. If the average of the bootstrapping results is significantly lower than that of the result obtained from a single test, we know that the single test was performing better by chance. We can also calculate the standard error using Equation 1.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

Where σ is the standard deviation of the overall accuracy for all bootstrap classification results and n is the number of bootstrap iterations.

In our dataset of 3,729 patients, a new dataset of 2,500 patients was created for each bootstrap iteration. Bootstrapping should be performed with a high number of iterations. In our evaluation, 100 bootstrap iterations were performed. For each iteration, the data was sampled from the base data using a 90:10 percentage split. 50% of patients in the new dataset were selected from the 90% of the base data and 50% were selected from the remaining 10%.

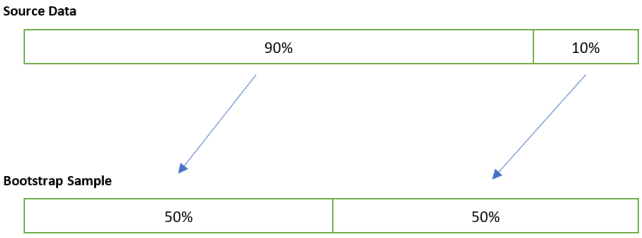


Fig. 4. Illustration of bootstrap sampling proportions.

		Predicted	
		No	Yes
Actual	No	**TN**	**FP**
	Yes	**FN**	**TP**

TABLE III
CONFUSION MATRIX LAYOUT. TN = TRUE NEGATIVE, TP = TRUE POSITIVE, FP = FALSE POSITIVE AND FN = FALSE NEGATIVE.

C. Model Selection Criteria

When testing a binary classification model, a confusion matrix can be generated which is a 2×2 matrix as shown in table III. From the figures in the confusion matrix, various performance indicators can be calculated.

The overall accuracy (a) of a model is the proportion of correctly classified cases among all cases in the test set. The sensitivity, also known as recall (r) is the proportion of all correctly classified positive cases among all actual positive cases. The specificity (n) is the proportion of all correctly classified negative cases among all actual negative cases. Precision (p) is the proportion of correctly predicted positives among all positive predictions. The F-measure (f) is the harmonic mean of precision and sensitivity. The f-measure provides a single integrated score of both precision and sensitivity together.

To judge which would be our best classification model, this is the criteria we use:

Accuracy - The model must be one of the most accurate models overall, i.e. accuracy (a) must be among the highest.

Predicted Results Bias - In our data we have 77% negative (patient not marked) and 23% positive (patient marked) results. We could have models which appear to be 77% accurate, for example. However, it is possible for such a model to only be predicting a negative result for every patient. Hence, we need to examine other accuracy scores such as sensitivity and specificity to check that there is no a significant bias towards one type of result. Hence sensitivity (r) and specificity (n) must be similar.

Consistency – models need to be tested sufficiently using different cuts of the dataset for training and testing to check that the variance is low between different tests.

Appropriateness – for use in a clinical context. Ideally, we want a model to be 100% accurate. Realistically, it is rare for a machine learning model to achieve 100% accuracy.

Errors are made up of false positive and false negative predictions. Depending on the application, certain errors may be less acceptable than others. In a clinical context, incorrectly predicting a negative result can lead to serious consequences for patients who could miss out on critical interventions to protect their health. Hence, among models with a similar level of overall accuracy we would firstly prioritise models with the highest sensitivity (r). Incorrectly predicting a positive result will result in unnecessary worry and inconvenience for patients and unnecessary costs to the health service. Hence, we would look for a high precision (p) score. In the interest of patient safety, a higher false positive rate would be favourable to a higher false negative rate.

1) Comparison of Best Models: Our top 10 best performing models are listed in table V. The training time was recorded so that models with an excessively long training time are discarded since it may not be feasible to use such models in practice. All our top 10 performing models took under a minute to train.

2) Bootstrap Results: For the top 10 overall best performing models listed in Table V, we additionally verified them with bootstrapping to check if the models still perform as well with more rigorous testing. With the bootstrapping tests, 100 test iterations were conducted.

In bootstrap testing for the top 10 models, the logistic regression models performed to a similar level of accuracy to the tests using k-fold cross validation. Our SVM based models, generally performed worse. One model, Model 1 which was the best performing model when tested in k-fold cross validation, performed significantly less well. Model 1 was assessed to have an accuracy of 90.64% accuracy when tested under k-fold cross validation. The accuracy level fell to 85.25% when tested more rigorously using bootstrapping, a drop of 5.39%. This model also showed a high standard error compared to the Logistic Regression based models. The model does not perform very consistently. The results varied more than for other models in repeated tests. The accuracy for Model 8 also noticeably dropped by 12.34% in predictive accuracy with bootstrapping compared to k-fold testing. All other logistic regression classifiers varied in accuracy by up to +/- 1%. Most of the best performing classifiers were biased towards making false negative predictions.

From the results of bootstrap testing, we judge our best prediction model to be Model 5 in table VI. It had an overall accuracy of 88.41%. The sensitivity was 86.67.41% and the specificity was 89.02%. The mean accuracy for Model for under k-Fold cross validation was 87.96%. The standard error was much lower compared with Model 1. Model 5 has been shown to perform consistently.

3) Receiver Operating Characteristic Curve: For each iteration in our machine learning pipeline function, a Receiver Operating Characteristic (ROC) curve was generated. Figure 5 shows a ROC curve from our best performing model Model 5. With the curve being far away from the chance line, we can see that the model is not simply making random predictions.

#	Classifier and Feature Set Description	Date Align
1	**SVM(LK)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including age	L
2	**SVM(LK)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including age and sex	L
3	**ANN (1024,256,2)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including age	L
4	**LogReg** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including age	L
5	**LogReg** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including age and sex	L
6	**ANN (1024,256,2)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings	L
7	**LogReg** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings	L
8	**ANN (512,64,2)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including sex	L
9	**LogReg** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings, including sex	L
10	**ANN (512,64,2)** matrix of interpolated eGFR at equal chronological time intervals, interpolation by value between 2 real readings	L

TABLE IV
DESCRIPTION OF TOP 10 PERFORMING MODELS TESTED WITH K-FOLD CROSS VALIDATION. THE DATE ALIGNMENT ABBREVIATIONS ARE L = LATEST DATE, F = FIRST DATE

Our model is intended for use in a clinical setting. If we find multiple models with high overall accuracy, we would favour a model that is favours false positives over false negatives. A curve which shows that the model is favouring false positives over false negatives would have the curve rising more steeply from the origin.

From the comparative graph shown in Figure 6, there is not a significant observable difference between our 10 best performing models in terms of bias towards either false positive predictions and false negative predictions.

4) Confusion Matrix: Table VII shows the confusion matrix from the bootstrap testing. The figures TP, TN, FP and FN are based on the average for all tests for each model. We can observe that while the most accurate models overall are the ones based on the logistic regression classifier which range from 88.05% to 88.48%. However, it is noticeable that the ANN classifiers tend to be more accurate at predicting true positive results. The sensitivity ranges from 88.36% to 89.30%, which is higher than the highest sensitivity for the

Model #	Avg Training Time (s)	Avg Overall Accuracy	Avg Sensitivity	Avg Specificity
1	1.95	90.64%	81.40%	93.37%
2	1.94	89.54%	91.86%	88.83%
3	37.48	89.11%	72.09%	94.07%
4	1.43	88.01%	88.95%	87.61%
5	1.38	87.96%	88.37%	87.78%
6	21.80	87.64%	87.79%	87.61%
7	1.32	87.53%	88.95%	87.09%
8	18.49	87.34%	86.63%	87.43%
9	1.47	87.18%	88.95%	86.74%
10	22.45	86.91%	81.98%	88.31%

TABLE V
TOP 10 OVERALL RESULTS FROM USING LOGISTIC REGRESSION, SVM OR ANN CLASSIFIER USING LINEAR INTERPOLATION VALUES BETWEEN READINGS. THESE RESULTS ARE TAKEN FROM TESTING WITH K-FOLD CROSS VALIDATION.

Model #	Avg Training Time (s)	Avg Overall Accuracy	Avg Sensitivity	Avg Specificity	Overall Accuracy Standard Error
5	6.47	88.48%	86.67%	89.02%	0.094115353
9	6.71	88.14%	86.50%	88.63%	0.104084011
4	6.42	88.09%	86.03%	88.71%	0.085024655
7	6.50	88.05%	86.08%	88.64%	0.080905047
8	48.21	87.12%	88.36%	86.74%	0.442102252
6	31.93	86.94%	89.01%	86.31%	0.549315377
10	15.21	86.60%	89.74%	85.65%	0.605641588
3	15.69	86.61%	89.30%	85.81%	0.544823282
1	0.40	85.29%	85.51%	85.23%	0.827364253
2	0.39	84.78%	80.29%	86.14%	0.843021240

TABLE VI
TOP 10 RESULTS FROM USING ANN CLASSIFIER USING LINEAR INTERPOLATION VALUES BETWEEN READINGS.

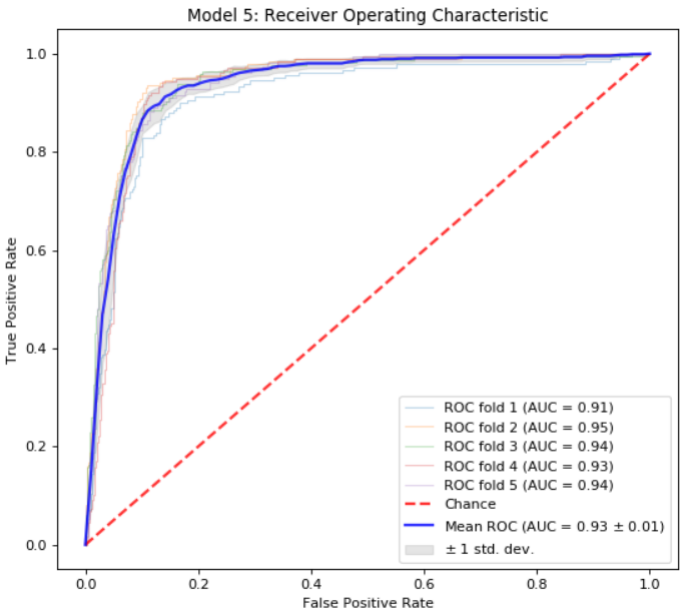


Fig. 5. ROC Curves for all k-Fold iterations for Model 5.

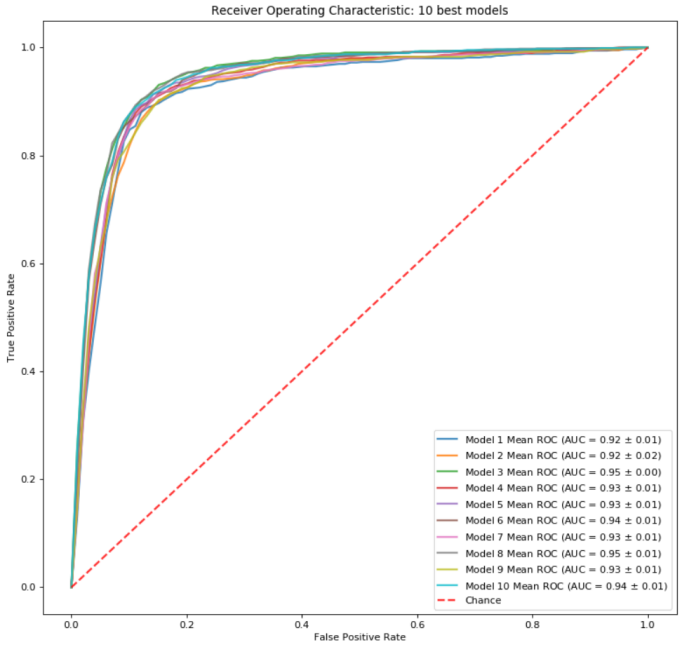


Fig. 6. Mean ROC Curves for 10 best performing Models for comparison. All 10 models were of a similar shape.

logistic regression classifiers 86.67%.

5) Reading Alignment: In section V-B2, we posed the hypothesis that aligning all the charts to the last reading for comparison would most likely produce the best results in terms of classification accuracy. Figure 7 shows a chart comparing all the models which we experimented with.

Among the best classifiers with an accuracy of over 80%, it is clear, that aligning the readings to the last reading resulted in the more accurate classification models. Among the classifiers with a predictive accuracy of below 80%, there is no trend favouring aligning the readings either way. From this we conclude that when experimenting with other models in future, and the models require aligning the charts for comparability, both alignments should be experimented with. However, aligning the data to the last readings are likely to generate the most accurate models.

VII. FUTURE RESEARCH

The imputation interval was arbitrarily chosen. A more systematic method should be used to re-run the experiments at different imputation time intervals to ascertain whether

Model #	TP	TN	FP	FN	r (%)	n (%)	a (%)	p (%)	f (%)
5	811	234	100	36	86.67	89.02	88.48	70.06	77.48
9	803	237	103	37	86.50	88.63	88.14	69.71	77.20
4	809	234	103	38	86.03	88.71	88.09	69.44	76.85
7	804	235	103	38	86.08	88.64	88.05	69.53	76.92
8	785	243	120	32	88.36	86.74	87.12	66.94	76.18
6	782	243	124	30	89.01	86.31	86.94	66.21	75.94
10	776	245	130	28	89.74	85.65	86.60	65.33	75.62
3	780	242	129	29	89.30	85.81	86.61	65.23	75.39
1	773	236	134	40	85.51	85.23	85.29	63.78	73.07
2	783	220	126	54	80.29	86.14	84.78	63.58	70.97

TABLE VII
CONFUSION MATRICES FOR BEST PERFORMING MODELS FROM BOOTSTRAP TESTING.

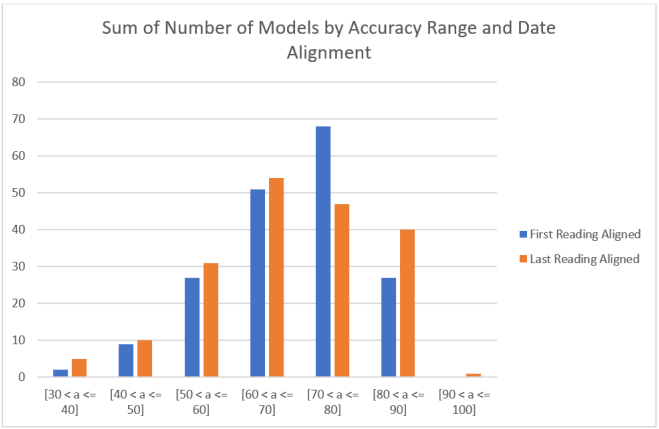


Fig. 7. Chart to Analyse Reading Alignments. The models included in this chart are all the models where we considered reading alignment could have an effect on the results.

the most accurate classification results are arrived at when converging to a particular time interval.

Convolutional neural networks (CNN) are a recommended avenue of future experimentation. CNNs may offer the benefit of removing the feature extraction step of having to impute results if the number of inputs are set to the highest number of days covered by any patient chart. This would also remove the potential need to search for an optimal chronological imputation interval.

VIII. CONCLUSION

Our ANN models using the linear interpolation imputed readings between actual eGFR blood test readings merit serious consideration. While they are not the most accurate models overall, they have the highest rates of sensitivity meaning that they are less likely to miss positive cases. They could become more accurate overall as more training data is made available. Their main disadvantage is that they are black box models. They have also been shown to be less stable than the Logistic regression models. However, their stability can be improved by providing more training data and by using an ensemble of ANN models.

Our most accurate model performs to an accuracy of 88.48% in the more rigorous bootstrap testing. It has been shown to perform consistently. This model uses linear interpolation imputed readings between actual eGFR blood test readings with age and sex. It is one of our simpler models. Originally tested with fewer samples it performed to an accuracy of 86.4%. We believe that with more data the accuracy will improve. Since it is using the Logistic Regression classifier, it was shown to be quick to train and quick to generate predictions with the model. The logistic regression classifier is also easy to explain the results with. Our tests have shown Logistic regression models to be very stable. This is the model we would recommend to use to develop an application initially.

REFERENCES

[1] K. J. et. al., "Prevalence and management of chronic kidney disease in primary care patients in the uk," https://www.ncbi.nlm.nih.gov/pubmed/24852335, 2014.
[2] H. P. et. al., "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," https://www.ncbi.nlm.nih.gov/pubmed/28243816, 2017.
[3] M. D. et. al., "Patient classification and outcome prediction in iga nephropathy," https://www.ncbi.nlm.nih.gov/pubmed/26453758, 2015.
[4] J. N. et. al., "Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system," https://www.hindawi.com/journals/cmmm/2016/6080814/abs/, 2016.
[5] K. J. et. al., "Machine-learning approaches to assist in accurate and extensive chronic kidney disease screening," http://www.ijcea.com/machine-learning-approaches-assist-accurate-extensive-chronic-kidney-disease-screening/, September 2017.
[6] C. M. Bishop, "Pattern recognition and machine learning," 2006.
[7] B. Z. Riccardo Bellazzi, "Predictive data mining in clinical medicine: Current issues and guidelines," https://www.ncbi.nlm.nih.gov/pubmed/17188928, 2008.
[8] T. D. N. et. al., "An end stage kidney disease predictor based on an artificial neural networks ensemble," https://www.sciencedirect.com/science/article/pii/S0957417413000778, 2013.
[9] W. G. Baxt, "Application of artificial neural networks to clinical medicine," https://www.ncbi.nlm.nih.gov/pubmed/7475607, 1995.
[10] R. P. Lippmann, "An introduction to computing with neural nets," https://ieeexplore.ieee.org/abstract/document/1165576, 1987, [Online; accessed 28/09/2019].