Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling

Tahsin M. Rahman
Department of Electrical and Electronics Engineering
American International University - Bangladesh
Dhaka, Bangladesh
tahsinmrahman@gmail.com

Saima Siddiqua
Department of Electrical and Electronics Engineering
American International University - Bangladesh
Dhaka, Bangladesh
saimasiddiqua25@gmail.com

Siam -E- Rabby
Department of Electrical and Electronics Engineering
American International University - Bangladesh
Dhaka, Bangladesh
siamerabbypiash@gmail.com

Nahid Hasan
Department of Electrical and Electronics Engineering
American International University - Bangladesh
Dhaka, Bangladesh
nahid.altair@gmail.com

Mohammad Hasan Imam
Department of Electrical and Electronics Engineering American International University - Bangladesh
Dhaka, Bangladesh
hasan.imam@aiub.edu

***Abstract*— This paper introduces the idea of detecting the presence of kidney disease through machine learning based classification modelling, by processing the patient's ECG signal. Recent studies and ongoing researches have showed that patients undergoing kidney problems start developing cardiac problems- scientifically known as the Cardio Renal Syndrome (CRS) which can lead to a sudden cardiac arrest in the last stages of their disease. Since cardio-vascular diseases and the chronic kidney disease is inter-related, this model can be used for patients undergoing cardio-vascular problems to determine whether their kidneys have been effected or not. If the Chronic Kidney Disease (CKD) can be diagnosed at an earlier stage, it may give the patient some time to help reverse the disease or at least slow its progression by taking necessary medical steps. For this model, digitized ECG data was collected from open access databases such as PTB (for kidney patients) and Fantasia (for healthy people) from Physionet Database (www.physionet.org) and the model was later validated using different data from the same online database. The validation process gave satisfactory results, as the model could successfully classify the users from being healthy or a kidney patient. In our study, we found an accuracy level of 97.6% which was the highest using both features QT and RR interval, in comparison to the accuracy that was found when either one of the features was used.**

Keywords—chronic kidney disease (CKD), cardio-vascular problems, cardio renal syndrome (CRS), sudden cardiac death (SCD), MATLAB

I. INTRODUCTION

Chronic Kidney Disease (CKD), or the chronic renal failure, is a disease where the kidneys start to lose their functionality [1]. It is a chronic condition which causes the kidney to deteriorate and lose their ability to function properly and lastly lead to the fifth and final, fatal stage- the End Stage Renal Disease (ESRD), where the kidney functions drop to almost 10 to 15 percent of their healthy capacities [2]. When the disease has progressed to this stage, a kidney transplant or dialysis are the only options for the patient to survive. Studies have since shown that, amongst the CKD patients' death, 60% of the deaths are Sudden Cardiac Deaths (SCD) whereas the rest 40% are other cardiovascular mortalities [2]. In accordance to a report from the US National Kidney Foundation, almost 10% of the world population suffers from CKD, among which around 2 million people require dialysis or a transplant to live [3]. Narrowing down more locally, a leading daily reported that in Bangladesh, one out every seven people suffer from kidney problems and 3.24% of the population death can be traced back to kidney disease [4]. If these deaths are further traced down, it was found that the majority of these deaths were due to a sudden cardiac arrest [5]. Sudden Cardiac Death (SCD) is defined as the unexpected natural death due to a cardiac cause, in a person that does not have any prior potential fatal condition [6]. It can occur due to rhythm abnormalities in the heart, known as arrhythmias. When the heart undergoes ventricular fibrillation, an arrhythmia where the heart fires erratic, chaotic impulses from the ventricles of the heart which interrupts the sinus rhythm and thus the normal blood flow of the heart. This, in turn, causes a depletion of oxygen in all parts of the body. The ventricular fibrillation stage requires strict medical attention, otherwise the patient might succumb to SCD [6]. This phenomenon, of a kidney condition causing cardiac problems and vice versa is medically termed as the Cardio Renal Syndrome (CRS). Since the heart and kidneys are connected through various pathways that ensure a stable blood flow, a problem in one often induces a problem in the other [7].

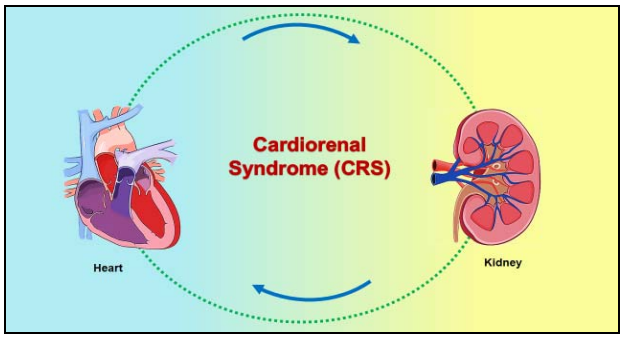


Fig. 1. Cardio Renal Syndrome (CRS)

978-1-5386-8014-8/19/$31.00 ©2019 IEEE

Thus, in reference to the CRS, it is possible to say that patients undergoing symptoms of cardio vascular diseases (CVD) may also be suffering from CKD. Since it is already known that different CVDs leave characteristics traces in the patient's ECG, it is also possible to detect the presence of CKD from the same ECG of the patient if the traces for CKD is known.

From various important studies, it was found that ECG of any patient undergoing CKD, shows some significant changes which can be traced back to CKD [9, 10]. The dynamic changes that can be observed from the ECG of CKD patients are summarized in this section. Although there are a number changes that occur, but not all the patients suffer through all the changes and this paper will only highlight the major changes that can be observed in the majority of the kidney patients. All the information here are collected from different journals where each patient underwent 12 lead ECGs or the standard Holter Monitoring Method. When interpreted by a qualified physician, it was found that the majority of the patients showed decreased Heart Rate Variability (HRV) in hemodialysis patients, the frequency component of the digital ECG was also found in lower values [9]. The QT duration was seen prolonged in CKD patients such that 460ms in female while 450ms in men and the QRS amplitude was also seen to be increased by approximately 0.18mV. A good section of the patients also showed the amplitude of the ST segment to be above 0.2mV in leads of V1 and V2 [9, 10].

The traditional method of detecting kidney diseases are the invasive methods of blood tests (e.g. GFR test) or kidney biopsy, and the non-invasive methods are ultra sound imaging or urine tests to check for creatinine levels, and they can all be uncomfortable procedures for the patient especially for very elderly patients or end stage CKD patients. Thus, this model opens up an option for patients to detect their kidney disease through a simple non-invasive way by means of their available bio signals i.e. ECG signal. ECG signal is widely used for CVD analysis as almost all types the heart abnormalities can be detected from this. Since, any patient with CVD are very likely to undergo an ECG test, through this model, the presence of CKD can also be detected from the same ECG, by checking for prolonged QT intervals and also checking RR intervals which can be used to measure the HRV. The following sections of this paper will discuss about the ECG data processing, model formation and validation stages.

II. DATA INFORMATION

For the training of the model, all the data was collected from two open access online database, PTB (used for digitized ECG of the kidney patients) and Fantasia (used for the digitized ECG of healthy people) from Physionet database (www.physionet.org). The database each contained two minute long digitalized ECG signal, of the patients, from which the two required features- QT interval and the RR interval was extracted using Berger's algorithm [13]. The patients, whose ECG were taken had an average age between 50 to 70 years old [11, 12].

For the PTB database, 290 subjects were considered, both male and female with a mean age of 57.2, 549 records were collected by taking 5 records for each subject. Each of these records included 15 simultaneously measured signals (the standard 12 leads, 3 Frank lead ECGs), which were digitized at a sampling rate of 1000 samples per second with 16-bit resolution. The sampling rate was also varied up to 10 KHz. The final data uploaded contained a clinical summary of the patient's age, gender, diagnosis [11, 12]. From this database we choose 7 subjects who were diagnosed with both CVD and CKD as presented in [15].

For the Fantasia database, each sub-groups of subjects, an equal number of men and women were taken whose ECG was recorded. This was then digitized at a sampling frequency of 250Hz and using an automated arrhythmia detection algorithm, each heart beat was annotated and finally verified by visual observation [11 ,12].

The following table summarizes brief information of the seven patients whose times series ECG were used for the training of the model.

TABLE I. INFORMATION OF KIDNEY PATIENTS [15]

Patient Index	Age	Cardiovascular Problem
Patient 13	71	Myocardial infarction
Patient 12	67	Myocardial infarction
Patient 140	75	Myocardial infarction
Patient 201	39	Cardiomyopathy
Patient 216	73	Hypertrophy
Patient 79	75	Myocardial infarction
Patient 78	68	Myocardial infarction

III. METHODOLOGY

The following figure describes the basic work flow of the work done for this paper.

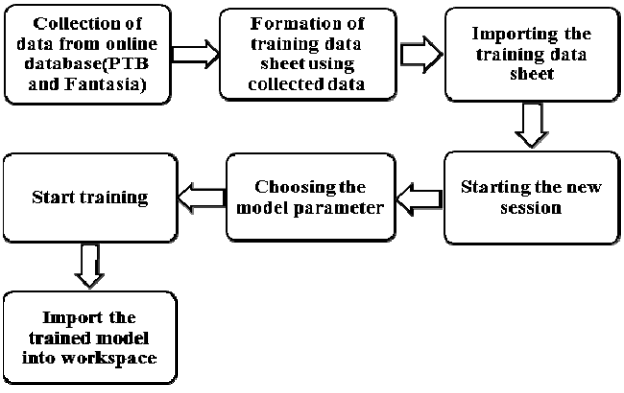


Fig. 2. Basic work flow of the model

To form the model, the first step was to extract digitalized ECG data from database. As discussed in the previous section, digitized ECG was collected from two databases - the PTB database for the kidney patients' ECG and the Fantasia database for the healthy elderly patients' ECG as the CKD patients taken form the PTB database were all elderly, to reduce the ageing effect on CVD. The ECG

signals were then processed using the Berger's algorithm, to find the required features- the QT interval and the RR interval. Using the extracted feature from the digitalized ECG, a training set was build where the patients were already labelled as 'kidney' or 'healthy'. Here, the two features act as the predictors while the label is the decision that is expected from the model to make with new unknown data.

MATLAB based application- Classification Learner app was used as for the training and validation of the model. The version of MATLAB used was 2017a since the previous versions might not have the above mentioned application. The created training set was then imported into the Classification Learner application, where the parameters for the model, i.e. the algorithm- linear Support Vector Machine (SVM), features used- QT interval and RR interval, the kernel function- linear, automatic and the cross-validation scheme was selected and the model was trained. It is very important to note that, the supervised machine learning was used because the training data set contained pre-labelled data and the ECG for both groups of patients were already known. Under supervised machine learning, SVM was chosen because it showed good performance in many studies for classification purposes. For the model, approximately 700 observations for each group were used to design a non-biased model and the cross validation scheme was chosen as the holdout validation with a degree of 50%. The trained model that is returned can be imported into the workspace and it would then appear as a user defined function, which can be used for classifying new unknown data.

The trained model, once imported to the workspace was in the form of a function where a new table format data was passed, and the function was used to classify the new data. For the validation of the model, unused data from PTB and Fantasia were again taken and tabulated into another excel file where the features were there but the decision was missing. Once passed through the function, the model delivered the correct decision in a categorical data type.

IV. RESULTS ANALYSIS

A summary of the model that was returned, is that the accuracy of the model is 97.6% and it only took approximately 10-15 seconds to get trained. The validation for the model was done using data from the aforementioned online database and it was seen that the model could classify most of the patients correctly.

TExpanded

1392x3 table

	1 QTinterval	2 RRinterval	3 Decision	4	5	6
1	0.2830	0.7040	kidney			
2	0.2890	0.6970	kidney			
3	0.3070	0.6980	kidney			
4	0.2880	0.7100	kidney			
5	0.3040	0.7050	kidney			
6	0.2880	0.7070	kidney			
7	0.3040	0.7090	kidney			
8	0.2860	0.7020	kidney			
9	0.2900	0.6990	kidney			

Fig. 3. An example of the training datasheet in the MATLAB workspace

The model can further be explained through the various plots that can describe its characteristics such as the scatter plot and the confusion matrix.

The following figure is the scatter plot of the model, the red color is used to mark the points that represent kidney patients, while the blue color is the representatives of the healthy patients. The circles are the correctly classified points where as the crosses are the points that have been misclassified, i.e. the points that were supposed to be in one class got labelled as the other. Since the model has a an accuracy of 97.6%, it can be seen that the majority of the points have been correctly classified.

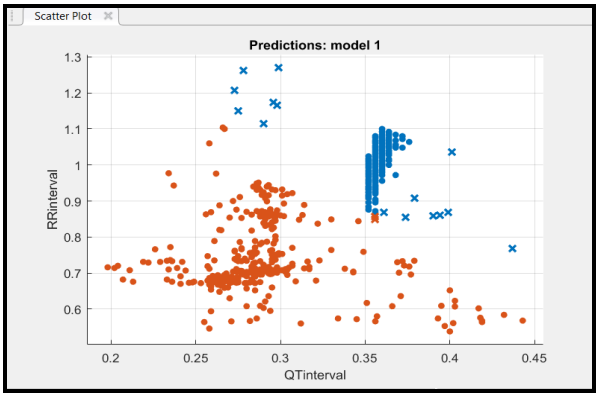


Fig. 4. The scatter plot of the model

The following figure is the confusion matrix for the model. Here the green boxes are the true values and then pink boxes are the false values. The upper green box represents the true positives i.e. the correctly classified healthy users and the lower green values are the true negatives which are the correctly classified kidney patients. The upper pink box represents the false negatives which are healthy people misclassified as kidney patients and the lower pink box represents the false positives which are kidney patients misclassified as healthy people. The three very important parameters to judge the performance of any algorithm are the sensitivity, selectivity and the accuracy and they can be all calculated easily from the plot below.

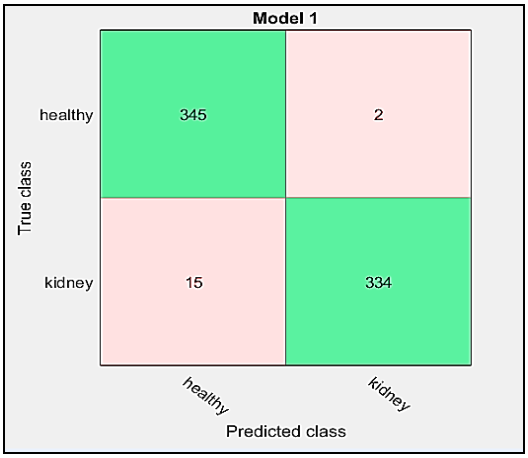


Fig. 5. The confusion matrix for the model

The above analysis was done for the model with the following chosen settings i.e. algorithm- SVM Linear, and using both the features- QT interval and RR interval. During the formation of the model, different algorithms and the number of features were also varied to see the effect that they had on the accuracies.

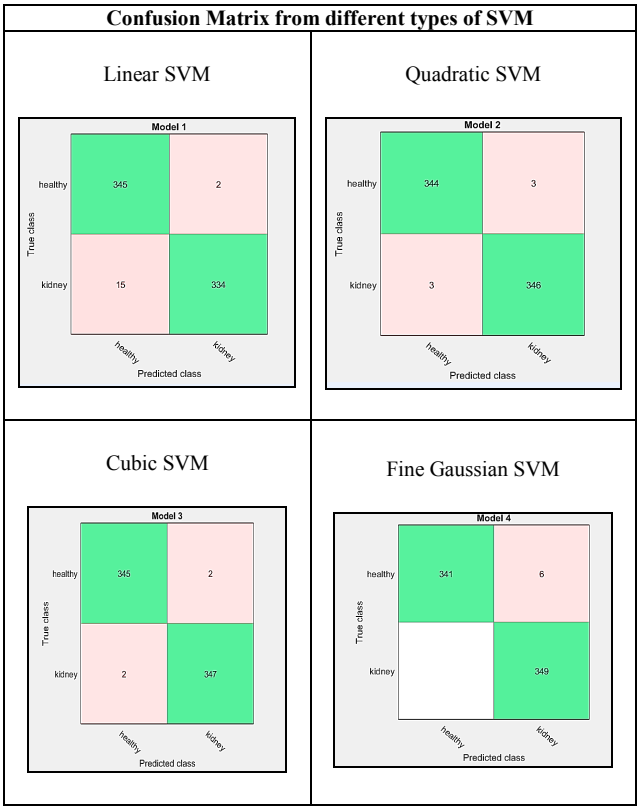
The following table shows the variation of the accuracy when the SVM type is varied:

TABLE II. VARIATION OF ACCURACY DUE TO THE TYPE OF SVM

Type of SVM	Accuracy
Linear SVM	97.6%
Quadratic SVM	99.1%
Cubic SVM	99.4%
Fine Gaussian SVM	99.1%

The effect of changing the type of SVM can be seen more clearly from the confusion matrix produced by the different models. The following table shows a combination of all the confusion matrix.

TABLE III. CONFUSION MATRIX FROM DIFFERENT TYPES OF SVM



From the above table of the four types of confusion matrix produced from the different type of SVM used, it can be seen that the Fine Gaussian SVM produced the best confusion matrix since here the false positive rate is zero because it had zero false positive points.

The model was also altered in such that, the number of features were reduced to either one of the two features used, and it was seen that the model had much better accuracy when both the features acted as predictors in comparison to when only one of them were chosen. Thus, the final model chosen was a multi-parameter model since it is clearly seen that the model can classify patients more accurately if both the features were present.

TABLE IV. CHANGES IN ACCURACY DUE TO CHANGE IN FEATURES SELECTED

Features Selected	Accuracy of the model
QT interval	94.1%
RR interval	93.4%
QT & RR interval	97.6%

V. CONCLUSION

Chronic Kidney Disease (CKD) is a worldwide epidemic in these days, which leads to millions of deaths each year. This is because CKD, when advances can lead to many CVD such as SCD in the last stages. In reference to CRS, many people suffering from cardio-vascular problems may also be a victim of the CKD, and sometimes the treatment for CVDs may also get limited due to the presence of CKDs [14]. Thus a classification based model has been developed, to be able to detect kidney disease in the early stages, from their digitalized ECG, using machine learning algorithms. It takes two concerning features- QT interval and RR interval to detect the presence of CKD in the patients, with an accuracy of 97.6%. The primary advantage of this model is the fact that it provides a safe non-invasive way for patients to determine the state of their kidneys. Since, all types of CVD can be diagnosed from the ECG of the patient and any patient suffering from any sorts of CVD, must undergo an ECG test, the same test signal can be passed onto the model to determine if the patient's kidneys are getting affected or not. If the patient's kidneys are indeed affected, then the doctor could prescribe a different medication route or any other necessary medical steps to stop the progression of the disease or even reverse its direction.

REFERENCES

[1] "About Chronic Kidney Disease", *National Kidney Foundation*, 2018. [Online]. Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease. [Accessed: 26- Sep- 2018].
[2] B. Franczyk-Skóra, A. Gluba, M. Banach, D. Kozłowski, J. Małyszko and J. Rysz, "Prevention of sudden cardiac death in patients with chronic kidney disease", BMC Nephrology, vol. 13, no. 1, 2012.
[3] "Global Facts: About Kidney Disease", the National Kidney Foundation, 2018. [Online]. Available: https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease. [Accessed: 08- Apr- 2018].
[4] "Kidney Disease in Bangladesh", World Life Expectancy, 2018. [Online]. Available: http://www.worldlifeexpectancy.com/bangladesh-kidney-disease. [Accessed: 08- Apr- 2018].

[5] National Institutes of Health. "National Institute of Diabetes and Digestive and Kidney disease". Annual Data Report. Retrieved 22 November 2013.
[6] "Sudden Cardiac Death (Sudden Cardiac Arrest) | Cleveland Clinic", Cleveland Clinic, 2018. [Online]. Available: https://my.clevelandclinic.org/health/diseases/17522-sudden-cardiac-death-sudden-cardiac-arrest. [Accessed: 12- Apr- 2018].
[7] Ronco, C.; McCullough, S.D. (2010). "Cardio-renal syndromes: Reports from the consensus conference of the acute dialysis quality initiative". European Heart Journal. 31 (6): 703–711. doi:10.1093/eurheartj/ehp507. PMC 2838681 Freely accessible. PMID 20037146
[8] Science Learning Hub. (2018). *Label the heart*. [Online] Available at: https://www.sciencelearn.org.nz/labelling_interactives/1-label-the-heart [Accessed 26 Sep. 2018].
[9] "Sudden Cardiac Death (Sudden Cardiac Arrest) | Cleveland Clinic", Cleveland Clinic, 2018. [Online]. Available: https://my.clevelandclinic.org/health/diseases/17522-sudden-cardiac-death-sudden-cardiac-arrest. [Accessed: 12- Apr- 2018].
[10] D. Zachariah, P. Kalra and P. Roberts, "Sudden cardiac death in end stage renal disease: unlocking the mystery", Journal of Nephrology, vol. 28, no. 2, pp. 133-141, 2014.
[11] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
[12] F. Schoonjans, "ROC curve analysis with MedCalc", MedCalc, 2018. [Online]. Available: https://www.medcalc.org/manual/roc-curves.php. [Accessed: 09- Aug- 2018].
[13] R. D. Berger, E. K. Kasper, K. L. Baughman, E. Marban, H. Calkins, and G. F. Tomaselli, “Beat-to-beat QT interval variability: Novel evidence for repolarization liability in ischemic and non-ischemic dilated cardiomyopathy,” Circulation, vol. 96, pp. 1557–1565, 1997.
[14] Evans, Frank. "Cardio-Renal Connections in Heart Failure and Cardiovascular Disease". NHLBI Working Group. Retrieved 22 November 2013.
[15] M. Corzo-Cuesta and C. Alvarado-Serrano, "An Algorithm for QT Dispersion Analysis: Validation and Application in Chronic Kidney Disease", 2016.