Introducing novel classification methodology for health science; A case study in kidney disease in Sri Lanka

K.A.A. Chathurangi[1], R.M.K.T. Rathnayaka[2], L.L.G. Chathuranga[3]

[1,3]Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

[2]Department of Physical Sciences and Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

Abstract:

***Background**:The healthcare sector has huge amount of medical data which are information rich and still not properly analyzed; especially, discovering useful information to predict future patterns is very limited. By using data mining techniques, the current study introduced a novel classification methodology and successfully applied it in Sri Lankan domain for the Chronic Kidney Disease (CKD) classifications.*

***Methods**: This study is carried under the two phases. In the first phase, Artificial Neural Network (ANN) method namely multilayer feed-forward neural network was used to detect whether a person has a risk of having a kidney disease or not and their risk level. In the second phase, a novel forecasting methodology is proposed using multiple algorithms, which is a combination of Random Forest algorithm and an ANN hybrid methodology to detect whether a patient has fallen into a CKD or not. The performances of the models are evaluated by using the confusion matrix using three different data samples.*

***Results**: According to the error rates and accuracy results, a modified ANN with 2 hidden layers is better to detect kidney disease and a model with the combination of Random Forest and ANN with 3 hidden layers is better to predictwhether a person is fallen in CKD or not. The constructed models give high accuracy and minimum error rates when comparing with the other data mining algorithms. Model built in first phase and second phase gives 80.952% accuracy and 19.047% error rate and 81.395% accuracy and 18.604% error rate for testing data respectively.*

***Conclusion:** The proposed model will help to both doctors and patients to improve the quality of clinical decisions including reduce medical errors and enhance patient safety.*

***Key Word**:Artificial Neural Network, Data Mining, Kidney Disease, Random Forest*

Date of Submission: 05-07-2020 Date of Acceptance: 21-07-2020

I. Introduction

The healthcare industry is producing huge amounts of data which need to be mine to discover hidden information for effective prediction, diagnosis, exploration and decision making. Analyzing these huge amounts of data is complex and makes a huge challenge with available traditional methods. As a result of these confusions, Healthcare Information Technology (HIT) has been developed as an interdisciplinary study of the design, development, adaptation, and application of Information Technology (IT) based innovations in health services for management and planning. HIT is a huge area comprising a multitude of components, solutions and technologies. The HIT plays a vital role in terms of improving the quality and effectiveness of healthcare, reducing healthcare costs and paperwork, improves the efficiency of both administrative and clinical processes, increases the accuracy of diagnoses, prevents medical errors, improve patients satisfaction and enabling better health outcomes[1][2]. As well as the benefits of HIT includes the ability to use data analytics and big data for effective management of population health plans and lower the occurrence of expensive chronic health conditions, the ability to share health data among academic researchers to introduce novel medical therapies and drugs, and the privilege of patients to acquire and use their own health data and work together in their own care with clinicians. The HIT can be applied in several health domains which generates huge amount of data like diabetes, heart disease, dengue, cancer and etc.[3]. Data mining is the most widely used technique to transform these vast amounts of data into useful information and patterns for decision making and generate relationships amongst the attributes. Also it can be defined as the method of analyzing data from various perspectives and summarizing it into information that are typically used to increase and enhance the revenue or reduce costs or to provide a new understanding and solution to a problems; especially, in several industries such as healthcare, e-commerce, retail and social media. One of the important applications of data mining techniques that can be used in healthcare sector is anticipating patients future behavior on the given history or symptoms[1].

The data mining provides automatic pattern recognition and attempts to uncover patternsin data that are difficult to detect with traditional statistical methods. Most statistical methods require a hypothesis which is proved or disproved by the data. In data mining however, patterns are automatically recognized and the strength of the associations is evaluated too. As well as it is quite possible that there is more knowledge hidden in the data than what is visible outside. Data mining techniques have a group of various tools and techniques that can be used for different purposes. Data mining goes beyond statistics and uses various algorithms to understand underlying relations between data. However, algorithms ought to be modified and scaled before they are applied to data mining. Data mining involves using different algorithms to do different tasks. Basically, algorithms are used to try to fit to a model closest to the characteristics of data under consideration. Theoretically, available models can be categorized into two as predictive or descriptive. The predictive models are used to make predictions, for example, to make a quick diagnosis of certain element. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. Classification, regression, neural networks and time series analysis are some of the tasks of predictive modeling. Descriptive models are used to identify patterns in data. Clustering rules, association rules, and visualization rules are some are the tasks of descriptive modeling.

The usage of data mining in the healthcare domain is still at a medium stage in Sri Lanka; because the scope of medical data mining is huge. To apply data mining in the medical data that data should be recorded in electronic format. But it is difficult to find properly stored medical data in Sri Lankan hospitals. Most of them are stored in books. So we need to convert these paper based data into an electronic form that can be applying data mining. Applying data mining techniques on such domain can help medical practitioners to predict even the crucial diseases with ease. Currently there is no any mechanism to classify kidney diseases in Sri Lanka. The major objective of this research is predicting the patients who have a risk on kidney disease to move towards treatments. And classify the patients according to the kidney disease type.

The current case study, introduce a novel classification and forecasting methodologyusing data mining techniques for kidney disease in Sri Lanka. Basically there are two types of kidney diseases that can be found in Sri Lanka namely Chronic Kidney Disease (CKD) and Acute Kidney Disease (AKD). CKD is kidney damage and a decline function that lasts more than three months. This is a very serious condition, because patients may not have any symptoms until severe kidney damage, which is incurable, has occurred. Most common symptoms of kidney disease includes swelling of the body, nausea, vomiting, itching of body, fatigue and exhaustion, sleeping problems, metallic taste in mouth, dizziness, breathing difficulty and chest pain, pain in the back and the sides and loss of appetite[4][5][6]. This study is carried under the two phases. In the first phase, the classification model predicts whether a person has a risk in kidney disease or not. As well as it shows risk level like high risky or low risky. This prediction was made by using an Artificial Neural Network (ANN) considering on 11 attributes. In the second phase, model predicts whether a person is fallen in CKD or not. This prediction was made by a model which is a combination of Random Forest algorithm and ANN considering on 30 attributes. Data for this research was collected from special nephrology unit in provincial general hospital in Badulla, Sri Lanka. Dataset contains 1080 instances and consists with 31 attributes. All these attributes represented in numeric or binary format.

ANNs are the subfield of Artificial Intelligence (AI) systems. It is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data. Their ability to correlate input and corresponding output data, based on vector mapping, has established themselves as a powerful tool in various applications. Such as Classification (including pattern and sequence recognition, novelty detection and sequential decision making) and Data processing (including filtering, clustering, blind source separation and compression), Robotics and etc. Also ANNs have been applied in various medical fields, constituting themselves as a useful technique in clinical practice. Medicine is a field that ANNs can be proven as a powerful tool to enhance current medical techniques. Random Forest algorithm is a supervisedclassification algorithm. It enables to classify huge amount of data with an acceptable accuracy. At the training time it forms number of decision trees and outputting the class that is the mode of the classes output by individual trees. So this study is focus on the development of novel ANN model with the combination of Random Forest algorithm for kidney disease classification and prediction.

Major results of this study is a modified ANN with 2 hidden layers to detect kidney disease which gives 80.952% accuracy and 19.047% error rates for testing data. As well as a model with the combination of Random Forest and ANN with 3 hidden layers to predict whether a person is fallen in CKD or not which gives 81.395% accuracy and 18.604% error rate for testing data. The constructed models gives high accuracy and minimum error rate when comparing with the other data mining algorithms. The novel model built with the combination of RF and ANN gives high performance in the CKD prediction instead of using RF and ANN separately. When considering about previous research works related to this domain, they conducted their research works to find the most suitable algorithm that can be applicable for kidney disease prediction. But, models introduce in this study can be reusable and applicable for predict future patients status on kidney disease. So, the proposed solution will save time of both patients and doctors, improve patient satisfaction and keep data for future predictions. So it will be useful for patients, doctors and researchers.

II. Literature Review

In 2008, Palaniappan and Awang used three data mining modelling techniques namely Decision Trees, Nave Bayes and Neural Network to discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database to develop a prototype Intelligent Heart Disease Prediction System [7]. In 2014, Kaur and Chhabra conducted a research work for predicting the diabetes from medical records of patients using modified J48 classifier [8]. Features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. The data mining tool WEKA has been used as an API of MATLAB for generating the J-48 classifiers. Shakil et al. discuss various algorithm approaches of data mining that have been utilized for dengue disease prediction. Such as Nave Bayes, J48, Sequential Minimal Optimization (SMO) and Random tree [9]. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, time taken to build model, mean absolute error and Receiver Operating Characteristic (ROC) area. From the findings of this paper Nave Bayes and J48 are the best performance algorithms for classified accuracy. Kourou et al. present a review of recent machine learning approaches employed in the modelling of cancer progression [10]. Based on the analysis of their results; it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. Among the most common applied machine learning algorithms relevant to the prediction outcomes of cancer patients, they found that Support Vector Machine (SVM) and ANN classifiers were widely used. Kalaiselvi and Nasira discussed data mining techniques that used to diagnose the diseases like heart disease and cancer in diabetic patients [11]. This work consists of two stages. In the first stage, the attributes are identified and extracted using Particle Swarm Optimization (PSO) algorithm. In the second stage, Adaptive Neuro Fuzzy Inference System (ANFIS) with Adaptive Group based K-Nearest Neighbor (AGKNN) algorithm has been used to classify the data. The experiment is done using MATLAB 7.14 and the simulation results are obtainedand grouped into four categories. The experimental results show a very good accuracy and signify the ANFIS with AGKNN along with feature subset selection using PSO. The performance is evaluated using performance metrics and proved this classifiers efficiency for the prediction of heart disease and cancer in diabetic patients.

In 2012, Kumar and Abhishek conducted a study to diagnose kidney stone disease by using three different neural network algorithms namely Learning Vector Quantization (LVQ), two layers feed forward perceptron trained with Back Propagation training algorithm (BPA) and Radial Basis Function (RBF) networks which have different architecture and characteristics [12]. They aim to compare the performance of all three neural networks on the basis of its accuracy, time taken to build model, and training data set size. Finally they come to conclusion that multilayer perceptron trained with 2 hidden layers and back propagation algorithm is the best model for diagnosis of kidney stone disease. Its accuracy is 92% to diagnosis the kidney stone disease. In 2013, Babu et al. reported a method for the diagnosis of polycystic kidney disease (PKD) using ANN [13]. A multi-layered feed forward neural network with one hidden layer is constructed, trained and tested by applying back propagation learning rule for the diagnosis of PKD based on physical symptoms. A neural network is built with 20 input nodes, 10 hidden nodes and one output node. 75% of the data used for training and remaining 25% of the data are used for testing purpose. In the training process weights are adjusted till the target is reached. Once the network is trained then it does the classification automatically for a new pattern. In 2014, Bala and Kumar demonstrated the utility of classification techniques for predicting kidney disease with different data mining tools [14]. Also they stated that number of factors which increase the risk of kidney disease, symptoms of kidney diseases and types of kidney diseases. Outcomes of their research is most commonly used data mining techniques such as Decision Trees, ANN, Nave Bayes, Logistic Regression, Genetic Algorithms resulting as well-performing on medical databases. Also shows that Decision Trees, ANNs and Naive Bayes are the well-performing algorithms used for kidney disease. But at the end they said that it is very difficult to name a single data mining technique as the best for the kidney disease prediction. In 2016, Kumar used the sixmachine learning algorithms, namely: Random Forest classifiers, SMO, Nave Bayes, RBF, Multilayer Perceptron Classifier (MLPC) and Simple Logistic for prediction of CKD [15]. The experiments were conducted for the prediction task of CKD obtained from UCI Machine Learning repository. Prediction task for CKD is performed by separately applying six machine learning algorithms using Weka 3.7.12. The classification performances of the classifiers were analyzed with respect to the standard performance parameters, namely: Accuracy, Specificity, Sensitivity, Precision, ROC Areaand Matthews Correlation Coefficient (MCC) besides time taken for training (learning). The experimental results of this proposed method have demonstrated that Random Forests has produced superior prediction performance in terms of classification accuracy, Area Under Curve (AUC) and MCC respectively for the considered dataset. In 2016, Patil conducted a review of several research papers on prediction of chronic kidney disease using data mining classifiers[16]. She stated that CKD can be very well predicted using many classifiers in data mining. As well as one can also predict the level of CKD using classifiers. According to observation of different experiments; Multilayer Perceptron, Random Forest, Nave Bayes, SVM, KNN and Radial Basis Function are some classifiers which gave highest accuracy. In 2017, Astha et.al has carried out a study to predict and analysis kidney diseases and their treatments based on data mining methods[17]. At last they stated that data mining technique, like classification, is highly efficient in prediction of kidney related diseases. Various classification techniques, like SVM, ANN, Bayesian network, Nave Bayes, Decision Tree, Rule Based, Discriminant Analysis, Clustering technique like k-means, and temporal abstraction techniques are efficient in prediction of different kidney diseases depending on the data.

In Sri Lankan domain there are several research works conducted for identify kidney disease types, risk factors, population of kidney patients and environmental factors that affect to kidney disease. But there is not any kidney disease prediction models were built in Sri Lankan domain. Previous research works and publications on applications of data mining techniques in healthcare have already received a lot of attention in many researchers around the world. Different models and methodologies have been developed for different health domains like diabetes, heartdisease, cancer, dengue, kidney disease and etc. In those models they have used several algorithms like Nave Bayes, Artificial Neural Network (ANN), SVM, Decision Tree, K-Nearest Neighbour (KNN), Random Forest and etc. According to the literature most of the research works are focused on finding a best algorithm or method that can be used in kidney disease classification and prediction and to identify the risk factors for kidney disease, symptoms of kidney disease, types of kidney disease and etc. Not only to find best applicable data mining techniques for disease prediction, it is better to apply them in real scenarios. So in this study introduce a novel hybrid methodology for kidney disease classification and prediction using random forest algorithm and ANN that can be used for real kidney patients.

III. Material And Methods

The current study introduced a novel classification methodology and successfully applied it in Sri Lankan domain for the CKD classifications. It is carried under the two phases. In the first phase, ANN method namely multilayer feed-forward neural network was used to detect whether a person has a risk of having a kidney disease or not and their risk level. In the second phase, a novel forecasting methodology is proposed using multiple algorithms, which is a combination of Random Forest algorithm and an ANN hybrid methodology to detect whether a patient has fallen into a CKD or not.

Dataset: Data for this research was collected from special nephrology unit in provincial general hospital in Badulla, Sri Lanka through a questionnaire. Dataset contains 1080 instances and 31 attributes. Such as Clinic number, Age, Gender, Occupation type(Whether the occupation of a patient is related with agriculture or not), Drinking water source, Smoking, Alcohol usage, Diabetes mellitus, Taken any drugs for a long time (Whether a person taken any drug including ayurvedic medicines for a long time (continuously more than 3 months) for any disease), Hypertension, Snake bites, Family history (Whether a persons any relation has diabetes, kidney disease or hypertension or not), Swelling of body, Fatigue and exhaustion, Nausea, Vomiting, Itching of body, Metallic taste in mouth, Back of flank pain, Shortness of breath and chest pain, Loss of appetite, Dizziness,Sleeping problems, Systolic pressure, Diastolic pressure, Albumin, Creatinine, eGFR (Glomerular Filtration Rate), Sodium, Potassium and CKD. All these attributes represented in numeric or binary format.

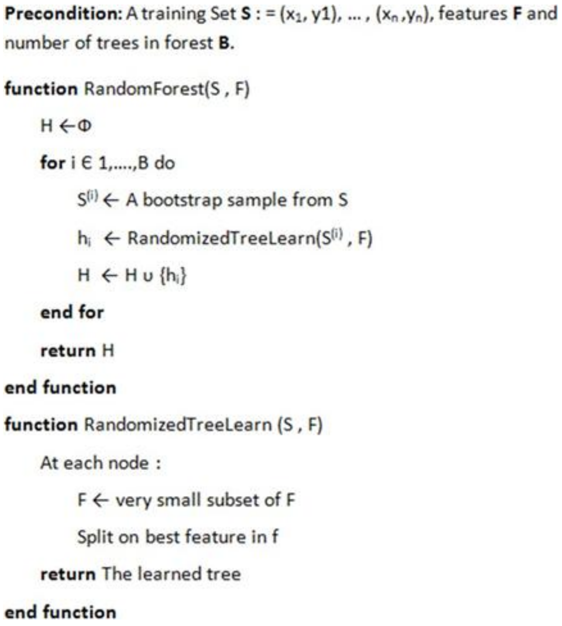
Data pre-processing:Data pre-processing is an important step in the data mining process. Data gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. The product of data pre-processing is the final training set. In this study WEKA data mining tool was used for data pre-processing. Initially data quality was made sure by checking missing, incomplete, outlier values in the dataset.

Building training and testing datasets: When modelling neural networks and other machine learning algorithms first it should be trained. Then the trained models should evaluate. Therefore dataset should divide into two as training and testing. Training data are used to optimize the weights in the neural network and other parameters in the model. Test data are used to evaluate the quality of estimates and forecasts respectively. Test dataset was not used for training models. Test data realistically simulated the model in the case where there was no information about the future. The test data is randomly selected. So that all data had an equal chance to participate in the selection process. In this study, dataset was separated as training dataset and testing dataset in 3 ways as shown in Table no 1.

Table no 1: Sample datasets for algorithm training and testing.

Sample No	Training Dataset	Testing Dataset
1	60%	40%
2	70%	30%
3	80%	20%

Algorithms used: In this study, Random Forest algorithm and Back propagation algorithm were used in the model creation. Random Forestalgorithm is a supervised classification algorithm. It enables to classify huge amount of data with an acceptable accuracy [18]. At the training time it forms number of decision trees and outputting the class that is the mode of the classes output by individual trees [19]. The pseudo code for Random Forest algorithm is shown in Figure no 1.



```
**Precondition:** A training Set **S** : = $(x_1, y1), ... , (x_n ,y_n)$, features **F** and
number of trees in forest **B**.

**function** RandomForest(S , F)
    H ←Φ
    **for** i ∈ 1,....,B do
        $S^{(i)}$ ← A bootstrap sample from S
        $h_i$ ← RandomizedTreeLearn($S^{(i)}$ , F)
        H ← H ∪ {$h_i$}
    **end for**
    **return** H
**end function**
**function** RandomizedTreeLearn (S , F)
    At each node :
        F ← very small subset of F
        Split on best feature in f
    **return** The learned tree
**end function**
```

Figure no 1: Pseudo code for Random Forest algorithm.

ANNs tries to capture the structure and procedure of the human brains problem solving skill and apply them to information systems. Nowadays ANNs have become most widely used tool for diagnosis of disease[2][20]. Because of the fault tolerance, generalization and learning from environment like capabilities of ANN, it is becoming more and more popular in medical diagnosis and many more other areas. The feed-forward neural networks are inspired by the information processing of one or more neural cells which is called as neuron [21]. The fundamental of the Back propagation method is to create a given function by adjusting internal weightings of input signals to compose a desired output signal. The neural network model is trained using a supervised learning method. In here potential outputs of the algorithm are already recognized and the data set used to learn the algorithm is already identified with correct results. Technically, the Back-propagation algorithm is a method for training the weights in a multilayer feed forward neural network. A standard network structure is one input layer, one hidden layer and one output layer as shown in Figure no 2.

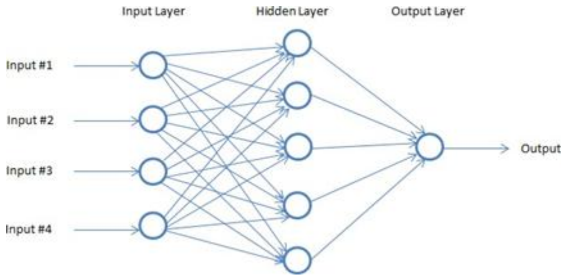
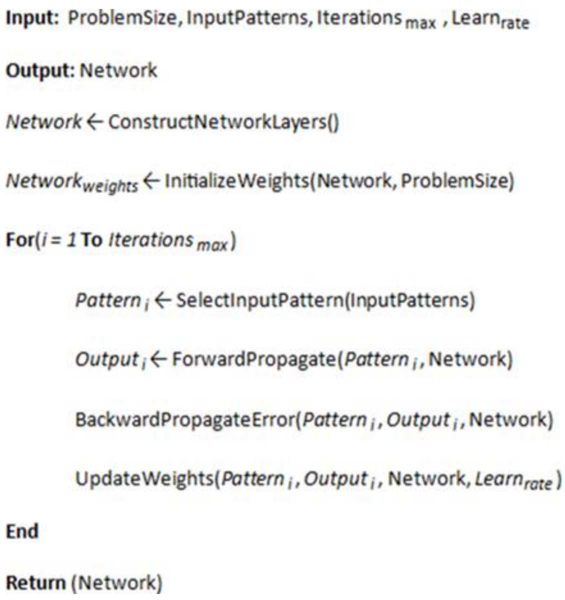


Figure no 2:A standard neural network structure.

First layer is input layer which provides an interface with the environment, second layer is hidden layer where computation is done and last layer is output layer where output is stored. Data is propagated through successive layers, with the final result available at the output layer. Pseudo code for preparing a network using the Back-propagation training method is shown in Figure no 3.



```
Input: ProblemSize, InputPatterns, Iterations_max , Learn_rate
Output: Network
Network ← ConstructNetworkLayers()
Network_weights ← InitializeWeights(Network, ProblemSize)
For(i = 1 To Iterations_max)
        Pattern_i ← SelectInputPattern(InputPatterns)
        Output_i ← ForwardPropagate(Pattern_i, Network)
        BackwardPropagateError(Pattern_i, Output_i, Network)
        UpdateWeights(Pattern_i, Output_i, Network, Learn_rate)
End
Return (Network)
```

Figure no 3: Pseudo code for preparing a network using the Back-Propagation training method.

Proposed model: The current study is carried under the two stages. In the first stage, The Back propagation algorithm which is a supervised learning method for multilayer feed-forward networks in the field of ANNs used to detect whether the person has a risk on having a kidney disease or not. As well as it shows risk level like high risky or low risky. This prediction was made by considering 11 attributes which are symptoms of kidney disease and that can be taken without any medical tests. In the second stage, a novel forecasting methodology is proposed usingmultiple algorithms which is a combination of Random Forest algorithm and an ANN hybrid methodology to detect whether a patient is fallen in CKD or not. Because there is no one data mining method to resolve the issues in the healthcare data sets or make predictions.

In order to obtain the highest accuracy among the classifiers which is important in medical diagnosing with the characteristics of data being taken care, a hybrid model should be designed [3]. As an input data for this model, 30 attributes were used; which is a collection of general data about person, symptoms of kidney disease, results of medical tests and prediction results of Random Forest algorithm. Flowchart of the proposed model is shown in Figure no 4.

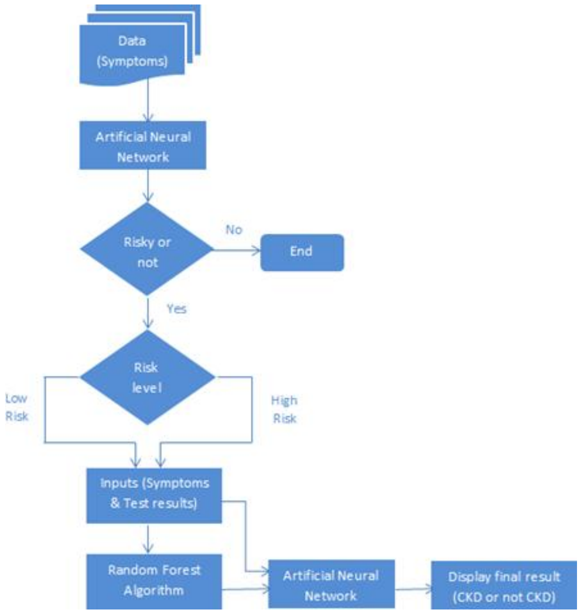

Data (Symptoms)
Artificial Neural Network
Risky or not
No
End
Yes
Risk level
Low Risk
High Risk
Inputs (Symptoms & Test results)
Random Forest Algorithm
Artificial Neural Network
Display final result (CKD or not CKD)

Figure no 4: Flowchart of proposed solution.

In the training and testing the models build in this study, machine learning approach was used. It is a methodology which uses to build mathematical models in order to understand data. Basically it can be divided into 2 phases namely build phase or modeling phase and operational phase. Training dataset is used in build phase. First features and labels should be extracted from the dataset. Then the selected machine learning algorithm should be trained until the model has learned enough. Once it learned, it should be saved. That saved finalized model can be used in the operational phase. Testing dataset is used in this phase. In operational phase you can ask from learned algorithm to explain newly observed data. This process is illustrated inthe Figure no 5.

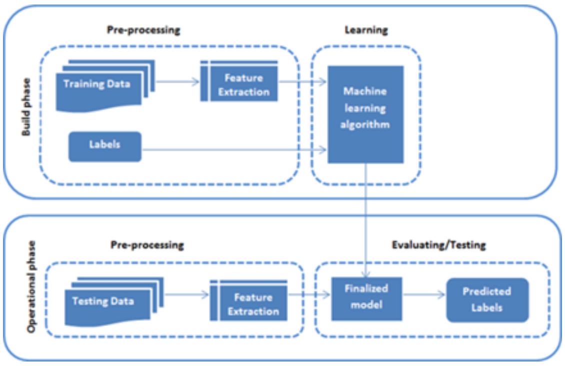


Figure no 5: Machine learning workflow.

In the process of ANN training; number of neurons in hidden layers and an epoch are adjusted until the target (known output) is reached. The training is stopped when the output result is consistence with the original result with least error rate. The output value of the models is in between the range 0.0 to 1.0. If the acquired output value is near to 1.0 then the person is having a risk of kidney disease or the acquired value is near to 0.0 then the person is normal person. The neural networks are trained and tested using three data samples and a neural network model with high performance was selected and saved. Then it can be used to perform the classification automatically for a new pattern.

Machine learning algorithms are driven by parameters. Outcome of learning process of algorithms are highly depend on these parameters. So in here parameter tuning was applied to discover the best value for each parameter to enhance the accuracy of the algorithm or model. By repeating this process with a number of well performing models; optimum model can be selected. When training the Random Forest algorithm adjust the parameter values until best accuracy comes for three data samples. Then the highest accuracy shown model was selected and saved for model the CKD prediction model.

Model evaluation techniques: In machine learning process, performance evaluation is an essential task. A confusion matrix is a specific table layout that allows visualization of the performance of a classification model (or "classifier") on a set of test data for which the true values are known [22]. It is formed from the four outcomes produced as a result of binaryclassification. A binary classifier predicts all data instances of a test dataset as either positive or negative. This classification (or prediction) produces four outcomes namely True positive (TP): correct positive prediction, True negative (TN): correct negative prediction, False positive (FP): incorrect positive prediction and False negative (FN): incorrect negative prediction [15][23]. The columns represent the predictions, and the rows represent the actual class as shown in Table no 2.

Table no 2: Confusion Matrix description

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Some of the confusion matrix-based measures are used to evaluate the performance of the models constructed in this study using testing data. Such as accuracy, recall or sensitivity, precision and F1 score shown in Table no 3.

Table no 3: Performance measurements

Metric	Description	Formula
Accuracy	Number of correct predictions from all predictions made	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
Precision	Positive predictive values	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
Recall (Sensitivity)	Proportion of positive predictions that are correctly identified	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
F1 score	Combination of precision and recall	$\text{F1 score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$

To do further comparisons ROC analysis and Mean Absolute Error (MAE) are used. ROC curve is a commonly used graph that visualizes the performance of a binary classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis). AUC is the best way to summarize its performance in a single number.

The MAE is a quantity used to measure predictions of the eventual outcomes [21]. The mean absolute error is given by below equation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

The MAE is an average of the absolute errors e_i= |f_i – y_i|, where, f_i and y_i represent predictiontrue value respectively. The Table no 4, the scale of judgment of forecast accuracy regarding to MAE indicated that, the minimum values of MAE make more accuracy for forecasting future predictions.

Table no 4: Model accuracy

MAE	Judgment of forecast accuracy
<10%	Highly accurate
10% to 20%	Good forecast
21% to 50%	Reasonable forecast
>51%	Inaccurate forecast

IV. Results

A four layered feed forward neural network is constructed in this stage. It consists of 11 input neurons in input layer which is the physical symptoms of kidney disease, 2 hidden layers and one output in the output layer. ANN was trained by using three data samples. In the training process epochs, number of neurons in first hidden layer and number of neurons in second hidden layer are adjusted till the target is reached. When the output result is matched with the original resultant with minimum error rate and then the training is stopped. The output value is in between the range 0.0 to 1.0. If the obtained output value is near to 1.0 then the patient is having a risk on kidney disease or the obtained value is near to 0.0 then the patient is normal person. Once the network is trained using these samples then it does the classification automatically for new pattern.

In the training process, an ANN with highest accuracy was selected for the prediction. Accuracy of the algorithms varies according to the dataset used. So in here 3 data samples were used. Such as a data sample with 60% for training and 40% for testing, a data sample with 70% for training and 30% for testing and a data sample with 80% for training and 20% for testing.

After a successful training of an ANN in the first phase using 80% data and testing that model using 20% data gives best performance. It gives 80.952% accuracy and 19.047% error rates for testing data. Final model that constructed for make predictions for new data consists 11 input neurons in input layer, 14 neurons in first hidden layer, 9 neurons in second hidden layer and one output neuron in output layer. An epoch is set to 200. Figure no 6 shows ANN structure of the model in first stage.

Model in second stage was built to perform CKD predictions based on physical symptoms, general information (like age, gender) and results obtained from different medical tests. The model was built with a combination of Random Forest and ANN. Machine learning algorithms are driven by parameters.

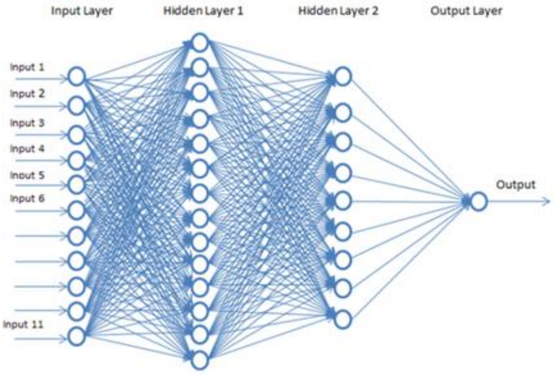


Figure no 6: An ANN structure to predict risk on having a kidney disease in first stage.

These parameters majorly influence the outcome of learning process. Parameters in random forest are either to increase the predictive power of the model or to make it easier to train the model. The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model. It is better to repeat this process with a number of well performing models. Some features tuned in this study to improve the predictive power of the Random Forest model are n_estimators, min_samples_split, max_depth and max_features. n_estimators is the number of trees you want to build before taking the maximum voting or averages of predictions. min_samples_split describes the minimum number of samples required to split an internal node. max_depth defines the maximum depth of the tree. max_features defines the number of features to consider when looking for the best split. The model was trained by adjusting these parameter values until high accuracy comes for 3 training data samples mentioned in Table no 1.

When tuning the parameters to improve the predictive power of the model it is necessary to consider the impact of these attributes to performance and speed of the model. Increasing max_features generally improves the performance of the model. But decrease the speed of algorithm by increasing the max_features. So it is best to strike the right balance and choose the optimal max_features in training process. When considering on n_estimators, higher number of trees give better performance but makes code slower. So it is better to choose as high value as the processor can handle because this makes your predictions stronger and more stable. The best way is trying with multiple parameter values to find the most optimum for the case. A model with n_estimators = 45, min_samples_split = 25, max_depth =10 and max_features =29 shows high performance when comparing with other models used in testing. So that model was selected and saved for future prediction.

In the second phase, a four layered feed forward neural network with 30 input neurons in input layer, 2 hidden layers and one output in the output layer was trained by using three data samples. In the training process epochs, number of neurons in first hidden layer and number of neurons in second hidden layer are adjusted till the target is reached. So the model was trained and tested for 3 hidden layers by adjusting the neurons in each layer and epoch for 3 data samples. A model with the combination of Random Forest and ANN including 30 input neurons in input layer, 10 neurons in first hidden layer, 9 neurons in second hidden layer, 6 neurons in third hidden layer and one output neuron in output layer was constructed for CKD prediction. It gives 81.395% accuracy and 18.604% MAE for testing data.

Figure no 7 shows performance of CKD prediction of Random Forest algorithm, ANN and novel model that built in the phase 2.

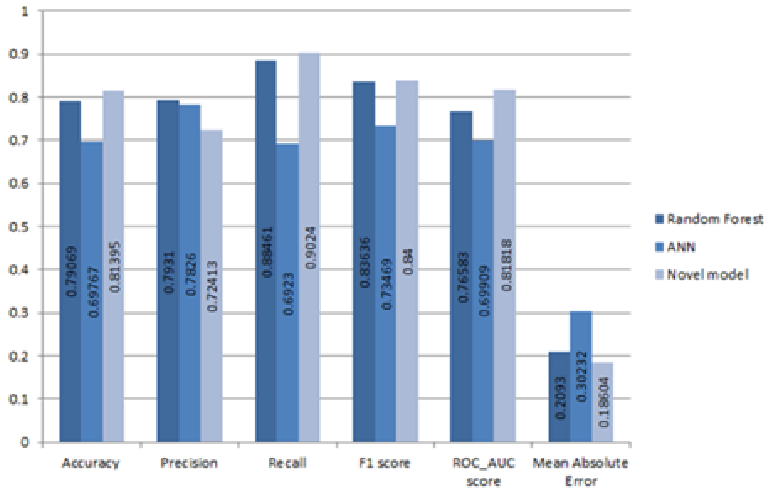


Figure no 7: Performance comparison of algorithms.

The results clearly shows the novel model built with the combination of Random Forest and ANN gives high performance in the CKD prediction instead of using Random Forest and ANN separately for CKD prediction.

Figure no 8 shows the ROC curve for the model built in the first phase for testing data. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect model; an area of 0.5 represents a poor model. According to the Fig.8 risk prediction model gives high performance for the data sample which is 80% and 20%. AUC of this is 0.78.

Figure no 9 shows the ROC curve for the model built in the second phase. According to the Figure no 9 CKD prediction model gives high performance for the data sample which is 60% and 40% for testing. AUC value is 0.82.

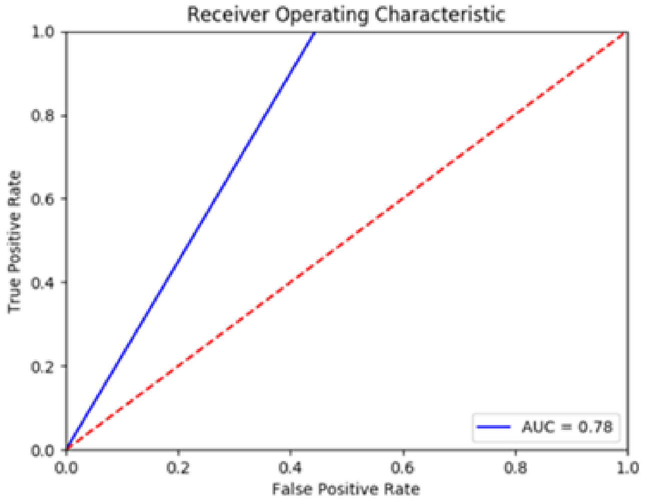


Figure no 8:AUC for model in the first phase for test data.

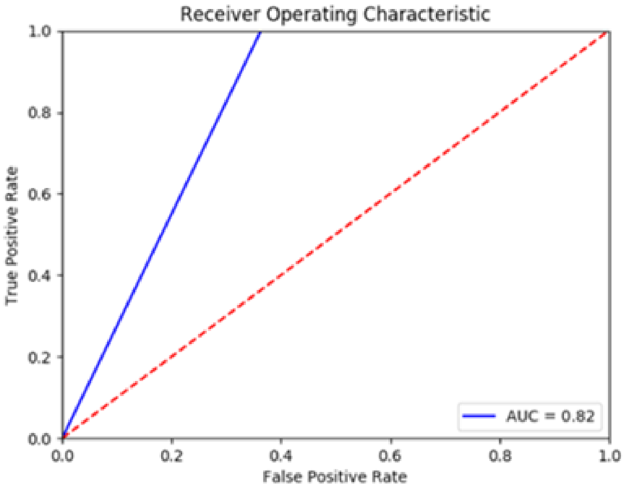


Figure no 9:AUC for model in the second phase for test data.

V. Conclusion

Healthcare sector is producing massive data. It includes patient centric data, resource management and many more. Healthcare organizations must have ability to analyse these data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in decision making, forecasting and answering several important and critical questions related to healthcare. Also clinical decision support and computer based patient records can be integrate by using data modelling and analysing tools. It will significantly improve the quality of clinical decisions including reduce medical errors, enhance patient safety and enhance patient outcome. ANNs have been used in different medical fields and constitute useful techniques in clinical practice. Also they are frequently used as a strong discriminating classifier for tasks in medical diagnosis for early detection of diseases. So in this study, introduces a novel model for kidney disease classification and prediction using Random Forest algorithm and ANN. It is carried under two phases. In the first stage an ANN with one input layer, two hidden layers and one output layer was constructed to detect whether a person has a risk on having a kidney disease or not. In the second stage, a model was built with the combination of RF and ANN with one input layer, three hidden layers and oneoutput layer to predict whether a person is fallen in CKD or not CKD.

Acknowledgement

The authors like to acknowledge consultant nephrologist and all the staff members of nephrology unit in Provincial General Hospital, Badulla, Sri Lanka.

References

[1]. Srinivas, K., Rani, B.K. and Govrdhan, A., "Applications of data mining techniques in healthcare and prediction of heart attacks.," International Journal on Computer Science and Engineering, vol. 2, no. 02, pp. 250-255, 2010.

[2]. Kraft, M.R., Desouza, K.C. and Androwich, I.,, "Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population," in 36th Annual Hawaii International Conference on System Sciences, 2003.

[3]. Jothi N, Husain W, "Data mining in healthcare–a review.," in Procedia Computer Science, 2015.

[4]. Abdel-Kader K, Unruh ML, Weisbord SD, "Symptom burden, depression, and quality of life in chronic and end-stage kidney disease.," Clinical Journal of the American Society of Nephrology, vol. 4, no. 6, pp. 1057-1064, 2009.

[5]. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS, "A predictive model for progression of chronic kidney disease to kidney failure," Jama, vol. 305, no. 15, pp. 1553-1559, 2011.

[6]. Fisher MA, Taylor GW, "A prediction model for chronic kidney disease includes periodontal disease.," Journal of periodontology, vol. 80, no. 1, pp. 16-23, 2009.

[7]. Palaniappan S, Awang R, "Intelligent heart disease prediction system using data mining techniques.," in IEEE/ACS international conference on computer systems and applications, 2008.

[8]. Kaur G, Chhabra A, "Improved J48 classification algorithm for the prediction of diabetes.," International Journal of Computer Applications, vol. 98, no. 22, 2014.

[9]. Shakil KA, Anis S, Alam M, "Dengue disease prediction using weka data mining tool," arXiv preprint, 2015.

[10]. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI, "Machine learning applications in cancer prognosis and prediction.," Computational and structural biotechnology journal, pp. 8-17, 2015.

[11]. Kalaiselvi C, Nasira G, "Prediction of heart diseases and cancer in diabetic patients using data mining techniques.," Indian Journal of Science and Technology, vol. 8, no. 14, 2015.

[12]. Kumar K, Abhishek B, "Artificial neural networks for diagnosis of kidney stones disease.," I.J. Information Technology and Computer Science, 2012.

[13]. Babu A, Sumana G, Rajasekhar M, "Computer Aided Diagnosis of Polycystic Kidney Disease Using ANN.," World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, vol. 7, no. 12, pp. 933-937, 2013.

[14]. Bala S, Kumar K, "A literature review on kidney disease prediction using data mining classification technique.," International Journal of Computer Science and Mobile Computing, vol. 3, no. 7, pp. 960-967, 2014.

[15]. K. M, "Prediction of chronic kidney disease using random forest machine learning algorithm.," International Journal of Computer Science and Mobile Computing, vol. 5, no. 2, pp. 24-33, 2016.

[16]. P. PM, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques.," International Journal of Computer Science and Mobile Computing, vol. 5, no. 5, p. 135, 2016.

[17]. J. MAAMK, "Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review.," International Journal Of Engineering And Computer Science, vol. 6, no. 2, 2017.

[18]. Abdullah AS, Rajalaxmi R, "A data mining model for predicting the coronary heart disease using random forest classifier," in International Conference in Recent Trends in Computational Methods, Communication and Controls, 2012.

[19]. Khalilia M, Chakraborty S, Popescu M, "Predicting disease risks from highly imbalanced data using random forest," BMC medical informatics and decision making , vol. 11, no. 1, 2011.

[20]. B. J, "Clever algorithms," in Nature-Inspired Programming Recipes, 2011, p. 436.

[21]. Rathnayaka RKT, Seneviratna D, Jianguo W, Arumawadu HI, "A hybrid statistical approach for stock market forecasting based on Artificial Neural Network and ARIMA time series models," in International Conference on Behavioral, Economic and Socio-cultural Computing, 2015.

[22]. Chatterjee S, Dzitac S, Sen S, Rohatinovici NC, Dey N, Ashour AS, Balas VE, "Hybrid modified Cuckoo Search-Neural Network in chronic kidney disease classification.," in 14th International Conference on Engineering of Modern Electric Systems, 2017.

[23]. Chang C-D, Wang C-C, Jiang BC, "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors," Expert systems with applications, vol. 38, no. 5, pp. 5507-5513, 2011.

K.A.A. Chathurangi. "Introducing novel classification methodology for health science; A case study in kidney disease in Sri Lanka." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22(4), 2020, pp. 19-30.