ARTIFICIAL NEURAL NETWORK BASED NEW CLASSIFICATION METHODOLOGY FOR IDENTIFYING KIDNEY DISEASE RISK LEVELS

KAA Chathurangi[1] and RMKT Rathnayaka[2]

[1]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

[1]*ayeshachathurangi1110@gmail.com*

Abstract- The healthcare sector has vast amount of medical data which are still not properly analysed; 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 Chronic Kidney Disease (CKD) classifications. The current 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.

Keywords- Artificial Neural Networks, Data Mining, Random Forest

I. INTRODUCTION

The healthcare industry is producing huge amounts of data which need to be mined to discover hidden information for effective prediction, diagnosis, exploration and decision making. Analysing 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 patient satisfaction and enabling better health outcomes.

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 like diabetes, heart disease, dengue, cancer and etc. Through these areas, huge amount of data are generated. The data mining techniques provide the methods and technology to generate useful information and patterns in large data sets for decision making and generate relationships amongst the attributes. Also it can be defined as the method of analysing 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 e-commerce, retail and social media.

In this case study, introduce a new classification methodology for kidney disease in Sri Lanka. As an initial step, 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. Then the model predicts whether a person is fallen in CKD or not.

Kidneys in human body play an important role, with various functions that are critical to life. The main job of kidney is to filter and clean our blood. Kidney disease is an increasingly serious problem. When our kidneys are incapable to accomplish their functions properly; it may cause to occur kidney disease. Basically there are two types of kidney diseases that can be found in Sri Lanka.

A. Types of Kidney Disease

Chronic Kidney Disease (CKD) is kidney damage and a decline function that lasts more than three months. When someone is sick, injured kidney function, rapid changes in the cause, or taking certain medications, this is called Acute Kidney Disease (AKD). This can happen in ordinary kidneys or people who already have kidney problems.

B. Symptoms of Kidney Disease

Most common symptoms of kidney disease includes swelling of the body, specially noted in the face, feet, legs, hands and ankles, changes in urine output, frequency and colour, 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.

The rest of the paper is organized as follows. Section 2 explains about related works under this topic. Section 3 explains about the methodology of the research including technologies and techniques used in this study. Section 4 explains about the experimental design. Section 5 explains about experimental results. Section 6 is discussion and Section 7 ends up with conclusion.

II. RELATED WORKS

When considering about previous research publications it clearly shows this topic has already received a lot of attention in many researchers around the world. Different models and methodologies have been developed which is in related to this subject. In those models they have used several algorithms like Naïve Bayes, ANN, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbour (KNN), Random Forest and etc. 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.

III. METHODOLOGY

The current study is carried out under 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 Artificial Neural Networks used to detect whether the person has a risk on having a kidney disease or not. Also 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. The feed-forward neural networks are inspired by the information processing of one or more neural cells which is called as neuron. 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 (Arumawadu et.al, 2015; Arumawadu et.al, 2016).

In the second stage, 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 is fallen in CKD or not (Rathnayaka et.al, 2012). 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. Random Forest algorithm is a supervised classification algorithm (Rathnayaka et.al, 2014; Rathnayaka et.al, 2015).It is able 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.

A. Dataset Used

Data for this research was collected from special nephrology unit in provincial general hospital in Badulla. Dataset contains 108 instances and consists with 31 attributes. All these attributes represented in numeric or binary format.

B. Building Training and Testing Datasets

When modelling neural networks and other machine learning algorithms first it should be train. 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 1.

Table 1. Sample datasets for algorithm training and testing

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

C. Evaluate the Novel Model

In machine learning process, performance evaluation is an essential task. 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, F1 score Receiver Operating Characteristic (ROC) analysis and Mean Absolute Error (MAE).

IV. PEXPERIMENTAL DESIGN

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 modelling 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 in the Figure 1.

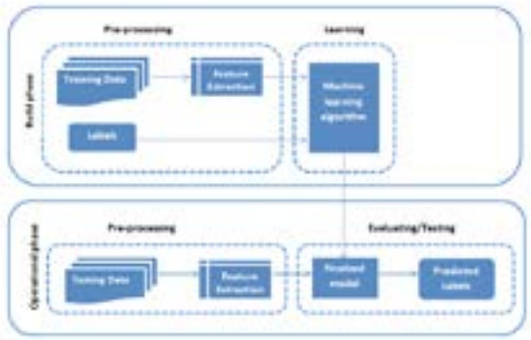

Figure 1. Machine learning workflow

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 on 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 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.

V. RESULTS

In the first phase, after a successful training of an ANN using 80% training data and testing that model using 20% testing data gives best performance. It gives 0.80952 accuracy and 0.19047 error rates for testing data. Final neural network 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. This model can be used to detect kidney disease and risk level. Figure 2 shows ANN structure of the model in first stage.

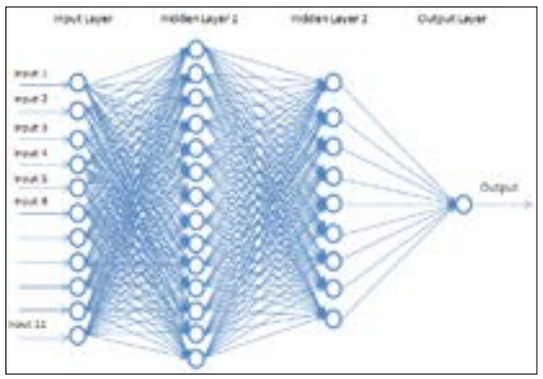

Figure 2. ANN structure for the model in first phase

In the second phase, 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 0.81395 accuracy and 0.18604 error rate for testing data.

Figure 3 shows the performance comparison of model in phase 2 with other algorithms.

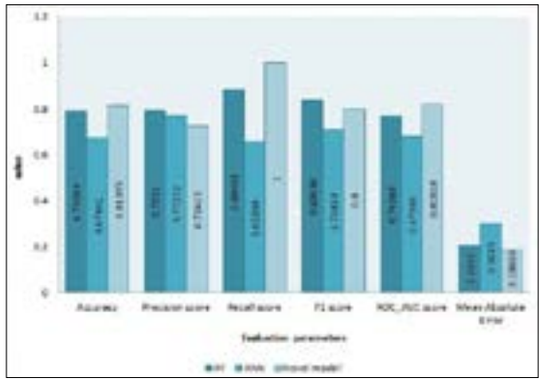

Figure 3. Performance comparison of algorithms

VI. DISCUSSION

This study introduces novel models for kidney disease prediction instead of finding a best or suitable algorithm that can be used to kidney disease classification and prediction. Figure 3 shows performance of CKD prediction of Random Forest algorithm, ANN and novel model built in the phase 2. The results clearly show 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.

VII. CONCLUSION

ANNs have been used in different medical fields and constitute useful techniques in clinical practice. This study introduces a novel model for kidney disease classification and prediction using Random Forest algorithm and ANN. Instead of using one method or one algorithm, the novel model was built by combining Random Forest and ANN. It gives high performance when comparing with algorithms separately. Artificial Neural Networks are frequently used as a strong discriminating classifier for tasks in medical diagnosis for early detection of diseases.

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 model was built by with the combination of RF and ANN with one input layer, three hidden layers and one output layer to predict whether a person is fallen in CKD or not CKD.

According to the health reports population of kidney patients in Sri Lanka is very high. But facilities required to treat them is very low. So this system will help to make the treatment process efficient. Sometimes persons want to know whether they are fallen into kidney disease or not, even though they have not fallen into kidney disease actually. So this system is very useful to them. Also this system is useful to kidney patients in initial stage. Also it reduces the cost and time.

ACKNOWLEDGEMENT

Success of this research depends on the encouragement and guide that I got from the many others. I would like to thankful Consultant Nephrologist and all the staff members of Nephrology unit in Provincial General Hospital, Badulla.

REFERENCES

Arumawadu, H.I., Rathnayaka, R.K.T. and Illangarathne, S.K., 2015. K-Means Clustering For Segment Web Search Results. *International Journal of Engineering Works*, 2(8), pp.79-83.

Arumawadu, H.I., Rathnayaka, R.K.T. and Illangarathne, S.K., 2015. Mining profitability of telecommunication customers using k-means clustering. *Journal of Data Analysis and Information Processing*, 3(3), pp.63-71.

Arumawadu, H.I., Tharanga, R.K. and Seneviratna, D.M.K.N., 2016. New Proposed Mobile Telecommunication Customer Call Center Roster Scheduling Under the Graph Coloring Approach. *International Journal of Computer Applications Technology and Research*, 5(4), pp.234-237.

Babu, A., Sumana, G. and Rajasekhar, M., 2013. 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,* 7(12), pp.933-937.

Bala, S. and Kumar, K., 2014. A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing,* 3(7), pp.960-967.

Celik, E., Atalay, M. and Kondiloglu, A., 2016. The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods. *International Journal of Intelligent Systems and Applications in Engineering,* 4(Special Issue-1), pp.27-31.

Gunatilake, S., Samaratunga, S. and Rubasinghe, R., 2015. Chronic Kidney Disease (CKD) in Sri Lanka-Current Research Evidence Justification: A Review. *Sabaragamuwa University Journal,* 13(2).

Jain, M.A.A.M.K., 2017. Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review. *International Journal Of Engineering And Computer Science*, 6(2).

Kalaiselvi, C. and Nasira, G.M., 2015. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. *Indian Journal of Science and Technology*, 8(14).

Kaur, G. and Chhabra, A., 2014. Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*, 98(22).

Khan, S.H., 2010. Predictive models for chronic renal disease using decision trees, naïve bayes and case-based methods.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, pp.8-17.

Kumar, K. and Abhishek, B., 2012. Artificial neural networks for diagnosis of kidney stones disease.

Patil, P.M., 2016. Review on Prediction of Chronic Kidney Disease using Data Mining Techniques. - *International Journal of Computer Science and Mobile Computing*, 5(5), pp.135-141.

Rathnayaka, R.K.T., Jianguo, W. and Seneviratna, D.N., 2014, October. Geometric Brownian Motion with Ito's lemma approach to evaluate market fluctuations: A case study on Colombo Stock Exchange. In Behavior, *Economic and Social Computing (BESC), 2014 International Conference* on (pp. 1-6). IEEE.

Rathnayaka, R.K.T., Seneviratna, D.M.K.N., Jianguo, W. and Arumawadu, H.I., 2015, October. A hybrid statistical approach for stock market forecasting based on Artificial Neural Network and ARIMA time series models. In Behavioral, *Economic and Socio-cultural Computing (BESC), 2015 International Conference on* (pp. 54-60). IEEE.

Ratnayaka, R.K.T., Wang, Z.J., Anamalamudi, S. and Cheng, S., 2012. Enhanced greedy optimization algorithm with data warehousing for automated nurse scheduling system. *E-Health Telecommunication Systems and Networks*, 1(04), p.43.

Saumya, T.M.D., Rupasinghe, T. and Abeysinghe, P., 2014. A Literature Review in Data Mining Models Used for Survivability Prediction of Cancer Patients.

Shakil, K.A., Anis, S. and Alam, M., 2015. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv*:1502.05167.

A MACHINE LEARNING BASED SOLUTION FOR FINDING PERFECT MARITAL PARTNER

BKTP Wickramasinghe[1], DU Vidanagama, and N Wedasinghe
Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
[1]*wickramasinghe95@gmail.com*

Abstract- Marriage is a socially or ceremonially perceived joining between mates that sets up rights and commitments between those life partners. Finding a good marriage partner is one of the main reasons for the delay in marriage. Therefore, there is a need for a solution that can get user details and expected partner preferences and suggest proper matches based on their preferences. The objective of this paper is to discuss the necessity of the proposed model for Sri Lankans. The proposed solution will maintain user details and get user's preferences for their matrimonial partner. Based on the preferences, appropriate matches will be displayed to the user using a clustering algorithm along with the matching percentages. Horoscope of the user will be generated based on the planet details of the user. Furthermore, previous birth connection and 'dosha' identification will also be done. The proposed solution will also enable of sending messages to the matches and get email notifications about those matches. The final aim of this solution is to ease the matchmaking business by providing proper matches.

Keywords- Matrimonial Partner, Horoscope, Clustering, Matches

I. INTRODUCTION

Marriage is the union of two people. The definition of marriage varies around the world not only between cultures and between religions, but also throughout the history of any given culture and religion. A marriage ceremony is known as a wedding. In Sri Lanka, during the present century female age at marriage has increased almost by seven years. The delay in marriage has an enormous impact on the birth rate. Unlike in the west, where marriage is not necessarily the prosecutor of childbearing or the responsibility of running a household, in Sri Lanka procreation is almost entirely within the marriage. The United Nations World Fertility Report of 2003 reports that 89% of all people get married before the age of forty-nine. The percent of women and men who marry before the age of forty-nine drops to nearly 50% in some nations and reaches near 100% in other nations. (Un.org,2000) Finding a good marriage partner is one of the main reasons for the delay in marriage. In early times, people were not allowed to fall in love and get married. They had to find their matrimonial partner using traditional matchmaker approach ("Kapuwa") or Matrimonial advertisements. Even though those approaches were famous, there were some disadvantages. According to (Vreede-de Stuers,1969), the matrimonial advertisements that appear in English-language newspapers have attracted scholarly attention in recent years, and quite justifiably, for these items provide an abundance of information admirably adapted to statistical analysis of some of the variables determining mate selection. Yet the limitations of this material are also obvious. Verifying the accuracy of the contents is the main limitation. For instance, the beauty of a girl or the earnings of a boy can be exaggerated, and that may be misleading those who search the columns for an attractive candidate. Furthermore, the marriages happened based on these advertisements have led to divorce in most of the cases. With the advent of the Internet, a new channel in the form of matrimonial Web sites has emerged as an alternative way to find partners for marriageable members of the family. The introduction of technology in the form of matrimonial Web sites in an otherwise socially-enabled process provides the setting for a fascinating exploration of changing social mores and the interaction of technology and society. (Patnayakuni and Seth,2008)

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 model was built by with the combination of RF and ANN with one input layer, three hidden layers and one output layer to predict whether a person is fallen in CKD or not CKD.

According to the health reports population of kidney patients in Sri Lanka is very high. But facilities required to treat them is very low. So this system will help to make the treatment process efficient. Sometimes persons want to know whether they are fallen into kidney disease or not, even though they have not fallen into kidney disease actually. So this system is very useful to them. Also this system is useful to kidney patients in initial stage. Also it reduces the cost and time.

ACKNOWLEDGEMENT

Success of this research depends on the encouragement and guide that I got from the many others. I would like to thankful Consultant Nephrologist and all the staff members of Nephrology unit in Provincial General Hospital, Badulla.

REFERENCES

Arumawadu, H.I., Rathnayaka, R.K.T. and Illangarathne, S.K., 2015. K-Means Clustering For Segment Web Search Results. *International Journal of Engineering Works*, 2(8), pp.79-83.

Arumawadu, H.I., Rathnayaka, R.K.T. and Illangarathne, S.K., 2015. Mining profitability of telecommunication customers using k-means clustering. *Journal of Data Analysis and Information Processing*, 3(3), pp.63-71.

Arumawadu, H.I., Tharanga, R.K. and Seneviratna, D.M.K.N., 2016. New Proposed Mobile Telecommunication Customer Call Center Roster Scheduling Under the Graph Coloring Approach. *International Journal of Computer Applications Technology and Research*, 5(4), pp.234-237.

Babu, A., Sumana, G. and Rajasekhar, M., 2013. 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,* 7(12), pp.933-937.

Bala, S. and Kumar, K., 2014. A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing,* 3(7), pp.960-967.

Celik, E., Atalay, M. and Kondiloglu, A., 2016. The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods. *International Journal of Intelligent Systems and Applications in Engineering,* 4(Special Issue-1), pp.27-31.

Gunatilake, S., Samaratunga, S. and Rubasinghe, R., 2015. Chronic Kidney Disease (CKD) in Sri Lanka-Current Research Evidence Justification: A Review. *Sabaragamuwa University Journal,* 13(2).

Jain, M.A.A.M.K., 2017. Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review. *International Journal Of Engineering And Computer Science*, 6(2).

Kalaiselvi, C. and Nasira, G.M., 2015. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. *Indian Journal of Science and Technology*, 8(14).

Kaur, G. and Chhabra, A., 2014. Improved J48 classification algorithm for the prediction of diabetes. *International Journal of Computer Applications*, 98(22).

Khan, S.H., 2010. Predictive models for chronic renal disease using decision trees, naïve bayes and case-based methods.

Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, pp.8-17.

Kumar, K. and Abhishek, B., 2012. Artificial neural networks for diagnosis of kidney stones disease.

Patil, P.M., 2016. Review on Prediction of Chronic Kidney Disease using Data Mining Techniques. - *International Journal of Computer Science and Mobile Computing*, 5(5), pp.135-141.

Rathnayaka, R.K.T., Jianguo, W. and Seneviratna, D.N., 2014, October. Geometric Brownian Motion with Ito's lemma approach to evaluate market fluctuations: A case study on Colombo Stock Exchange. In Behavior, *Economic and Social Computing (BESC), 2014 International Conference* on (pp. 1-6). IEEE.

Rathnayaka, R.K.T., Seneviratna, D.M.K.N., Jianguo, W. and Arumawadu, H.I., 2015, October. A hybrid statistical approach for stock market forecasting based on Artificial Neural Network and ARIMA time series models. In Behavioral, *Economic and Socio-cultural Computing (BESC), 2015 International Conference on* (pp. 54-60). IEEE.

Ratnayaka, R.K.T., Wang, Z.J., Anamalamudi, S. and Cheng, S., 2012. Enhanced greedy optimization algorithm with data warehousing for automated nurse scheduling system. *E-Health Telecommunication Systems and Networks*, 1(04), p.43.

Saumya, T.M.D., Rupasinghe, T. and Abeysinghe, P., 2014. A Literature Review in Data Mining Models Used for Survivability Prediction of Cancer Patients.

Shakil, K.A., Anis, S. and Alam, M., 2015. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv*:1502.05167.