Survival Time Analysis of Liver Cancer Patients using Machine Learning

Md. Ferdous, Md. Monowar Hossain, F.M. Rahat Hasan Robi

Abstract— Survival time analysis is a great factor in biomedical fields. In the medical sector, Machine Learning (ML) is extensively used for cancer research. Liver Cancer is one of the leading cause of death in Bangladesh. Survival time prediction might be an urgent concern for both patients and doctors mainly when a patient goes to the difficult spotlight. In this paper, we try to show patients survival time using their external body condition. This analysis can also be done using statistics technique, but inadequate when dealing with complex and highly nonlinear data.We will match the predict output with the real data.

Index Terms— Artificial Neural Networks, Backpropagation Algorithm, Liver Cancer, Medical Diagnosis, Prediction.

1 Introduction

The use of ML is increasing day by day. We can use it in the field of science and technology. In the field of medical and biological research, ML is popular than the previous year. The popularity of the use of ML in the medical sector for disease detection and prediction tremendously high. Cancer is a deadly disease all over the world. Cancer patients are increasing day by day and the risk of death is also high. Bangladesh Bureau of Statistics described cancer as the six leading cause of death. Liver cancer is a serious disease among many types of cancer. Many patients die from this disease. Survival prediction and classification have been used in biomedical fields. In the medical sector, the ML is extensively used for cancer research. Artificial neural networks (ANN) a remarkably popular class of Computational Intelligence (CI) models have been extensively applied to various predicting problems because they are very generic, accurate and useful mathematical patterns able to efficiently simulate numerical design elements. The dilemma of nonlinearities a comprehensive dataset can be solved smoothly through ANN. The major participation of this thesis are outlined as follows: to conduct this work we choose a dataset from the real world which is taken by the National Institute of Cancer Research and Hospital (NICRH). Our dataset contains 1200(both male and female) patients' data. We choose a supervised machine learning algorithm in this context. The ANN is used which is trained by levenberg marquardt backpropagation algorithm. Utilizing this algorithm we try to predict patients' survival time. From Fuzzy Network (FN) we can determine level of cancer. We use K-means clustering for bunching patients. To do this we break the total dataset into three isolated parts whereas the total error rate can be minimized. Three parts are training part, testing part and validation part. We use such type of model for the purpose of the use of ML for helping patients and doctors to a better understanding of risk analysis for decision making.

The dissertation is organized as follows: In Section 2 this chapter describe some related works, while Section 3 shows a model which we use as our system accomplishment. Implementation of our proposed technique has been described in Section 4. Section 5 comprise experiments and evaluation. From which we can see the utility of our proposed system. Concluding remark and future work are narrated in Section 6.

2 Related Work

M.-H. Zheng et al. [1] build an artificial neural network to predict 3- month mortality risk. For this, they took liver failure patients who are attacked with hepatitis B. Chen-Chiang Lin et al. [2] choose two algorithm Artificial Neural Network (ANN) and Logistic Regression to vindicate a predictive model for hip fractured patients. They discover the ANN has a greater predictive capacity. Filippo Amato et al. [3] use artificial neural networks as artificial intelligence methods to avoid misdiagnosis in the diagnosis process.D.Shanthi et al. [4] shows the application of Artificial Neural Networks (ANN) for stroke disease prediction. They exhibit ANN gives better accuracy to predict stroke disease.Alessandro cucchetti et al. [5] try to shows traditional linear model is less accurate than preoperative variables. This variable can be used in the diagnostic center.They also found that ANN correctly identify tumour grades.Andrue IR Mass et al. [6] use a classification of patients according to traumatic brain injury(TBI).They use morphological irregularities of the CT scan.Martin T. Hagan et al. [7] focused on training feedforward networks with the Marquardt algorithm and they found that the Marquardt algorithm is much more efficient than either of the other techniques.Dursun Delen et al. [8] select three different types of algorithm one is Artificial Neural Network and the other is Dession tree.They also use logistic regression to predict the model. In this regard, they found the decision tree is better than the other two algorithms for breast cancer survivability prediction.Farid E Ahmed [9] found Artificial Neural Networks have exalted accura-

- *Md. Ferdous, Dept. of CSE, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Goplalganj, Bangladesh*
- *Md. Monowar Hossain, Lecturer, Department of CSE, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh, murad0904045@gmail.com*
- *F.M. Rahat Hasan Robi, Lecturer, Department of CSE, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh, rahatcse10@gmail.com*

cy for classification and survival prediction of colon cancer patients than clinicopathological ways and any other statistical methods.Andrew Hunter et al. [10] apply the ANN to the inspection survival problem. They also do sensitivity analysis.Ananya Das et al. [11] try to examine critical lowergastrointestinal using ANN. They also predict clinically consequence with the disorder of patients.

3 METHODOLOGY

Figure 1 sketches the system which we governance as our system implementation. This is the representation of the neck and crop conception of how the prediction, classification and determination process will appear. At first, through the medical institution we assemble our required data.After getting the data we process it for the data mining tool. In this illustration, we apply python programming for algorithm implementation , tensorflow for visualization and MATLAB tool which is a fourthgeneration multi-paradigm programming language and data mining tool. Utilizing this tool we adjusted the data into ANN, k-means Clustering and Fuzzy Network. From unsupervised learning K-means Clustering we determine a patients category. We also determine the level of cancer i.e high-level (upto 70%), mid-level(40% to 60%), low-level(bellow 40%) using FN. Then we sketch a supervised neural network using tensorflow library which trains the data and extorts information from the data. We also use Levenberg- Marquardt backpropagation algorithm to train the neural network. Using this method, we predict how long (in month) the patient can survive attacked with liver Cancer and then we attempt to obtain the accuracy rate comparing with original data. Each component of our proposed system is illustrated in the bellowing subdivision.

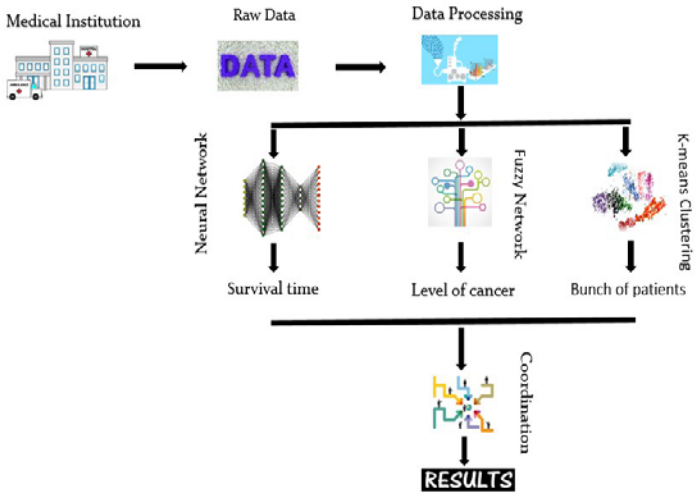

Figure 1: System Overview

3.1 Data Collection

We try to gather qualitative data from National Institute of Cancer Research and Hospital (NICRH). It is a difficult job to obtain qualitative data from the real world. Here we use 1200 patients data for the research. In this study, we examine eleven factors are Yellowing of the skin(Generally, the skin and eyes of the person suffering from liver cancer have a yellow color. It has been trying to measure how much amount of yellow color in his skin and eye).Dark urine(We tried to see here that the patients urine with more black or middle black or black color. Because the color of his urine exposes his physical condition).Itching(It may be due to various causes of itch in the human body, but it is exposed to the physical condition of the person suffering from liver cancer so that we have tried to see here that the level of itching in the patients body is less or more moderate).Jaundic(Generally the first condition of lever cancer is the condition of jaundice that has been affected by jaundice.Therefore, we have taken this parameter for how badly the patient has a jaundice effect).Amnesia(Patients with cancer often go to coma. Usually, more affected patients go to this coma, which means that those who go to the quake are more susceptible to liver cancer or the last stage of liver cancer).Abdominal pain(For the final stage of cancer, the amount of abdominal pain generally high, it is very important to have a symptom because the poorly affected patients are more likely to have abdominal pain).Darker stools(Dark stools is one of the most common symptoms of liver cancer affected patients it tells the internal issues of patients and so on.TABLE I and TABLE II shows the the parameters which we use for our work. We will try to collect this type of data from National Institute of Cancer Research and Hospital (NICRH).

TABLE I: Input and output data transformation

S/N	Factors	Domain
1	Yellowing of the skin	1.Less 2.Medium 3.More
2	Dark urine	1.Medium 2.More 3.Partial
3	Itching	1.Less 2.Medium 3.More
4	Jaundic	1.Partial 2.Medium 3. Highly
5	Amnesia	1.Less 2. Partial 3.More
6	Coma	1. Yes 2. No
7	Abdominal pain	1. Partial 2. Medium 3.Highly
8	Darker stools	1. Yes 2. No
9	Type of person they are?	1. Emotional 2. As usual 3. Reality
10	Age	Classified
11	Admitted date	Classified
12	Death certificates date	Classified

TABLE II: Target data transformation

S/N	Factors	Domain
1	Difference between Admitted date and Death certificates date	1.Month

3.2 Data Processing

Data processing is the way to create a meaningful information from gathered data. Consequently, we demand to separate out

some irrelevant data so that we have a correct date for Neural Network training, testing, and validation. We split the eleven factors into two portions. one portion contains ten factors as the input section amongst from eleven factors. The other portion consists of one factor as out output/target variable. This target variable has a great significance to implement Artificial Neural Network because it contains our desired value, which tells about the network that how well the network is trained.Gradation of data processing displays the Figure 2.

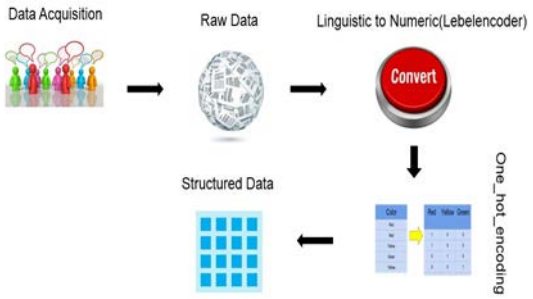

Figure 2: Data processing.

3.3 Neural Network

The concept and structure of Artificial Neural Network based on the human brain.The process(neuron) notes one at a time and learns by analyzing their prediction of the record (largely arbitrary) with the known genuine record. It performs its job by error calculating in every state until minimize the error to a threshold value.Weights are updated in every iteration. The error is recorded from the initial arrangement is feedbacked to the network and this mechanism is done for every evolution. The neural network is a layered architecture like our brain. The first and last layer called input and output layer respectively. All layers between input and output are mentioned as the hidden layer. Each set of input is modified by a set of unique weights and biases. Each edge has a unique weight and each node has unique biases. This means that the combination is used for activation is also unique. We want that accuracy to be high meaning that neural network predict value that is close to the desired output as possible every single time.

4 Implementation Details

4.1 Network adoption

Firstly data has been prepared and modified. Then method of training has been taken. It is the most crucial matter to then fix the topology of the neural network. There are several types of network. We can elect any of this as our appetite. Due to the aspects of our work, the Multilayer Perceptron network was chosen.

$$y = \varphi(\sum_{i=0}^{n} w_i x_i + b) = \varphi(W^T X + b) \quad (1)$$

where **w** denotes the vector of weights, **x** is the vector of inputs, b is the bias and φ is the activation function.

$$X = f(s) = B\varphi(As + a) + b \quad (2)$$

where **s** is a vector of inputs and **x** a vector of outputs. **A** is the matrix of weights of the first layer, **a** is the bias vector of the first layer. **B** and **b** are, respectively, the weight matrix and the bias vector of the second layer. The function φ denotes an element wise non linearity. A typical multilayer perceptron (MLP) network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes. The input signal propagates through the network layer-by-layer2. It is so tough building a neural network with the number of nodes and hidden layers. Because a small number of the hidden layer lower the processing capability. Comparatively the system will slow down if a large number of hidden layer. We come to a conclusion from this paradigm, determine a network with a hidden layer and ten input processing elements, one output. Train this network with different learning rate and by reducing or increasing the number of neurons in the hidden layer. For which learning rate and hidden layer our prediction came closer we took this network.

4.2 Allience dataset

In supervised training, the data is divided into 3 categories: training, testing and verification set. The training set permits the system to perceive correlations between input data and producing outputs so that it can improve the correlation between the input and the desired output. A total of 1200 patients documents were applied in the investigation. About 70% of the entire data (i.e.840 candidates) were used as the training set, 15% (i.e. 180 patients) as the testing set, and 15% (i.e. 180 patients) used for cross-validation.

4.3 Network training and validation process

First of all, we take our processed data. Thereafter we create a neural network. We use a hidden layer where twenty hidden neurons are used. The number of hidden neurons are fixed by changing the learning rate and hidden neurons. We take different learning rate and different hidden neurons for choosing the best network for our system. Which learning rate and hidden neurons give the best result we select that network. Through the dataset, we train the network. Levenberg- Marquardt backpropagation algorithm is used to train our neural network. Trainlm function is used for applying it to MATLAB.

5 Experiment And Evolution

The network performance validation is executed in Figure 3. Figure shows best validation comes out at epochs number 4 out of 6 epochs.

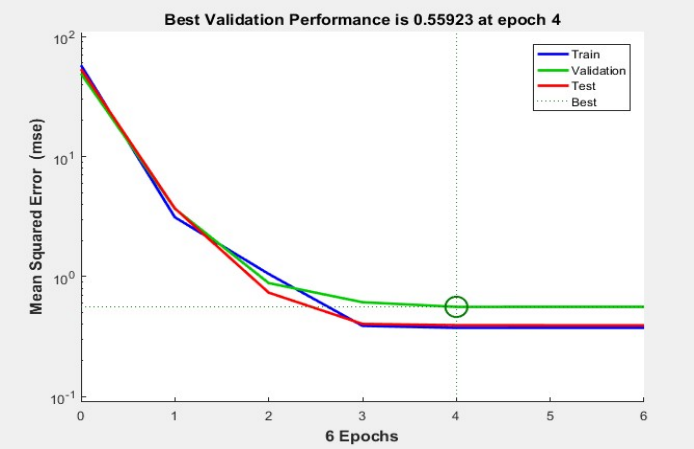

Figure 3: Performance validation.

In the Figure 4 shows the different errors histogram with 20 bins. Error histogram shows maximum data are trained up with less error.

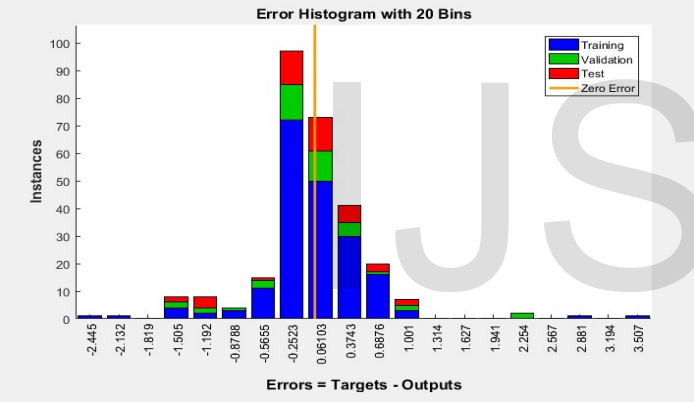

Figure 4: Error histogram.

Regression analysis of Figure 5 shows the consistency between input and output. Which tells us how well neural network are trained, validate and tested.It increase the reliability about network and datatset.

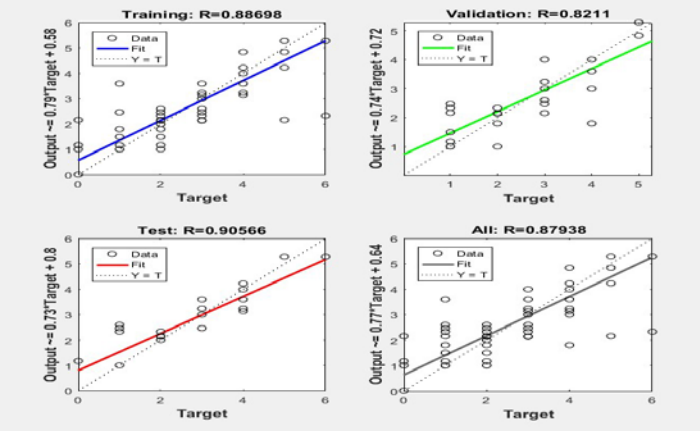

Figure 5: Regression

Figure 6 shows when the validation fail at the time of repetition.In this case at 6 epochs validation are failed.

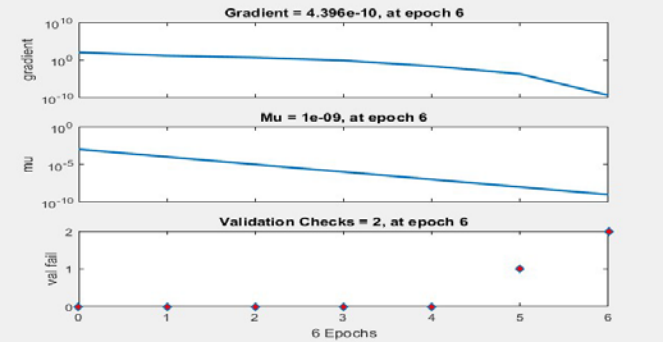

Figure 6: State

Figure 7 shows a relation among 10 patients survival time, level of cancer and their bunching. We see that whose patients survive more time their level of cancer is less or vice versa.

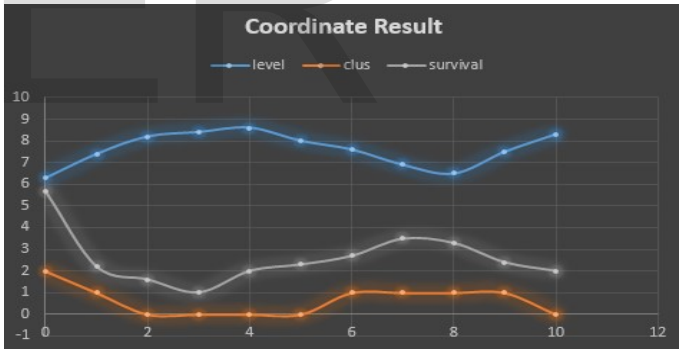

Figure 7: Coordinate Result

We do 3 clusters of patients. We see from above figure that whose patients survival time approximately 3 to 4 month and level of cancer is approximately 65% to 79% their bunch is 1, survival time approximately 1 to 2 month and level of cancer is approximately upto 80% their bunch is 0, survival time approximately 5 to 6 month and level of cancer is approximately bellow 65% their bunch is 2.

TABLE III: Result from ANN

Age	Survival(In month)	Predicted Survival(In month)	Error%
48	6	5.7289	0.27%
63	3	2.2258	0.77%
55	2	1.6288	0.37%
62	1	1.0000	0.0%
62	2	2.0002	0.0098%
55	2	2.3333	0.3333%
75	3	2.7456	0.2544%
63	4	3.5762	0.4238%
44	4	3.3561	0.6439%
50	3	2.4568	0.5432%
58	2	2.0356	0.0356%

In the TABLE III shows the error between patients actual survival and predicted survival (in month). It shows that our propose Neural Network predicted survival time(in month) and real dataset survival time(in month). It also show the error percentage between real data value versus our proposed systems predicted value.TABLE IV shows the Fuzzy logic result and classification result which we achive from K-means clustering.

TABLE IV: Result from FN and K-means

Age	Survival(In month)	Level of cancer	Bunch
48	6	63.2%	2
63	3	74.6%	1
55	2	82.7%	0
62	1	86.4%	0
62	2	84.5%	0
55	2	80.1%	0
75	3	76.5%	1
63	4	69.3%	1
44	4	65.2%	1
50	3	75.6%	1
58	2	83.9%	0

6 Conclution

In the present world artificial neural network is being tried to use everywhere.I think the artificial neural network in the medical sector can lead to breakthrough movements.The number of deaths due to cancer is increasing day by day.A good prediction last stage of cancer patients may be helpful for doctor and his family members. We try to investigate whether other efficient algorithms would lead to better discovery of predicting a patient survival time. We try to increase data volume and apply data mining for analyzing collected data.

Acknowledgment

The authors are grateful to the anonymous reviewers for their comments that improved the quality of our paper. This research was supported by the research fund of Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh. The authors thanks to the National Institute of Cancer Research Center and Hospital in Bangladesh.

References

[1] M.-H. Zheng, K.-Q. Shi, X.-F. Lin, D.-D. Xiao, L.-L. Chen, W.-Y. Liu, Y.-C. Fan, and Y.-P. Chen, "A model to predict 3-month mortality risk of acute-on-chronic hepatitis b liver failure using artificial neural network," *Journal of viral hepatitis*, vol. 20, no. 4, pp. 248–255, 2013.

[2] C.-C. Lin, Y.-K. Ou, S.-H. Chen, Y.-C. Liu, and J. Lin, "Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture," *Injury*, vol. 41, no. 8, pp. 869–873, 2010.

[3] F. Amato, A. Lopez, E. M. Pe´ na-M˜ endez, P. Va´ nhara, A. Hampl, and˘ J. Havel, "Artificial neural networks in medical diagnosis," 2013.

[4] D. Shanthi, G. Sahoo, and N. Saravanan, "Designing an artificial neural network model for the prediction of thrombo-embolic stroke," *International Journals of Biometric and Bioinformatics (IJBB)*, vol. 3, no. 1, pp. 10–18, 2009.

[5] A. Cucchetti, F. Piscaglia, A. D. Grigioni, M. Ravaioli, M. Cescon, M. Zanello, G. L. Grazi, R. Golfieri, W. F. Grigioni, and A. D. Pinna, "Preoperative prediction of hepatocellular carcinoma tumour grade and micro-vascular invasion by means of artificial neural network: a pilot study," *Journal of hepatology*, vol. 52, no. 6, pp. 880–888, 2010.

[6] A. I. Maas, C. W. Hukkelhoven, L. F. Marshall, and E. W. Steyerberg, "Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors," *Neurosurgery*, vol. 57, no. 6, pp. 1173–1182, 2005.

[7] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE transactions on Neural Networks*,

vol. 5, no. 6, pp. 989–993, 1994.

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[9] F. E. Ahmed, "Artificial neural networks for diagnosis and survival prediction in colon cancer," *Molecular cancer*, vol. 4, no. 1, p. 29, 2005.

[10] A. Hunter, L. Kennedy, J. Henry, and I. Ferguson, "Application of neural networks and sensitivity analysis to improved prediction of trauma survival," *Computer Methods and Programs in Biomedicine*, vol. 62, no. 1, pp. 11–19, 2000.

[11] A. Das, T. Ben-Menachem, G. S. Cooper, A. Chak, M. V. Sivak Jr, J. A. Gonet, and R. C. Wong, "Prediction of outcome in acute lowergastrointestinal haemorrhage based on an artificial neural network: internal and external validation of a predictive model," *The Lancet*, vol. 362, no. 9392, pp. 1261–1266, 2003.