Lung Cancer Survival Modelling using Adaptive Neuro-fuzzy Inference System

Plachikkad.A. Rehana Bhadhrika[1], Sri Kunjam Nageswara Rao[2]

[1]*M.Tech Student, Department of Computer Science and Systems Engineering,*
Andhra University College of Engineering,Andhra Pradesh, India
[2] *Assistant Professor, Department of Computer Science and Systems Engineering,*
Andhra University College of Engineering,, Andhra Pradesh, India

***Abstract*— The development of medical AI has been related to the growth of AI programs intended to help in the formulation of a diagnosis, prognosis and decision making. Approaches such as Artificial Neural Networks and fuzzy inference systems are widely used for expert behaviour modelling. The main aim of this project is to develop an approach for modelling lung cancer survival based on Adaptive neuro-fuzzy inference System where we combine both the learning capabilities of a neural network and reasoning capabilities of fuzzy logic in order to give enhanced prediction capabilities, as compared to using a single methodology alone.**

***Keywords*— Lung Cancer, ANFIS, Survival modelling, Lung Cancer Prognostic Index, Lung Cancer Prognostic Factors**

I. INTRODUCTION

Respiratory (lung) cancer is the second most common cancer, and the leading cause of cancer related deaths among men and women. Survival rate for lung cancer is estimated to be 15% after 5 years of diagnosis. Though Lung cancer is one of the most prevalent diseases and has a very high death rate when compared to other cancer related deaths, survival prediction and modelling techniques for lung cancer is not as developed as for other cancers like breast cancer and ovarian cancer. Earlier diagnosis and treatment should increase the survival rates, as the disease is much easier to control if it has not spread to other parts of the body.

Survival analysis is the analysis of data that relates to the time from when an individual enters a study until the occurrence of some particular event or end-point such as death or complete cure of the disease. It is deals with the comparison of survival function results for different combinations of risk factors. Analysis of survival data is complicated by the presence of censorship (patients leaving the study) and other unpredictable reasons. Statistical methods are commonly used in the analysis of survival data and lately artificial intelligence techniques have been considered as alternative methods for achieving this goal.

This paper proposes to discuss on a hybrid intelligent framework that uses both artificial neural networks (ANNs) and fuzzy inference. Artificial neural network has become very popular in the medical field following the discovery of the back-propagation algorithm, and this is extensively used in survival prediction [1],[2],[3]. We aim to use this advanced method of Adaptive fuzzy inference system (ANFIS) in predicting the hazard curve and survival curve of lung cancer patients. A specific form of data pre-processing has to be performed before a standard ANFIS model can be used for prognostic prediction of survival. A particular form of data pre-processing has to be performed before the ANFIS model can be used for prediction of survival using prognostic factors

II. PROBLEM DEFINITION

The challenge in survival analysis is how to best model the conditional hazard rate of failure times given certain covariate [6]. In addition there is abundant amount of Survival data available which are both censored and uncensored. Uncensored observations involved patients who are observed until they reach the end of the study. Censored observations on the other hand, involve only patients who survive beyond the end or who are lost to follow-up at some point. The presence of this type of data in the dataset makes the analysis of survival data more complicated. When proper data pre-processing procedures should performed and advanced survival modelling approaches are to be done on these data to make more accurate predictions by which the deaths due to lung cancer can be reduced immensely.

III. RELATED WORK

The use of Artificial Intelligence (AI) techniques in the medical field in the early 1970s emerged to model expert behaviour by utilising their knowledge and representing it in symbolic form. As it become more prevalent, attention has focused on the use of AI in modelling survival to improve upon the predictive power of proportional hazards in different kinds of cancers.

In modelling survival, a technique known as the partial logistic artificial neural network (PLANN) model was proposed by[9] to estimate smooth discrete hazards. It used a back-propagation ANN architecture in which the time interval is included as one of the inputs and is combined with multiple explanatory variables. The hidden representation of ANNs limited the understanding of the generated models to the average clinician. A number of rule extraction algorithms have been reported in the literature to provide trained artificial neural networks with explanation capability. Approaches based on decision trees, artificial immune systems, clustering, and many others, have been proposed to interpret the inputs, weights, biases, activation functions and outputs of ANNs.

This limitation encountered in symbolic ANN architectures was overcome by the use of a fuzzy inference system (FIS) Fuzzy inference seems to offer the capability to deliver the process of turning data into knowledge that can be understood by people. The framework proposed in this paper implements this effective solution of combining the fuzzy inference with ANN architecture to overcome the existing limitations.

IV. Proposed framework

Adaptive Neuro-Fuzzy Inference System(ANFIS)

The adaptive neuro-fuzzy inference system (ANFIS) proposed by Jang in 1993 [10], implements a Sugeno fuzzy inference method. The ANFIS architecture contains a sixlayer feed-forward neural network as shown in Figure 1.

Layer 1 is the input layer that passes external crisp signals to Layer 2, known as the fuzzification layer, to determine the membership grades for each input implemented by the given fuzzy membership function, for example the bell-shaped or gaussian curve. Layer 3 of ANFIS is the rule layer, which calculates the firing strength of the rule as the product of the membership grades. Layer 4 is called the ‘normalised firing strengths’, in which each neuron in the layer receives inputs from all neurons in Layer 3, and calculates the ratio of the firing strength of a given rule to the sum of firing strengths of all rules. Layer 5 is the defuzzification layer that yields the parameters of the consequent part of the rule. A single node in Layer 6 calculates the overall output as the summation of all incoming signals.

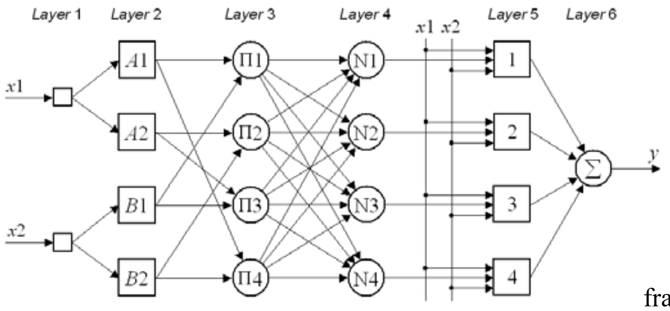

Fig. 1. Adaptive Neuro-Fuzzy Inference System (ANFIS*)*

ANFIS training can use alternative algorithms to reduce the error of the training. A combination of the gradient descent algorithm and a least squares algorithm is used for an effective search for the optimal parameters. The main benefit of such a hybrid approach is that it converges much faster, since it reduces the search space dimensions of the back-propagation method used in neural networks [10]. In the medical context, fuzzy approaches have been used in many areas, including in the prediction of patients’ survival rate and for relapse probability .

We now propose a step-by-step algorithm in modelling survival using the adaptive-neuro fuzzy inference system (ANFIS) as shown in Fig 2. This presents a complete framework to model survival which features automatic determination of the most suitable ANFIS model structure, and a modified ANFIS algorithm. The changes made to ANFIS are to address the specific constraint (in this context) of requiring non-negative values of hazard in certain iterations of the training. The explanation details the process in each step for use within any dataset.

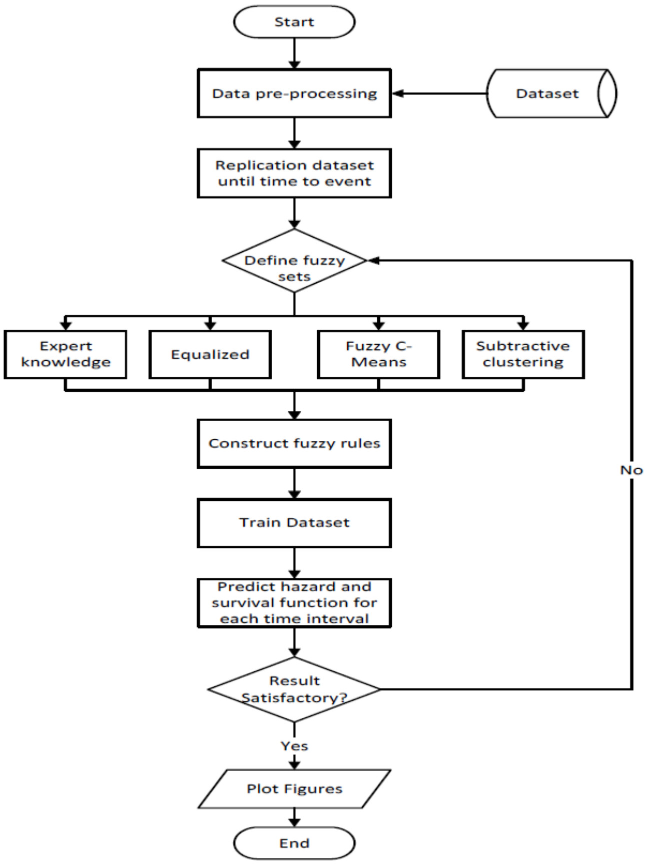

Fig 2: Flowchart for proposed framework

In general, the process is as follows. The proposed framework starts with data pre-processing by eliminating the data with missing values and the assigning of censorship (event status) after the period of study.

The fuzzy inference processes are performed by transforming each crisp input variable into a membership grade based on the membership functions defined, and conducting the fuzzy reasoning process by applying the appropriate fuzzy operators in order to obtain the fuzzy set. Data replication until time to an event is carried out to be used in the training process. The estimation of hazard and survival function in each time interval is obtained by performing the fuzzy inference calculations.

V. Framework process

A. Data

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute is an authoritative repository of cancer statistics in the United States. The data includes patient demographics, cancer type and site, stage, first course of treatment, and follow-up vital status. The SEER data attributes can be broadly classified as demographic attributes (e.g., age, gender, location), diagnosis attributes (e.g., primary site, histology, grade, tumour size), treatment attributes (e.g., surgical procedure, radiation therapy), and outcome attributes (e.g., survival time, cause of death), which makes the SEER data ideal for performing outcome analysis studies[11].

Lung Cancer Prognostic Index known as LCPI. The index is based on four prognostic factors namely age with 0.03 as coefficient[4], performance state (PS), primary tumour T, distant metastasis M and Regional Lymph nodes N. T,M, and N denotes the stage which has coefficient as 0.69 [4]. Descriptions are shown in Table I

TABLE I

PROGNOSTIC FACTORS FOR COMPUTING LCPI

PROGNOSTIC FACTORS	Computing value details LCPI	
Age	The numerical age should be considered	
PS	0	Fully active
	1	Restricted in physically strenuous activity but ambulatory
	2	Ambulatory and capable of all self care but unable to carry out any work
	3	Capable of only limited self care, confined to bed or chair
	4	Completely disabled. Cannot carry on any self care.
T	1	T1- Tumor 3 cm or less in greatest dimension
	2	T2- Tumor more than 3 cm but 7 cm or less
	3	T3- Tumor more than 7 cm
	4	T4- Tumor size is abnormally high
M	0	M0- No distant metastasis
	1	M1- Distant metastasis
N	0	N0
	1	N1
	2	N2
	3	N3

LCPI= 0.03(age)+1(if PS 1 or 2)+2 (if PS 3 or 4)+ 0.69 (1(if T 1 or 2) +2 (if T 3 or 4)+1(if N 0 or 1)+2(if N 2 or 3)+ M)

where, the categories of LCPI score as shown in Table II

TABLE II
CATEGORY OF LCPI SCORES

LCPI groups	LCPI cut-off
Good	Less than 3.69
Moderate	Between 3.7 to 6.25
Poor	Over 6.26

B. Data Pre-Processing

Pre-processing is a process that converts the raw inputs and outputs (target values) into a form understandable or acceptable before the training process. Often, this is used to reduce the dimensionality of input data and to optimise the generalization performance

The input of the network (survival time and LCPI groups) is replicated into t times which is the maximum survival time of an individual patient. The event attribute as a target of the network is also replicated and assigned as zero until the last time value is reached, where the event is 1 for occurrence and zero for censored. An example of replication is shown in Table III which shows the original data of three patients and Table IV which shows the replicated data suitable for input into

TABLE III

PRE-PROCESSING OF CATEGORICAL VARIABLES

LCPI Category			
Good	1	0	0
Moderate	0	1	0
Poor	0	0	1

TABLE IV

EXAMPLE FOR THREE PATIENT DATA SETS

	Time interval	LCPI category	Event
Patient 1	1	3	1
Patient2	2	1	1
Patient 3	3	3	0

TABLE V

REPLICATIONS OF ALL PATIENT DATA OBSERVED FOR EACH INTERVAL

	Time interval	LCPI category	Event
Patient 1	1	3	1
Patient2	1	1	1
	2	1	1
Patient 3	1	3	0
	2	3	0
	3	3	0

C. Define fuzzy sets

The second step of the framework is to transform each crisp input variable into a membership grade based on the memberships function defined, usually taken as a real value between 0 and 1. Often, when dealing with medical data, the grouping or cut-off of the variables has been defined by the clinical expert, thus, those techniques are chosen. LCPI cut-off scores are used for analyzing the survival and risk of the patient considered as shown in table II

D. Construct fuzzy rules

The next step is to conduct the fuzzy reasoning process by applying the appropriate fuzzy operators in order to obtain the fuzzy set to be accumulated in the output variable. In this framework, the grid partitioning method will be applied to generate the set of rules, by enumerating all possible combinations of membership functions of all inputs.

Example: IF survival time is less (< 3 years) AND stage group 3 THEN survival rate 0.58

THE SET OF RULES FOR LUNG CANCER SURVIVAL ESTIMATION	
Rule 1	IF survival time is less AND LCPI is good
Rule 2	IF survival time is less AND LCPI is moderate
Rule 3	IF survival time is less AND LCPI is poor
Rule 4	IF survival time is medium AND LCPI is good
Rule 5	IF survival time is medium AND LCPI is moderate
Rule 6	IF survival time is medium AND LCPI is poor
Rule 7	IF survival time is more AND LCPI is good
Rule 8	IF survival time is more AND LCPI is moderate
Rule 9	IF survival time is more AND LCPI is poor

E. Train Dataset

After the data have been prepared, and the parameters of the fuzzy inference system have been established, the data are ready to be trained. Batch training is applied in this framework, in which the updating of antecedents is performed after all the data have been trained. The gradient descent (GD) algorithm is used to update the antecedents, while the nonnegative least square (NNLS) algorithm is used to identify the consequent values. In this framework, implement a zeroth-order Sugeno model in which constants are used for the rule outputs. Further, the single output of the network known as the conditional event probability or hazard rate is the summation of all incoming signals with value obtained between 0 to 1. Two parameters have to be initialised before the training commences: the step size which was the length of each gradient transition in the parameter space; and the number of iterations to stop the training process.

F. Predict the hazard and survival function for each time interval

After the data have been trained, the final structure of the fuzzy inference system will be taken for the prediction process. The final membership function and the consequent values will be used into the fuzzy inference calculations to estimate the hazard function (hl) for the interval time defined by

$$S(t)=\prod_{l:tl<t}(1-h_l)$$

G. Result Satisfactory

The root mean square error (RMSE) was used as the performance measure. The RMSE is a measure between the desired target output and the actual current output. If the RMSE measure is not satisfactory, the adjustment of membership functions and the rule refinement procedure is activated towards better model optimisation. Therefore, the number of iterations for the training has to increase until convergence. Once the result is satisfactory, the hazard function and survival function can be plotted.

VI. Conclusions

Given the survival dataset into the framework, the estimation of hazard and survival curve can be plotted. In addition, the linguistic rules presented in this framework are to make it transparent to the clinician in the process of turning data into knowledge in modeling survival.

The representation of fuzzy inference systems, in which knowledge is encoded as a set of explicit linguistic rules that can be easily understood by people without technical expertise, it is hoped that this will allow the incorporation of expertise from clinicians into the selection of inputs and the modelling of rules. Thus, it is expected that such a technique may better address real clinical needs.

VII. FUTURE WORK

We also aim to create ANFIS models for other clinical data and cancer datasets. We also aim to repeat the study with the LCPI variable represented as a real number, utilizing between 3 and 7 membership functions. By doing so, we will be able to examine whether the membership functions can be trained to better match the data, rather than using the existing fixed clinical cut-offs presented in Table II.

References

[1] Adaptive Neuro-Fuzzy Inference System (ANFIS) in Modelling Breast Cancer Survival Hazlina Hamdan and Jonathan M. Garibaldi, Member, WCCI 2010 IEEE World Congress on Computational Intelligence July, 18-23, 2010 - CCIB, Barcelona, Spain

[2] An Exploration of the Adaptive Neuro-Fuzzy Inference System (ANFIS) in Modelling Survival By Hazlina Hamdan

[3] A. C. Joseph and S. W. David, “Applications of machine learning in cancer prediction and prognosis,” Cancer Informatics, vol. 2, pp. 59– 78, 2006.

[4] P. J. G. Lisboa, “A review of evidence of health benefit from artificial neural networks in medical intervention,” Neural Networks, vol. 15, no. 1, pp. 11– 39, 2002.

[5] H. Burke, P. Goodman, D. Rosen, D. Henson, J. Weinstein, F. Harrell, J. Marks, D. Winchester, and D. Bostwick, “Artificial neural network improve the accuracy of cancer survival prediction,” Cancer, vol. 79, no. 4, pp. 857–862, 1997.S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[6] Prediction of Clinical Outcome for All Stages and Multiple Cell Types of Non-small Cell Lung Cancer in Five Countries Using Lung Cancer Prognostic Index☆ Tiehua Chen,1, Luming Chen1

[7] J.-S. Jang, “Anfis adaptive-network-based fuzzy inference system,” Systems, Man and Cybernetics, IEEE Transactions on, vol. 23, pp. 665–685, May/Jun 1993.

[8] JIANQING, F., XIHONG, L. & LIU, J.S. (2009). New developments in biostatistics and bioinformatics. Higher Education Press , New Jersey.

[9] BIGANZOLI, E., BORACCHI, P., CORADINI, D., GRAZIA DAIDONE, M. & MARUBINI, E. (2003). Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors. Ann Oncol, 14, 1484–1493.

[10] J.-S. Jang, “Anfis adaptive-network-based fuzzy inference system,” Systems, Man and Cybernetics, IEEE Transactions on, vol. 23, pp. 665–685, May/Jun 1993.

[11] Lung cancer survival prediction using ensemble data mining on SEER data Ankit Agrawal ∗, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi and Alok Choudhary