Hepatocellular Carcinoma Survival Prediction Using Deep-Neural Network

Authors Name/s per 1st Affiliation (*Author*)

line 1 (of *Affiliation*): dept. name of organization

line 2-name of organization, acronyms acceptable

line 3-City, Country

line 4-e-mail address if desired

Authors Name/s per 2nd Affiliation (*Author*)

line 1 (of *Affiliation*): dept. name of organization

line 2-name of organization, acronyms acceptable

line 3-City, Country

line 4-e-mail address if desired

*Abstract*— Hepatocellular carcinoma, fifth most common type of liver cancer in adults. In patients with hepatocellular carcinoma, prediction of survival is very difficult. Through this prospective experiment, the authors have proposed a new improved classification approach using DNN (Deep Neural Network) for predicting survival of patients with hepatocellular carcinoma. The dataset was obtained at a University Hospital in Portugal and contains several demographic, risk factors, laboratory and overall survival features of 165 real patients diagnosed with HCC. Authors have selected fifteen risk factors out of forty-nine risk factors which are significantly responsible for HCC in this proposed method. The outcome of this experiment has proved to be of significant increase in accuracy of the prediction of survival over the conventional methods like multivariable cox model or unsupervised classification.

Keywords— Hepatocellular carcinoma, Classification, DNN, risk factors, survival, prediction.

# **Introduction**

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer in adults and it is the most common cause of death in people with cirrhosis, accounting for an estimated 500,000 deaths annually. HCC occurs due to liver inflammation and is most closely linked to chronic viral hepatitis infection (hepatitis B or C) or exposure to toxins such as alcohol or aflatoxin. The majority of HCC is prevalent in Southeast Asia and sub-Sahara Africa, the incidence of HCC has doubled in the United States over the past 25 years, and incidence and mortality rates are likely to double over the next 10–20 years.

# **RELATED WORKS**

In 1999, Josep M. Llovet along with Concepció Brú and Jordi Bruix [ ] proposed a method of classification using a new staging system, the Barcelona Clinic Liver Cancer (BCLC) staging classification, that comprises four stages that select the best candidates for the best therapies currently available.

In the same year, 1999 in July, Sylvie Chevret, Jean-Claude Trinchet [ ] proposed another method for prediction of HCC using a new prognostic classification, which selected Five prognostic factors, at the 0.0001 level. Seven hundred and sixty-one patients who presented with hepatocellular carcinoma from 24 Western medical centers were enrolled over a 30-month period. Patients were randomly assigned to either a training sample (n=506, with 418 deaths) from which a classification system was established, or a test sample (n=255, with 200 deaths) for validating its prognostic significance.

In the year of 2004, Ju-Seog Lee and his co-researchers [ ] published their work on classification and prediction of survival of HCC using gene expression profiling. The researchers analyzed global gene expression patterns of 91 human hepatocellular carcinomas (HCCs) to define the molecular characteristics of the tumours and to test the prognostic value of the expression profiles.

# **DATASET DESCRIPTION**

The dataset has been obtained from UCI Machine Learning repository, which has the data of 165 real patients diagnosed with HCC. The dataset contains 49 features selected according to the EASL-EORTC (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer) Clinical Practice Guidelines, which are the current state-of-the-art on the management of HCC.

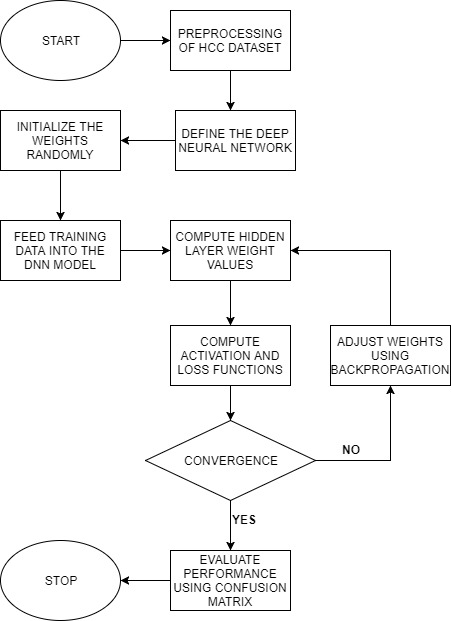
This is a heterogeneous dataset, with 23 quantitative variables, and 26 qualitative variables. Overall, missing data represents 10.22% of the whole dataset and only eight patients have complete information in all fields (4.85%). The target variables are the survival at 1 year and it was encoded as a binary variable: 0 (dies) and 1 (lives). A detailed description of the HCC dataset is presented in Table 1, which shows each feature’s type/scale, range, statistics (mean/mode) and missing rate percentage.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1  Characterization of CHUC’s hepatocellular carcinoma data. The dataset contains N = 165 records of n = 49 clinical variables, considered important to the clinicians’ decision process.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Prognostic factors | Type/scale | Range | Mean or mode | Missingness (%) | | Gender | Qualitative/dichotomous | 0/1 | 1 | 0 | | Symptoms | Qualitative/dichotomous | 0/1 | 1 | 10.91 | | Alcohol | Qualitative/dichotomous | 0/1 | 1 | 0 | | Hepatitis B Surface Antigen | Qualitative/dichotomous | 0/1 | 0 | 10.3 | | Hepatitis B e Antigen | Qualitative/dichotomous | 0/1 | 0 | 23.64 | | Hepatitis B Core Antibody | Qualitative/dichotomous | 0/1 | 0 | 14.55 | | Hepatitis C Virus Antibody | Qualitative/dichotomous | 0/1 | 0 | 5.45 | | Cirrhosis | Qualitative/dichotomous | 0/1 | 1 | 0 | | Endemic countries | Qualitative/dichotomous | 0/1 | 0 | 23.64 | | Smoking | Qualitative/dichotomous | 0/1 | 1 | 24.85 | | Diabetes | Qualitative/dichotomous | 0/1 | 0 | 1.82 | | Obesity | Qualitative/dichotomous | 0/1 | 0 | 6.06 | | Hemochromatosis | Qualitative/dichotomous | 0/1 | 0 | 13.94 | | Arterial Hypertension | Qualitative/dichotomous | 0/1 | 0 | 1.82 | | Chronic Renal Insufficiency | Qualitative/dichotomous | 0/1 | 0 | 1.21 | | Human Immunodeficiency Virus | Qualitative/dichotomous | 0/1 | 0 | 8.48 | | Nonalcoholic Steatohepatitis | Qualitative/dichotomous | 0/1 | 0 | 13.33 | | Esophageal varices | Qualitative/dichotomous | 0/1 | 1 | 31.52 | | Splenomegaly | Qualitative/dichotomous | 0/1 | 1 | 9.09 | | Portal hypertension | Qualitative/dichotomous | 0/1 | 1 | 6.67 | | Portal vein thrombosis | Qualitative/dichotomous | 0/1 | 0 | 1.82 | | Liver metastasis | Qualitative/dichotomous | 0/1 | 0 | 2.42 | | Radiological hallmark | Qualitative/dichotomous | 0/1 | 1 | 1.21 | | Age at diagnosis | Quantitative/ratio | 20–93 | 64.69 | 0 | | Grams/day | Quantitative/ratio | 0–500 | 71.01 | 29.09 | | Packs/year | Quantitative/ratio | 0–510 | 20.46 | 32.12 | | Performance status | Qualitative/ordinal | 0, 1, 2, 3, 4 | 0 | 0 | | Encephalopathy | Qualitative/ordinal | 1, 2, 3 | 1 | 0.61 | | Ascites degree | Qualitative/ordinal | 1, 2, 3 | 1 | 1.21 | | International Normalised Ratio | Quantitative/ratio | 0.84–4.82 | 1.42 | 2.42 | | Alpha-Fetoprotein (ng/mL) | Quantitative/ratio | 1.2–1,810,346 | 19299.95 | 4.85 | | Hemoglobin (g/dL) | Quantitative/ratio | 5–18.7 | 12.88 | 1.82 | | Mean Corpuscular Volume (fl) | Quantitative/ratio | 69.5–119.6 | 95.12 | 1.82 | | Leukocytes (G/L) | Quantitative/ratio | 2.2–13,000 | 1473.96 | 1.82 | | Platelets (G/L) | Quantitative/ratio | 1.71–459,000 | 113206.44 | 1.82 | | Albumin (mg/dL) | Quantitative/ratio | 1.9–4.9 | 3.45 | 3.64 | | Total Bilirubin (mg/dL) | Quantitative/ratio | 0.3–40.5 | 3.09 | 3.03 | | Alanine transaminase (U/L) | Quantitative/ratio | 11–420 | 67.09 | 2.42 | | Aspartate transaminase (U/L) | Quantitative/ratio | 17–553 | 69.38 | 1.82 | | Gamma glutamyl transferase (U/L) | Quantitative/ratio | 23–1575 | 268.03 | 1.82 | | Alkaline phosphatase (U/L) | Quantitative/ratio | 1.28–980 | 212.21 | 1.82 | | Total Proteins (g/dL) | Quantitative/ratio | 3.9–102 | 8.96 | 6.67 | | Creatinine(mg/dL) | Quantitative/ratio | 0.2–7.6 | 1.13 | 4.24 | | Number of nodules | Quantitative/ratio | 0–5 | 2.74 | 1.21 | | Major dimension of nodule (cm) | Quantitative/ratio | 1.5–22 | 6.85 | 12.12 | | Direct Bilirubin (mg/dL) | Quantitative/ratio | 0.1–29.3 | 1.93 | 26.67 | | Iron (mcg/dL) | Quantitative/ratio | 0–224 | 85.6 | 47.88 | | Oxygen Saturation (%) | Quantitative/ratio | 0–126 | 37.03 | 48.48 | | Ferritin (ng/mL) | Quantitative/ratio | 0–2230 | 439 | 48.48 | |

# **PROPOSED METHOD**

In the proposed method, the DNN model has been constructed using 3 hidden layers in between the input and output layer. The weight variables have been uniformly assigned in each layer for better optimization. Activation function used in the hidden layers is the commonly used “Linear Rectifier unit” (Relu) and at the output layer the sigmoid activation function has been used to retrieve predicted values in a probabilistic way. The basic flow of the experiment conducted by the authors is as follows:

1. Preprocessing of a database is a technique in the field of data-mining to transform raw untouched data into an understandable format. Usually, raw data contains missing values, inconsistent values and other lacking behavior. To avoid such issues, authors used preprocessing on the dataset. Normalizing, a process to deal with inconsistent values, is applied to scale the database values in a range of -1 and +1. Some missing values were filled using the mean strategy method and some by most frequent method.
2. The significant features which were having the most effect on the survival prediction value, were selected using the statistical correlation method to avoid unnecessary complexity in the model.
3. Then the whole dataset was divided into training and testing set in a ratio of 7:3. Both the training and testing set data have been shuffled to get the optimal result at the end of the experiment.
4. In the training phase, the training dataset was supplied to the DNN classifier model to train. The optimizer used to reduce the generated error is “Adam” optimizer with a learning rate of 0.03 and the loss function used was “Binary cross-entropy” function. The model was trained with a batch size of 20 and number of epochs being 100.
5. During the evaluation phase, the test dataset was supplied to the model, to predict the outcomes with a probabilistic value ranging from 0 to 1. The threshold value used to determine the output was 0.5.



**Fig 1.** Flowchart of the DNN training

# **RESULTS**

After the execution of both the training and testing phases, the performance of the proposed DNN model was evaluated based on some performance matrices. Theses performance metrics are such as: i) the accuracy, which is defined as a ratio of sum of the instances classified correctly to the total number of instances, (ii) precision, which is known as the ratio of correctly classified data in positive class to the total number of data classified as to be in positive class, (iii) recall (TP rate), which is defined as the ratio of tp to the total number of instances classified under positive class.

|  |  |  |
| --- | --- | --- |
| Predicted Class | PREDICTED: | PREDICTED: |
| Actual Class | 0 | 1 |
| TRUE: 0 | 11 | 6 |
| TRUE: 1 | 8 | 25 |

**Table 2**. Confusion Matrix of the DNN Model

|  |  |  |  |
| --- | --- | --- | --- |
|  | DNN (%) | KNN (%) | SVM (%) |
| ACCURACY | 72 | 64 | 58 |
| PRECISION | 57.89 | 66.66 | 0 |
| RECALL | 64.7 | 28.57 | 0 |

**Table 3**. Performance comparison between different models

# CONCLUSION

In this work, a new methodology capable of predicting the 1year survival for patients with HCC has been presented. To achieve that, a HCC dataset composed by 165 patients followed in an university hospital center was used. At the beginning of the study, this dataset presented three main challenges: dataset having data of only 165 patients thus resulting in the difficulty of training the neural network, the presence of heterogenous variables ( there are dichotomous, ratio scaled and ordinal features/variables ) and only eight patients having complete information .

In the model 15 most significant variables are selected to train the model. After training the dataset with 3 different models, the performance of these models are compared using 3 parameters and then the DNN model seemed to have outperformed the KNN and Kernel SVM model.

This topic could be a possibility for future work: extending our methodology to other contexts besides HCC disease, whether they are healthcare contexts or not.