Hepatocellular Carcinoma Survival Prediction Using Deep Neural Network

Chayan Kumar Kayal, Sougato Bagchi, Debraj Dhar, Tirtha Maitra and Sankhadeep Chatterjee

Abstract Hepatocellular carcinoma is one of the most common types of liver cancer in adults. In patients having this disease, prediction of survival is very strenuous. Through this eminent 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 15 risk factors out of 49 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 · Deep neural network Survival

C. K. Kayal · S. Bagchi · D. Dhar · T. Maitra · S. Chatterjee (✉)
Department of Computer Science & Engineering, University of Engineering and Management, Kolkata, India
e-mail: chatterjeesankhadeep.cu@gmail.com

C. K. Kayal
e-mail: chayankayal32@gmail.com

S. Bagchi
e-mail: sougato97@gmail.com

D. Dhar
e-mail: debrajdhar100@gmail.com

T. Maitra
e-mail: tirthamaitra0@gmail.com

M. Chakraborty et al. (eds.), *Proceedings of International Ethical Hacking Conference 2018*, Advances in Intelligent Systems and Computing 811,
https://doi.org/10.1007/978-981-13-1544-2_28

1 Introduction

Hepatocellular carcinoma (HCC) is one of the primary liver cancers found in adults, and it is the most common cause of death in people with cirrhosis, accounting for an estimated 5 lakhs deaths annually [1]. HCC happens as a result of liver inflammation and is most closely linked to chronic viral hepatitis infection (hepatitis B or C) or exposure to various toxins such as alcohol or aflatoxin. The majority of HCC is prevalent in the southern part of the world, mainly Southeast Asia and some suburban areas of Africa. The incidence of HCC has doubled in the USA over the past two and a half decades, and incidence and mortality rates are likely to double over the next two decades [2]. Today data-propelled statistical research has become an attractive complement for analytical research. Survival prediction is one of the most challenging tasks among these medical researchers. These statistical predictions help the analysis of the physician and also save huge resources. These researches are done with the help of some computational techniques/methods such as ANN, SVM and KNN as mentioned in [3]. These techniques are able to model unknown/complex relationships which are nonlinear or noisy and are difficult to analyze.

Recent researches have focused on application of machine learning in cancer prediction. Llovet et al. [4] proposed a method of classification using a new staging system, known as the Barcelona Clinic Liver Cancer (BCLC) staging classification, which selects the best candidate/patient for the most optimal therapy currently available using four stages of operation. Another method for prediction of HCC has been reported by Chevret et al. [5] using a new prognostic classification, which selected five major prognostic parameters. In their proposed work, they analyzed the data of 761 patients with HCC. The splitting was done randomly based on which they established a new classification system. Lee et al. [6] published their work on classification and prediction of survival of HCC using gene expression profiling in 2004. The primary aim of the researchers was to derive the molecular characteristics of any tumor and the secondary aim to test their prognostic value based on their expression profile. In 2015, Santos et al. [7] presented their work on the HCC using k-means clustering and SMOTE algorithm to build a representative dataset and use it as training example for different machine learning procedures.

2 Proposed Method

Artificial neural network (ANN) [8–13] can be considered as the mere replica of the biological neurons present in the human brain. The human brain is the most complex and powerful functional unit. It is capable of handling complex relations and taking important decisions in less than a fraction of second and also capable of modeling complex/unknown functional relationships with interconnected processing units (artificial neurons) [14–24]. That is the reason for the interest of replicating this enormous powerful model of computing, and this gave birth to ANN. ANN is different

for traditional hardcoded algorithms. These learn the functional relationships from a given dataset during its learning (training) stage. ANN mainly consists of an input layer, an output layer and one hidden layer in between input and output layers. The hidden layers are responsible for all the computations in a neural network model. In the proposed method, authors have used the DNN (deep neural network) model to perform the HCC classification task.

DNN is basically an extension of the ANN, where the number of hidden layers is more than one. To evaluate the effectiveness of the DNN model, authors have compared it with other models such as SVM (support vector machine) and KNN (k-nearest neighbor).

In SVM, the variables are mapped on a 3d/2d plane using some mapping functions and an optimal hyperplane is drawn to classify the variables. This hyperplane is drawn by considering the worst type of variables of different kinds. Therefore, this model is more robust as this takes into account the worst conditions. The variables are classified on the basis of their location inside/outside the kernel, or on the side of the hyperplane.

KNN is a nonparametric method used for classification and regression problems in machine learning. Here, a new variable is categorized by comparing the distance (mainly Euclidian distance) between that point and other points of different categories. The category having the maximum number of neighboring variables (or maximum number of variables with least distance) with that new variable is termed as the type of that variable.

In the current study, the DNN model has been constructed using the well-known "keras" library. The model consists of four hidden layers in between the input and output layers. Each layer containing, respectively, 1024, 512, 256 and 128 neurons which were chosen randomly based on trial-and-error approach by the authors. 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 optimizer used in the DNN architecture is one of the variants of the gradient descent optimizer, known as "Adam" optimizer, which was found to have better optimization result in the current study [1].

Figure 1 depicts the flow of the experiments in the current study. The basic flow of the experiment conducted by the authors is as follows:

1. Preprocessing: The following preprocessing is done in the dataset before the classification.

    (i) Feature extraction—This step involves extraction of significant features which are most important in classification. In the presented work, the feature extraction had been performed by finding out the correlation between the feature and the target. During this phase, out of 49 features, 14 unique features were extracted out of the dataset which were found to have the most significant effect on the class prediction result, resulting in lesser distortion and higher accuracy [3].

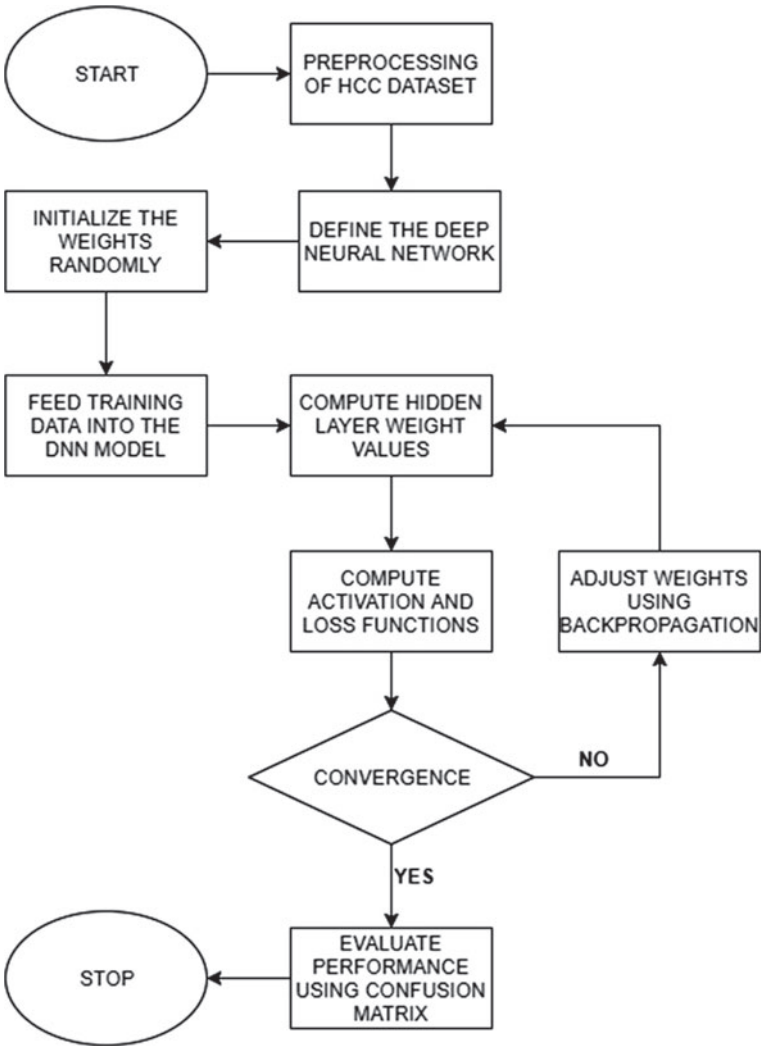

Fig. 1 Flowchart of the DNN training

(ii) Data cleaning—The dataset might contain missing or inconsistent values. To deal with such issues, statistical methods are used in this step of preprocessing.

(iii) Data normalization—The normalization of the dataset is carried out to reduce the distance between attribute values. It is generally achieved by keeping the value range in between −1 and +1.

2. Then, the whole dataset was divided into training and testing sets 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.
3. 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.
4. 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.

3 Dataset Description

The dataset has been obtained from UCI Machine Learning Repository, which contains data of 165 patients who were detected with HCC. The dataset contains a total of 49 features which were selected by a European Organization, EASL-EORTC, who are specialized in research and treatment of cancer [25]. The dataset is a divergent one, with a total of 23 quantitative variables and 26 qualitative variables. The number of missing values contained in the dataset is over 10%, while only eight instances of the patient record have complete information in all the attributes, with a percentage value of 4.85% among the whole datasets. The target variables are the survival at 1 year, and it was encoded as a binary variable: 0 (dies) and 1 (lives).

In the dataset, each feature has a missing value percentage between 0 and 48.48%. The conventional approaches for dealing with such missing values are deletion of a record or imputation of a value. One of the most commonly used strategies when dealing with missing values is the elimination of that particular instance, but this approach has been ruled out since the beginning due to the large number of incomplete data. Thus, the authors used the imputation-based approach to handle the records using the mean and most frequent strategy.

Table 1 depicts a detailed description of the HCC dataset showing each feature's type/scale and the range of values of each feature along with the missing values contained in each of the feature columns [2, 4].

4 Results and Discussion

After the execution of both the training and testing phases, the proposed DNN model performance was evaluated based on some matrices. Theses metrics are: (i) the accuracy, which is defined as a ratio of sum of the specimens classified correctly to the total number of specimens, (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

Table 1 Characterization of HCC data according to medical reports

Prognostic factors	Type/scale	Range	Missing (%)
Gender	Qualitative/dichotomous	0/1	0
Symptoms	Qualitative/dichotomous	0/1	10.91
Alcohol	Qualitative/dichotomous	0/1	0
Hepatitis B surface antigen	Qualitative/dichotomous	0/1	10.3
Hepatitis B e antigen	Qualitative/dichotomous	0/1	23.64
Hepatitis B core antibody	Qualitative/dichotomous	0/1	14.55
Hepatitis C virus antibody	Qualitative/dichotomous	0/1	5.45
Cirrhosis	Qualitative/dichotomous	0/1	0
Endemic countries	Qualitative/dichotomous	0/1	23.64
Smoking	Qualitative/dichotomous	0/1	24.85
Diabetes	Qualitative/dichotomous	0/1	1.82
Obesity	Qualitative/dichotomous	0/1	6.06
Hemochromatosis	Qualitative/dichotomous	0/1	13.94
Arterial hypertension	Qualitative/dichotomous	0/1	1.82
Chronic renal insufficiency	Qualitative/dichotomous	0/1	1.21
Human immunodeficiency virus	Qualitative/dichotomous	0/1	8.48
Nonalcoholic steatohepatitis	Qualitative/dichotomous	0/1	13.33
Esophageal varices	Qualitative/dichotomous	0/1	31.52
Splenomegaly	Qualitative/dichotomous	0/1	9.09
Portal hypertension	Qualitative/dichotomous	0/1	6.67
Portal vein thrombosis	Qualitative/dichotomous	0/1	1.82
Liver metastasis	Qualitative/dichotomous	0/1	2.42
Radiological hallmark	Qualitative/dichotomous	0/1	1.21
Age at diagnosis	Quantitative/ratio	20–93	0
Grams/day	Quantitative/ratio	0–500	29.09
Packs/year	Quantitative/ratio	0–510	32.12
Performance status	Qualitative/ordinal	0, 1, 2, 3, 4	0
Encephalopathy	Qualitative/ordinal	1, 2, 3	0.61
Ascites degree	Qualitative/ordinal	1, 2, 3	1.21
International normalized ratio	Quantitative/ratio	0.84–4.82	2.42
Alpha fetoprotein (ng/mL)	Quantitative/ratio	1.2–1,810,346	4.85
Hemoglobin (g/dL)	Quantitative/ratio	5–18.7	1.82
Mean corpuscular volume (fl)	Quantitative/ratio	69.5–119.6	1.82
Leukocytes (G/L)	Quantitative/ratio	2.2–13,000	1.82
Platelets (G/L)	Quantitative/ratio	1.71–459,000	1.82
Albumin (mg/dL)	Quantitative/ratio	1.9–4.9	3.64
Total bilirubin (mg/dL)	Quantitative/ratio	0.3–40.5	3.03
Alanine transaminase (U/L)	Quantitative/ratio	11–420	2.42

(continued)

Table 1 (continued)

Prognostic factors	Type/scale	Range	Missing (%)
Aspartate transaminase (U/L)	Quantitative/ratio	17–553	1.82
Gamma glutamyl transferase (U/L)	Quantitative/ratio	23–1575	1.82
Alkaline phosphatase (U/L)	Quantitative/ratio	1.28–980	1.82
Total proteins (g/dL)	Quantitative/ratio	3.9–102	6.67
Creatinine (mg/dL)	Quantitative/ratio	0.2–7.6	4.24
Number of nodules	Quantitative/ratio	0–5	1.21
Major dimension of nodule (cm)	Quantitative/ratio	1.5–22	12.12
Direct bilirubin (mg/dL)	Quantitative/ratio	0.1–29.3	26.67
Iron (mcg/dL)	Quantitative/ratio	0–224	47.88
Oxygen saturation (%)	Quantitative/ratio	0–126	48.48
Ferritin (ng/mL)	Quantitative/ratio	0–2230	48.48

Table 2 Confusion matrix of the DNN model

Actual class	Predicted class	
	Predicted: 0	Predicted: 1
True: 0	22	6
True: 1	8	27

class, (iii) recall (true positive rate), which is defined as the ratio of true positive to the total number of instances classified under positive class. Table 2 reveals the confusion matrix for DNN model where class "0" indicates negative survivability and "1" denotes positive survivability.

Table 3 reveals that the accuracy of the DNN model is 78%, which is certainly better compared with the accuracy of KNN model which is 64% and it is also better than support vector machine (SVM) model which is having an accuracy of 58%. On the other hand, the precision of the DNN model is 83.58%, whereas KNN model is having a precision of 63.41% and SVM model is having a precision of 58%. Recall value percentage of the DNN is at 81.25% in contrast to KNN being 89.65% and SVM having 100%. Another one of the important evaluation parameters F-measure percentage of DNN is 80% which is greater than the 74.28% and 73.41% of the KNN and SVM models, respectively. The SVM established the worst result among the other two models, whereas DNN model has been established as the most efficient model in the testing phase of the dataset.

The observations obtained from Fig. 2 show which errors are there in the predicted values of the test dataset. Corresponding red line denotes the wrong predicted values, and the blue line denotes the original target values.

Table 3 Performance comparison between different models

	DNN (%)	KNN (%)	SVM (%)
Accuracy	78	64	58
Precision	83.58	63.41	58
Recall	81.25	89.65	100
F-Measure	80	74.28	73.41

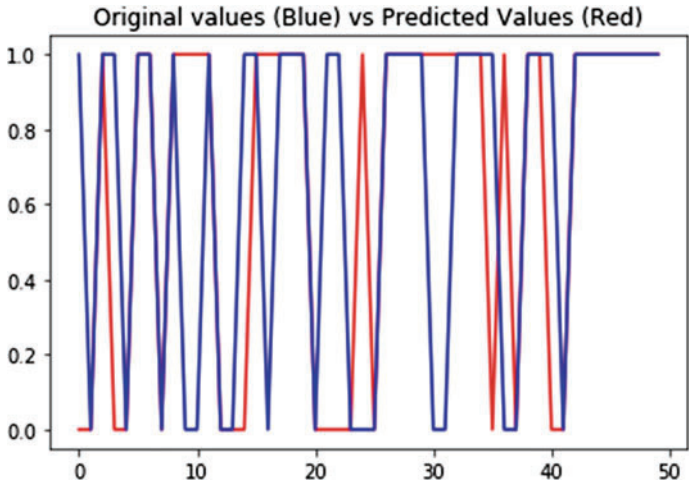

Fig. 2 Graph between original values and predicted values of DNN model

5 Conclusion

The current study proposed a deep neural network-based prediction of survival of patients suffering of hepatocellular carcinoma. Traditional methods such as SVM- and KNN-based methods have found to be not suitable for such predictions. Experimental results have revealed that the performance of DNN is superior to other classifiers. Nevertheless, future studies could be focused on appropriate feature selection to build more efficient and trustworthy classifier to predict survivability.

References

1. Parkin, D.M., Bray, F., Ferlay, J., Pisani, P.: Estimating the world cancer burden: Globocan 2000. Int. J. Cancer **94**, 153–156 (2001)
2. El Serag, H.B., Mason, A.C.: Rising incidence of hepatocellular carcinoma in the United States. N. Engl. J. Med. **340**, 745–750 (1999)
3. Thorgeirsson, S.S., Grisham, J.W.: Molecular pathogenesis of human hepatocellular carcinoma. Nat. Genet. **31**, 339–346 (2002)

4. Llovet, J.M., Brú, C., Bruix, J.: Prognosis of hepatocellular carcinoma: the BCLC staging classification. Semin. Liver Dis. **19**(3): 329–338 (1999)
5. Chevret, S., Trinchet, J.-C., Mathieu, D., AbouRached, A., Beaugrand, M., Chastang, C.: A new prognostic classification for predicting survival in patients with hepatocellular carcinoma. J. Hepatol. **31**(1), 133–141 (1999)
6. Lee, J.-S., Chu, I.-S., Heo, J., Calvisi, D.F., Sun, Z., Roskams, T., Durnez, A., Demetris, A.J., Thorgeirsson, S.S.: Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. **40**(3), 667–676 (2004)
7. Santos, M.S., Abreu, P.H., Garcia-Laencina, P.J., AdeliaSimao, A.C.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inform. **58**, 49–59 (2015)
8. Chatterjee, S., Sarkar, S., Hore, S., Dey, N., Ashour, A.S., Shi, F., Le, D.N.: Structural failure classification for reinforced concrete buildings using trained neural network based multi-objective genetic algorithm. Struct. Eng. Mech. **63**(4), 429–438 (2017)
9. Chatterjee, S., Dey, N., Shi, F., Ashour, A.S., Fong, S.J., Sen, S.: Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. Med. Biol. Eng. Comput. 1–12 (2017)
10. Chatterjee, S., Sarkar, S., Dey, N., Ashour, A.S., Sen, S., Hassanien, A.E.: Application of cuckoo search in water quality prediction using artificial neural network. Int. J. Comput. Intell. Stud. **6**(2–3), 229–244 (2017)
11. Chatterjee, S., Banerjee, S., Mazumdar, K.G., Bose, S., Sen, S: Non-dominated sorting genetic algorithm—II supported neural network in classifying forest types. In: 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech) (pp. 1–6). IEEE, April 2017
12. Chatterjee, S., Banerjee, S., Basu, P., Debnath, M., Sen, S.: Cuckoo search coupled artificial neural network in detection of chronic kidney disease. In: 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech) (pp. 1–4). IEEE, April 2017
13. Chatterjee, S., Dey, N., Ashour, A.S., Drugarin, C.V.A.: Electrical energy output prediction using cuckoo search based artificial neural network. In: Smart Trends in Systems, Security and Sustainability (pp. 277–285). Springer, Singapore (2018)
14. Chakraborty, S., Dey, N., Chatterjee, S., Ashour, A.S.: Gradient Approximation in Retinal Blood Vessel Segmentation
15. Chatterjee, S., Sarkar, S., Dey, N., Ashour, A.S., Sen, S.: Hybrid non-dominated sorting genetic algorithm: II-neural network approach. Adv. Appl. Metaheuristic Comput. **264** (2017)
16. Chatterjee, S., Sarkar, S., Hore, S., Dey, N., Ashour, A.S., Balas, V.E.: Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings. Neural Comput. Appl. **28**(8), 2005–2016 (2017)
17. Chatterjee, S., Ghosh, S., Dawn, S., Hore, S., Dey, N.: Forest type classification: a hybrid NN-GA model based approach. In: Information Systems Design and Intelligent Applications (pp. 227–236). Springer, New Delhi (2016)
18. Chatterjee, S., Hore, S., Dey, N., Chakraborty, S., Ashour, A.S.: Dengue fever classification using gene expression data: a PSO based artificial neural network approach. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 331–341). Springer, Singapore (2017)
19. Hore, S., Chatterjee, S., Sarkar, S., Dey, N., Ashour, A.S., Balas-Timar, D., Balas, V.E.: Neural-based prediction of structural failure of multistoried RC buildings. Struct. Eng. Mech. **58**(3), 459–473 (2016)
20. Chatterjee, S., Raktim C., Dey, N., Hore, S.: A quality prediction method for weight lifting activity 95–98 (2015)
21. Hore, S., Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Van Chung, L., Le, D.N.: An integrated interactive technique for image segmentation using stack based seeded region growing and thresholding. Int. J. Electr. Comput. Eng. **6**(6), 2773 (2016)

22. Chatterjee, S., Paladhi, S., Hore, S., Dey, N.: Counting all possible simple paths using artificial cell division mechanism for directed acyclic graphs. In: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1874–1879). IEEE, March 2015
23. Hore, S., Chatterjee, S., Chakraborty, S., Shaw, R.K.: Analysis of different feature description algorithm in object recognition. In: Feature Detectors and Motion Detection in Video Processing, p. 66 (2016)
24. Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Ashour, A.S., Shi, F., Mali, K.: Modified cuckoo search algorithm in microscopic image segmentation of hippocampus. Microsc. Res. Tech. **80**(10), 1051–1072 (2017)
25. Lichman, M.: UCI machine learning repository (http://archive.ics.uci.edu/ml). Irvine, CA: University of California, School of Information and Computer Science (2013)