A Deep Learning Strategy for Predicting Liver Cancer Using Convolutional Neural Network Algorithm

Sirajali Nagalpara[1] and Bhavesh M. Patel[2]*

Abstract

One of the common types of cancer is liver cancer, early detection and diagnosis of which are critical. Discovery, decision, and aggressive therapy can prevent most cancer deaths. We use data mining approaches (Convolutional Neural Networks) to build prediction models for liver cancer with the most widely used statistical analysis methodology. Around 579 records and 10 variables were included in the data collection. The model was built, evaluated, and compared using a k-fold cross-validation process. CNN was the best accurate predictor for this domain with a test set accuracy of 100%.

Keywords : **Cancer, Convolutional Neural Networks, data mining, statistical analysis**

I. INTRODUCTION

Recently, liver cancer has reached epidemic proportions. According to the World Cancer Research Journal, the incidence rate of liver cancer was 10.1 per one million persons in 2017, with a fatality rate of 9.5 per 0.1 million individuals. According to Cancer.net, 42,220 Americans are affected, including 11,610 females and 30,610 males. Since 1980, cases have increased thrice, posing a serious public health danger. According to the same site, Southeast Asia and Sub-Saharan Africa are the most affected regions[1].

There are two types of liver cancer, primary and secondary[1]. If we can detect cancer early, patients can get better and timely treatment [2]. Biological factors are complemented by data mining, a powerful tool used to find patterns in a particular cancer dataset [3].

In this research, we used the deep learning Convolutional Neural Network (CNN) technique to develop a model for predicting Liver Cancer. The dataset used for this study was obtained from UCI Machine Learning Repository.[2] The dataset includes 579 records and 10 variables. The data cleaning and preparation stage, the data set and removing records were missing values. The model was built using a k-fold cross-validation process for evaluation and comparison purposes.

The outcomes revealed that CNN is the best predictor, with the test data set accuracy of 100%. The CNN model is also analysed using sensitivity to determine the predicted elements' relative importance.

[1] https://www.rgcirc.org/cancer-knowledge-base/liver-cancer-signs-and-symptoms/

[2] https://archive-beta.ics.uci.edu/ml/datasets

Manuscript Received : May 5, 2022 ; Revised : May 14, 2022 ; Accepted : May 15, 2022. Date of Publication : June 5, 2022.

S. Nagalpara*[1], *Research Scholar*, Department of Computer Science, Hemchandracharya North Gujarat University, Patan - 384265, Gujarat. Email : sirajit.momin@gmail.com ; ORCID iD : https://orcid.org/0000-0001-8132-5795

B. M. Patel[2], *Assistant Professor*, Department of Computer Science, Hemchandracharya North Gujarat University, Patan - 384265, Gujarat. Email : bmpatel@ngu.ac.in ; ORCID iD: https://orcid.org/0000-0002-6722-3108

DOI : https://doi.org/10.17010/ijcs/2022/v7/i3/171270

II. BACKGROUND

A. Liver Cancer

Liver cancer is the most common cause of cancer deaths worldwide and the fifth most prevalent in the United States. It is also the only one of the top five deadliest cancers to experience an annual percentage increase in occurrence. The prognosis for liver cancer is poor. Only 5% to 15% of patients are eligible for surgical removal, which is suitable only for early-stage patients [4].

Although there are multiple established risk factors for liver cancer, no one knows why some men develop the disease while others do not. Alcohol, Diabetes Mellitus, obesity, and Autoimmune Hepatitis are some of the most critical risk factors for liver cancer [5] .

Hepatocellular Carcinoma (HCC), often known as primary liver cancer, is one of the leading causes of death in patients with chronic liver disease. HCC is becoming the primary cause of cancer in India, owing to an epidemic of non-alcoholic fatty liver disease [5].

B. Related Models and Theories

CNN : Many computer vision applications, such as object detection and recognition have lately seen advancements, thanks to deep learning methods, particularly Convolutional Neural Networks (CNNs). Deep Learning is highly effective in automating learning feature representation schemes from training data by attempting to mimic the structure and operation of neurons in the human visual cortex through hierarchical multi-layer networks. This eliminates the laborious feature engineering task [6].

Deep Learning (DL) is based on a traditional neural network, but it outperforms its predecessors significantly. Furthermore, DL simultaneously uses transformations and graph technologies [7].

The earliest levels of these algorithms extract low-level features, whereas the last layers extract high-level features resulting in a multi-layer data representation architecture. This style of architecture, which models the process that occurs in whole sensory regions within the human brain was inspired by Artificial Intelligence (AI). The human brain can automatically derive data representation from many scenes. The identified objects are the process's output, while the received scene information is the process's input. This procedure imitates the way the human brain functions. As a result, it stresses the primary advantage of DL. Fig. 1 shows the CNN model diagram.

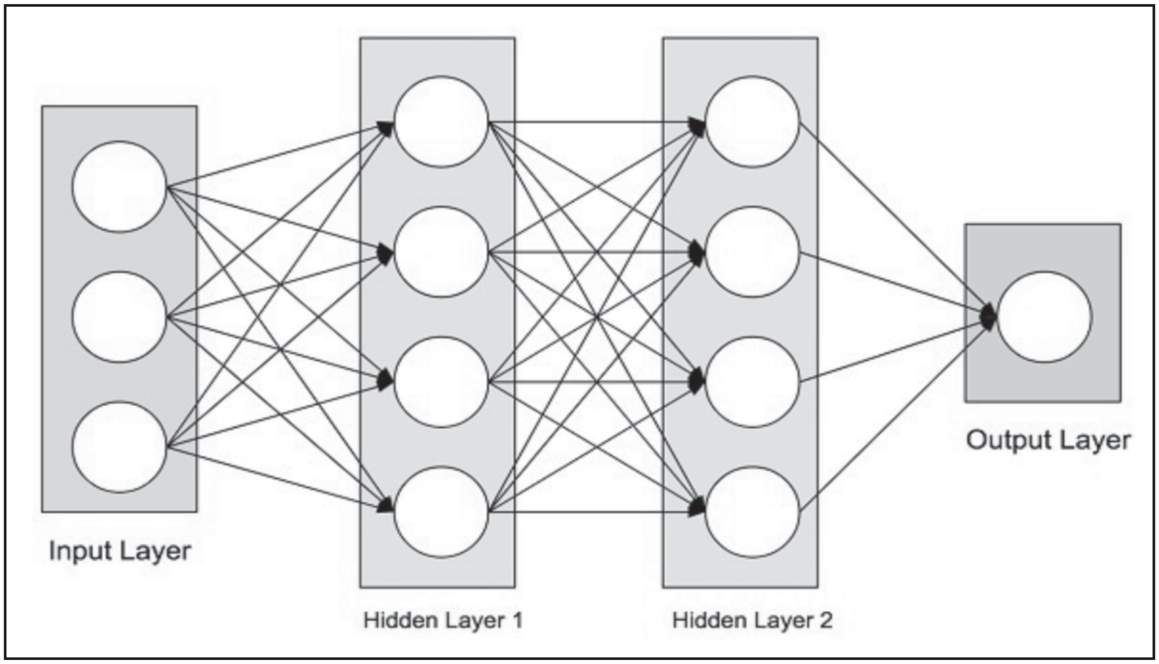

Fig. 1. CNN Model Diagram

$$h^k = f(W^k * x + b^k) \qquad (1)$$

where,

x Input of each layer in a CNN model is organized in three dimensions;

k Convolutional layer is denoted by k;

b^k Bias;

W^k Weight;

h^k Feature maps.

The next step is down-sampling every feature map in the sub-sampling layers. This leads to a reduction in network parameters, which accelerates the training process and in turn enables handling of the overfitting issue [7].

III. LITERATURE REVIEW

Random Forestwas used to predict Liver Cancer and it was found that the J48 algorithm performs better than all other classification algorithms.With the help of feature selection it has an accuracy of 95.04% [1]. K-Nearest Neighbour (KNN) outperformed Logistic Regression, Auto Neural, and Random Forest with an accuracy of 99.794% [8]. In this work Support Vector Machines for classifying liver disease using diabetes datasets and two liver patents datasets were used to examine accuracy, error rate, sensitivity, prevalence and specificity [9]. Naïve Bayes classification (NBC), Bagging, KStar, Logistic, and REP trees were used to evaluate the accuracy, precision, sensitivity, and specificity. In this case, UCLA and AP datasets were considered for finding the best algorithm. KStar algorithm proved to have the highest accuracy, precision, sensitivity, and specificity [10].

Naïve Bayes, ANN, Xero, 1BK, and VFI algorithms are used to classify these diseases and compare their effectiveness and correction rates. In terms of accuracy in predicting or describing liver disease, early detection is critical for its cure. It also results in better performance of the classification model.[2]. Naïve Bayes, FT Tree, and KStar algorithms were tested, and Naïve Bayes had the best classification accuracy of 96.52% [11]. Using the Support Vector Data Description (SVDD), data visualization techniques, and Glow Worm Swarm Optimisation (GSO) were used to improve diagnostic accuracy. As a result, the proposed method achieves 96% sensitivity, 86.28% specificity, and 84.28% accuracy. Thus, it is well suited for diagnosing early liver disease [12]. Naïve Bayesian and C4.5 classification data mining techniques were used to analyze liver disease. The experiment showed that C4.5 had an 99.20% accuracy [13]. Influenced symptoms in alcoholic fatty liver disease apply support and confidence to the Reverse Sequential Covering Algorithm [13].

Sequential Minimal Optimisation (SMO), Naïve Bayes, J48, Bayes Net data, and mining algorithms were used for detecting and predicting liver disease. The results showed that the SMO model was highly accurate with 97.3913% accuracy [14]. Applying J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding Trees showed that Decision Stump provided the highest accuracy compared to other techniques [15].

NWKNN and Fuzzy-NWKNN methods gave 73.91% accuracy (BUPA Dataset), 77.59% accuracy (ILPD Dataset), and 87.01% accuracy (MPRLPD Dataset).Furthermore, the TLRUS pre-processing method is used for improved accuracy of 78.46% (BUPA Datasets), 78.46% (ILPD Datasets), and 95.79% (MPRLPD Datasets) [16]. Gradient Boosting Decision Trees (GBDT) is used for feature selection and after two classification methods, Support Vector Machines (SVM) and GBDT were used for training and testing. According to the results, the Kappa index is almost perfect, and the accuracy is over 90% [17]. Decision Tree, Linear Discriminant, SVM Fine Gaussian, and Logistic Regression algorithms were applied. Based on linear discriminant analysis, the highest prediction accuracy was 95.8%, and ROC was 0.93 [18].

IV. RESEARCH METHODOLOGY

A. Dataset Description

The Indian Liver Patient Dataset (ILPD), which was retrieved from the UCI Machine Learning Repository was used in this study. There were 579 instances of 10 attributes in the sample. Positive and negative tests were conducted on the names of the two target classes. The number of positive target instances was 416, whereas, the number of negative target instances was 167. Age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos are among the 10 variables in this dataset.

B. Result Analysis

This ILPD dataset was given to the CNN algorithm as input. Scikit-learn is an open-source machine learning A library used in the Python programming language. Jupyter Notebook is an open-source web application that makes this possible. Live code, visualisation, Equations and narrated text are used to fulfill our goal.

C. Evaluation Metrics

The evaluation metrics used in DL challenges are critical for achieving the best classifier in four measures: Precision, Recall, F1-score, and accuracy (2) to (4) are the formulae for calculating these measurements.

1) *Precision* is used to compute the positive patterns in an upbeat class correctly, predicted by all predicted patterns.

$$Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \qquad (2)$$

where,

TP True Positive;

FP False Positive.

2) *Recall* is used to compute the percentage of correctly classified positive patterns.

$$Recall = TP/(TP + FP) \qquad (3)$$

where *FN* is False Negative.

3) *F*1-*Score* calculates recall and precision rates.

$$F1 = (2\,x\,Precision\,x\,Recall)/(Precision + Recall) \qquad (4)$$

4) Accuracy calculates the proportion of correctly predicted classes to the total number of samples tested.

$$\text{Accuracy} = (TP + TN)/Total\ Instances \qquad (5)$$

where, *TN* is True Negative.

V. EXPERIMENTAL RESULTS

Table I shows performance evaluation of Convolutional Neural Network. Fig. 3 shows model loss. Table II shows training and simulation error.

TABLE I.
PERFORMANCE EVALUATION OF CONVOLUTIONAL NEURAL NETWORK

Algorithm	Dataset	Accuracy	Recall	Precision	*F*1-score
CNN	Indian Liver Patient Dataset	100%	100%	100%	100%

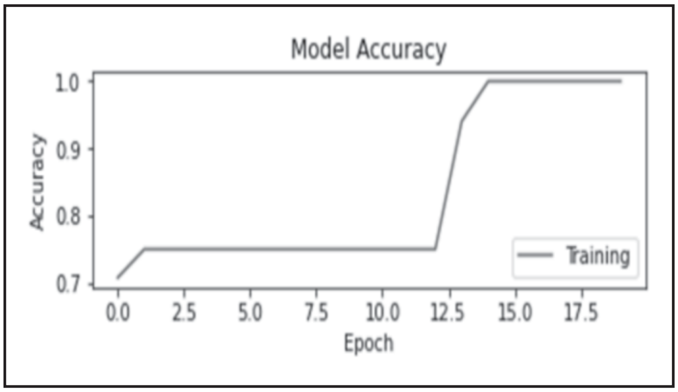

Fig. 2. Model Accuracy

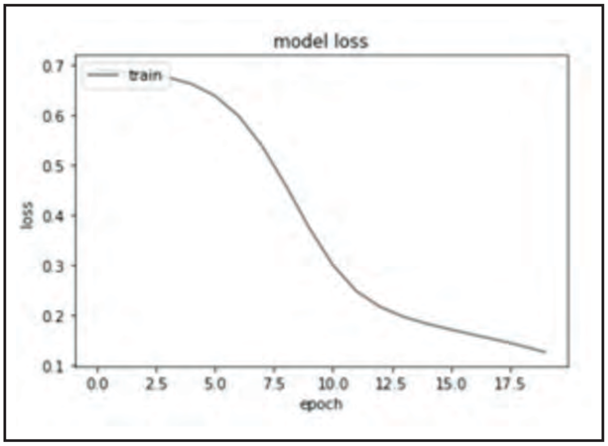

Fig. 3. Model Loss

TABLE II.
TRAINING AND SIMULATION ERROR

Algorithm	Kappa statistics	Mean absoluteerror	Root mean squared error	Median absolute error %	*R*-Squared
CNN	1.0	0.1054	0.1256	0.0921	0.9091

TABLE III.
VALUES CONFUSION MATRIX FOR LIVER CANCER

		Predicted	
		Negative	**Positive**
Actual	Negative	39	0
	Positive	0	135

Fig. 4 shows the ROC curves of our different classifiers in terms of their accuracy.

The ROC curve illustrates the performance of various classification methods. We can select the most suitable models and eliminate others by analyzing the plot. In Table III, each row represents the actual class rates while each column represents predictions since confusion matrices help evaluate classifiers.

VI. DISCUSSION AND CONCLUSION

Disease perdition has long been considered a severe issue. Deep learning algorithms are commonly used to handle this medical care challenge. This type of problem is typically solved using a neural network technique. This study demonstrates CNN's accuracy in predicting cancer. The CNN developed model is more consistent than any

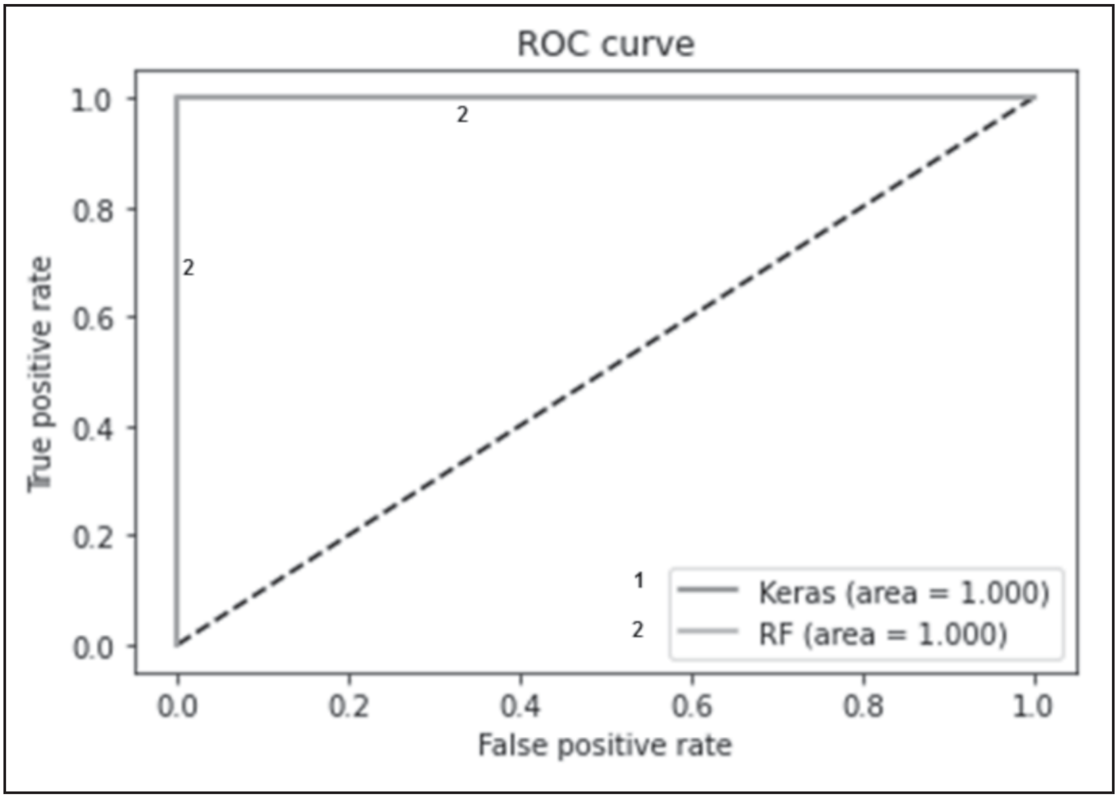

Fig. 4. ROC curve

other technique mentioned, and it has the potential to alter the area of liver cancer prediction. According to the findings, deep learning techniques can detect the disease automatically and accurately.

AUTHORS' CONTRIBUTION

Both the authors actively worked on this paper. Sirajali Nagalpara is the key person and has performed the entirety of the work described in this paper. He prepared the training and test dataset, ran the various models against the data, and analyzed the output. Dr. Bhavesh Patel guided Sirajali Nagalpara in this work.

CONFLICT OF INTEREST

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or material discussed in the manuscript.

FUNDING ACKNOWLEDGEMENT

The authors did not receive any financial support for the research, authorship, and/or for the publication of the article.

REFERENCES

[1] M. B. Priya, P. L. Juliet, and P. R. Tamilselvi, "Performance analysis of liver disease prediction using Machine Learning Algorithms," *Int. Res. J. Eng. Technol.,* vol. 5, no. 1, pp. 206–211, 2018, [Online]. Available: https://www.irjet.net/archives/V5/i1/IRJET-V5I142.pdf

[2] T. R. Baitharu, and S. K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset," *Procedia Comput. Sci.,* vol. 85, pp. 862–870, 2016, doi: 10.1016/j.procs.2016.05.276.

[3] B. M. Ramageri, "Data mining techniques and applications," *Indian J. Comput. Sci. Eng.* vol. 1, no. 4, pp. 301–305. [Online]. Available: http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf

[4] D. Anwanwan, S. K. Singh, S. Singh, V. Saikam, and R. Singh, "Challenges in liver cancer and possible treatment approaches," *Biochim. Biophys Acta Rev. Cancer*, vol. 1873, no. 1, p. 188314, 2020, doi: 10.1016/j.bbcan.2019.188314.

[5] A. Kumar, S. K. Acharya, S. P. Singh, A. Arora, R. K. Dhiman, R. Aggarwal, and A. C. Anand et al. "2019 Update of Indian National Association for Study of the Liver Consensus on Prevention, Diagnosis, and Management of Hepatocellular Carcinoma in India: The Puri II Recommendations," *J. Clin. Exp. Hepatol.,* vol. 10, no. 1, pp. 43–80, 2020, doi: 10.1016/j.jceh.2019.09.007.

[6] K. Nguyen,C. Fokes, A. Ross, and S. Sridharan., "Iris recognition with off-the-shelf CNN features: A deep learning practice," *IEEE Access*, vol. 6, pp. 18848–18855, 2018, doi: 10.1109/ACCESS.2017.2784352.

[7] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaria, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning?: Concepts , CNN architectures, challenges , applications, future directions.," *J. Big Data*, vol. 8, 2021, Art. No. 53, doi:10.1186%2Fs40537-021-00444-8.

[8] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data,"*Int. J. Data Sci. Adv. Analytics*, vol. 1, no. 1, 2019. [Online]. Available: http://ijdsaa.com/index.php/welcome/article/view/2

[9] E. Mamdouh and M. Mabrouk, "A study of support vector machine algorithm for liver disease diagnosis," 2014. [Online]. Available: https://www.researchgate.net/publication/272356715

[10] S. R. Ghosh and S. Waheed, "Analysis of classification algorithms for liver disease diagnosis," *J. Sci. Technol. Environ. Inf.*, vol. 5, no. 1, pp. 361–370, 2017, doi:10.18801, jstei.050117.38.

[11] P. Rajewari and G. S. Reena, "Analysis of liver disorder using data mining algorithm," *Global J. Comput. Sci. Technol.*,vol. 10, no. 14, pp. 48–52, 2010.

[12] P. Ramachandran, N. Girija, and T. Bhuvaneswari, "Early detection and prevention of cancer using data mining techniques," *Int. J. Comput. Appl.*, vol. 97, no. 13, pp. 48–53, 2014, doi: 10.5120/17069-7492.