International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST52310216

Comparison and Analysis of Liver Cancer Prediction Using ML

Dr. J Siva Prashanth[1], Vodapally Sriharsha[2], Salkapuram Sai Chaithanya Teja[2], Mohd Abdul Jabbar[2]

[1]Assistant Professor, Computer Science and Engineering, Anurag group of Institutions, Hyderabad, Telangana, India

[2]B. Tech Student, Computer Science and Engineering, Anurag group of Institutions, Hyderabad, Telangana, India

ARTICLEINFO

Article History:

Accepted: 01 March 2023
Published: 12 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
192-198

ABSTRACT

In diagnosis centers the importance of detecting a cancer on time is vital. With the help of tools like x-ray, MRI machine, medical professionals can detect somatic mutations easily (a somatic mutation is an acquired change in a genetic code of one or more cells). Here we chose a disease that is liver cancer. This deployed model is given data via google Collab, then analyzed in real-time with machine learning model which was pretrained and the result is shown in the google Collab. Models that are used in our project are Logistic regression, Naive Bayes classifier and Random Forest etc., is used to carry out computation for prediction. And we compare these machine learning models accuracies. But we got good accuracy for machine learning model (random forest classifier). Early detection can help in identifying the risk of liver cancer. Our model is helpful for doctors to give timely medications for treatment.

Keywords: Machine Learning, Logistic Regression, Naive Bayes, Random Forest.

I. INTRODUCTION

Millions of individuals all around the world are afflicted by liver cancer, a serious and frequently fatal condition. Early detection and diagnosis of liver cancer are crucial for improving patient outcomes and increasing the chances of successful treatment. Machine learning models have shown great promise in predicting the occurrence of liver cancer based on various input features such as age, sex, alcohol consumption, hepatitis B/C status, and liver function tests.

In this context, the use of machine learning algorithms such as Logistic Regression, Gaussian Naive Bayes, and Random Forest can provide accurate and reliable predictions of liver cancer occurrence. These algorithms are widely used in machine learning for classification and prediction tasks and have demonstrated excellent performance in a variety of domains.

This approach involves training the machine learning models using historical data of patients diagnosed with liver cancer, and then testing the models on a

separate dataset to evaluate their accuracy and predictive power. By analyzing the relationship between the input features and the occurrence of liver cancer, the models can predict the likelihood of a patient developing liver cancer in the future.

This paper seeks to give a general overview of the use of the Random Forest, Gaussian Naive Bayes, and Logistic Regression algorithms for forecasting the development of liver cancer. We will discuss the methodology, tools, and technologies used for the development of these models, as well as their performance and potential future enhancements. In order to enhance patient outcomes and potentially save lives, this work aims to contribute to the creation of precise and trustworthy machine learning models for forecasting liver disease.

II. LITERATURE SURVEY

The authors of the paper "DNorm complaint name normalisation with pairwise literacy to rank," Zhiyong Lu, Rezarta Islamaj Dogan, and Robert Leaman, proposed it. In this essay, we provoke Contrary to other normalising jobs in biological textbook mining exploration, there have been significant but lesser attempts in identifying the conditions that are stated in the textbook. In this composition we are combining MeSH and OMIM. Our framework for finding similarities from data is highly effective and depended on mathematical principles. The approach is depended on literacy to order, which is not applied in the issue of normalisation but is successfully solved significant optimization challenges for data reclamation. Our approach is grounded to speech normalisation and corresponding, it outperforms the highest performing system.[2]

The affiliations of Richard H. Scheuermann, Werner Ceusters and Barry Smith dissected the study "Toward an ontological treatment of complaint and opinion". In this investigation, many biomedical language norms are shown to be based on incomplete, contradictory, or muddled definitions of terms that refer to medical problems. This framework illustrates related realities with their relationships. We maintain the idea that complaints always stem from a physical foundation which carries a propensity for the advancement of processes. In order to provide a unified foundation, we present our viewpoint as a glossary of terminology and definitions. [5]

The study by Kate M. Dunn, Peter Croft, Harry Hemingway, Jonathan J. Deeks, Douglas G. Altman and Alastair D. Hay, among others, was based on the article "The wisdom of clinical practise complaint opinion or case prognostic? Evidence about "what is likely to be" ought to guide clinical treatment. In clinical practise, background opinion has historically served as the basis for decision-making. The benefits and harms to unborn children of these viewpoints for instances evaluated with and without complaint typically lack supporting evidence. We suggest that a clinical practise approach that focuses prognosis and forecasting fetal problems effectively. Discussions of complaint opinions can give data to judge and determine the course of significant acute disease. The regular usage of specific pointers with nonstop divisions, akin to glucose levels that are recognized in providing data of unborn outgrowth, challenges the complaint as a "yes" or "no". Moreover, ailments like habitual weariness, unables to accurately classify in view of complaint-opinion. Such situations call for the use of an anticipating model, which extends a crucial foundation for medical practise that surpasses complaint and perspective and considers a huge variety of facts to prognosticate further patient concerns and direct opinions towards their resolution. Similar information includes inheritable, non-disease variables, and other biomarkers that affect development. Patient prognosis provides the skeleton for cutting-edge scientific practise by integrating data via constantly amplifying scientific and medical databases for methodical care. [6]

The authors of the study "Disease vaticination with different types of neural network classifiers" are Ruo-Ping Han, Tony Cheng- Kui Huang, and Cheng-Hsiung Weng. Complaint vaticination has been regarded as a crucial component in this essay. In the past, methods for breaking this kind of medical care difficulty have been created using artificial intelligence and mechanical literacy. Recently, neural network combinations have been used successfully in several activities, including those that support medical judgement. The conceptualization potential of learning systems can be considerably improved by neural network accumulates by training limited neural networks and integrating outputs. Yet, it is currently unclear how well different classifiers perform in complaint vaticination. This study's main goal intends to examine the functioning of several classifiers, like solo classifiers working in an association. Additionally, we assess the effectiveness of these classifiers using eye-catching evaluation criteria using real-world datasets. Finally, we quantify the importance of the performance difference between the three classifiers using statistical testing. Yet, when built using a similar size training dataset, the solo classifier does not function as poorly as the ensemble classifier. [11]

III.PROPOSED SYSTEM

The proposed system includes liver cancer test values in the dataset consisting of input features like direct and indirect bilirubin, age of the patient and liver enzymes like ALB albumin, sgot aspartate aminotransferase, alkaline phosphatase, total proteins, sgpt alamine aminotransferase, globulin ratio albumin and a/g ratio and one output variable is result (0: not liver cancer,1: liver cancer).

In our work we shown the comparison of machine learning models using different algorithms with liver cancer dataset available in Kaggle website. From these algorithms we have chosen the best accuracy model as our final model. Logistic Regression, Gaussian Naive Bayes, and Random Forest Classifier are the algorithms that we have utilized.

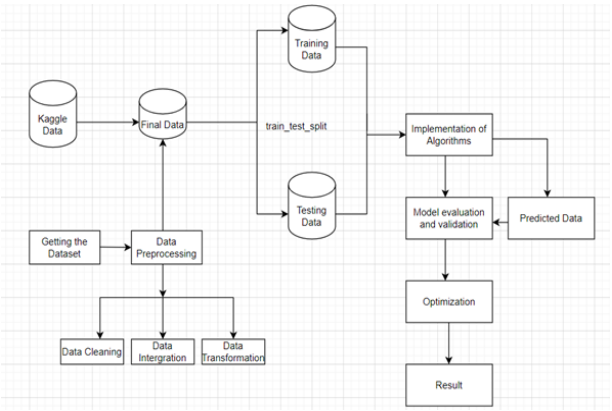

Fig 3.1 System Architecture

IV. IMPLEMENTATION

Logistic Regression:

It's statistical technique for binary categorization involves predicting a binary outcome, such as whether a patient has a disease or not. In logistic regression, the goal is to estimate the probability of the binary outcome given a set of input variables or features. Logistic regression works by fitting a logistic function to the input features. The logistic function, also referred to as the sigmoid function, accepts any input value and produces a result 0 or 1.

It is widely used in various applications, including medical diagnosis, credit scoring, and marketing analytics. It is a popular method because it is relatively simple, interpretable, and computationally efficient. On the other hand, it assumes that the input features and the log odds of the binary result are linearly related, and may not perform well when this assumption is broken. Additionally, logistic regression is limited to binary classification and may not be suitable for multi-class classification problems.

Gaussian Naive Bayes:

It is frequently employed for classification jobs. It is founded on the Bayes probability theorem. In

Gaussian Naive Bayes, it is supposed that the likelihood of the aspects given the class labels follows a Gaussian (normal) distribution. This means that the features are assumed to be continuous and the probability density function of each feature given a class label is modelled as a Gaussian distribution.

Using the training data, the method first calculates stastistical entities for each class label. The algorithm uses the Bayes theorem and the estimated entities to determine the likelihood of each class label given a fresh input sample with feature values. The output is then expected to be the label with highest possibility. The "naive" presumption is that given the class label, the characteristics are independent. This simplifies the probability calculation and makes the algorithm computationally efficient, but may not always hold true in practice.

Random Forest:

Several decision trees are combined in this ensemble learning technique to produce predictions. Multiple trees are created which are trained using a different subdivisions of input features and learning data, and their projections are then merged to provide a result. Any individual subgroup of input factors and learning data is used to train respective decision tree in the Random Forest technique. This improves the model's capacity for generalization while reducing overfitting. During training, each tree recursively splits the input space into smaller regions based on the values of the input features, and assigns a label or value to each region. The split points for each tree are chosen based on the feature that maximizes the reduction in impurity or error between the predicted and actual labels/values.

Each tree in the Random Forest algorithm independently predicts the label or value based on the input features to make a prediction for a new input. The result is then obtained by considering the highest votes or the average of all the trees' predictions. The great accuracy and resistance to noise and outliers in the data of Random Forest are well known. It can handle both categorical and numerical input features, and can capture complex nonlinear relationships between the features and the output. Additionally, Random Forest provides measures of feature importance, which can be useful for feature selection and interpretation. Many applications, including image analysis, bioinformatics, and finance, frequently use Random Forest. However, it may not perform well on tasks with highly imbalanced data or when the data contains significant outliers or missing values.

V. RESULTS

```
Logistic Regression Training Score:
 72.06
Logistic Regression Test Score:
 68.0
Coefficient:
 [[-0.00994993 -0.09851216 -0.30688722 -0.00082939 -0.01078829 -0.00275598
  -0.23899671  0.40208933  0.59475502  0.25335289  0.09115993]]
Intercept:
 [0.36100671]
Accuracy:
 0.68
Confusion Matrix:
 [[107  17]
 [ 39  12]]
Classification Report:
              precision    recall  f1-score   support

           1       0.73      0.86      0.79       124
           2       0.41      0.24      0.30        51

    accuracy                           0.68       175
   macro avg       0.57      0.55      0.55       175
weighted avg       0.64      0.68      0.65       175
```

Fig 5.1 Logistic Regression with accuracy 68%

```
Gaussian Score:
 56.13
Gaussian Test Score:
 53.14
Accuracy:
 0.5314285714285715
[[44 80]
 [ 2 49]]
              precision    recall  f1-score   support

           1       0.96      0.35      0.52       124
           2       0.38      0.96      0.54        51

    accuracy                           0.53       175
   macro avg       0.67      0.66      0.53       175
weighted avg       0.79      0.53      0.53       175
```

Fig 5.2 Gaussian Naive Bayes with accuracy 53.14%

```
Random Forest Score:
 100.0
Random Forest Test Score:
 70.29
Accuracy:
 0.7028571428571428
[[102  22]
 [ 30  21]]
              precision    recall  f1-score   support

           1       0.77      0.82      0.80       124
           2       0.49      0.41      0.45        51

    accuracy                           0.70       175
   macro avg       0.63      0.62      0.62       175
weighted avg       0.69      0.70      0.69       175
```

Fig 5.3 Random Forest with accuracy 70.29%

	Model	Score	Test Score
2	Random Forest	100.00	70.29
0	Logistic Regression	72.06	68.00
1	Gaussian Naive Bayes	56.13	53.14

Fig 5.4 Model Evaluation

VI. CONCLUSION

The prediction of liver cancer has shown satisfactory result using machine learning models including Logistic Regression, Gaussian Naive Bayes, and Random Forest. These models have been used to predict the occurrence of liver cancer based on a variety of input features such as age, gender and liver function tests.

A common approach for issues involving binary classification, such as determining whether liver cancer will develop or not, is logistic regression. It models the probability of the binary outcome as a function of the input features. Based on the Bayes theorem and assuming that the features are random and distributed normally, Naive Bayes is a probabilistic method. It is particularly useful when dealing with high-dimensional datasets. Many decision trees are combined in Random Forest, an ensemble learning approach, to increase the predictors' robustness and accuracy.

Studies have shown that these models can achieve high accuracy in predicting liver cancer, with Random Forest generally outperforming the other two algorithms. Yet, the quality and quantity of the input data might have an impact on how well these models work, as well as the specific parameters and settings used in each algorithm.

Overall, the development of accurate and reliable machine learning models for predicting liver cancer holds great potential for early detection and improved patient outcomes. Further research and development of these models will be essential to ensure their clinical relevance and usefulness in real-world settings.

VII. FUTURE ENHANCEMENT

Future machine learning improvements to prediction models for liver cancer will necessitate a multidisciplinary strategy that combines clinical, biological, and computational expertise. The precision and robustness of the models used to predict liver disease can be raised by integrating many varieties of omics data, including genomics, transcriptomics, and proteomics. To aggregate the predictions of various machine learning models and increase accuracy, ensemble learning techniques can be utilized.

VIII. REFERENCES

[1]. J. L. Scully, "What is a disease?" EMBO Rep., vol. 5, no. 7, pp. 650–653, 2004.

[2]. R. Leaman, R. Islamaj Dogan, and Z. Lu, "DNorm: Disease name normalization with pairwise learning to rank," Bioinformatics, vol. 29, no. 22, pp. 2909–2917, Nov. 2013.

[3]. N. Armstrong and P. Hilton, "Doing diagnosis: Whether and how clinicians use a diagnostic tool of uncertain clinical utility," Social Sci. Med., vol. 120, pp. 208–214, Nov. 2014.

[4]. A.-L. Bar abasi, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," Nature Rev. Genet., vol. 12, no. 1, pp. 56–68, Jan. 2011.

[5]. R. H. Scheuermann, W. Ceusters, and B. Smith, "Toward an ontological treatment of disease and

diagnosis," Summit Transl. Bioinformat., vol. 2009, p. 116, Mar. 2009.

[6]. P. Croft, D. G. Altman, and J. J. Deeks, "The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about 'what is likely to happen' should shape clinical practice," BMC Med., vol. 13, no. 1, p. 20, 2015.

[7]. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in Proc. Mach. Learn. Healthcare Conf., 2016, pp. 301–318.

[8]. C. C. Lee, "Fuzzy logic in control systems: Fuzzy logic controller. I,"IEEE Trans. Syst., Man, Cybern., vol. 20, no. 2, pp. 404–418, Mar./Apr. 1990.

[9]. J. Yen and R. Langari, Fuzzy Logic: Intelligence, Control, and Information, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[10]. H. D. Beale, H. B. Demuth, and M. Hagan, Neural Network Design. Boston, MA, USA: PWS, 1996.

[11]. C.-H. Weng, T. C.-K. Huang, and R.-P. Han, "Disease prediction with different types of neural network classifiers," Telematics Inform., vol. 33, no. 2, pp. 277–292, 2016.

[12]. K. Sumeet, J.J. Larson, B. Yawn, T.M. Therneau, W.R. Kim, Underestimation of liver-related mortality in the United States. Gastroenterology;(2013) 145:375–382, e371–372.

[13]. A.A. Mokdad, A.D. Lopez, S. Shahraz, R. Lozano, A.H. Mokdad, J. Stanaway, et al, Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. BMC Med 2014; 12:145.

[14]. Byass, Peter, The global burden of liver disease: a challenge for methods and for public health. BMC medicine 12.1 (2014); 159.

[15]. L. A. Auxilia, Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE (2018).

[16]. Hashem, M. Esraa, S. Mai, A study of support vector machine algorithm for liver disease diagnosis. American Journal of Intelligent Systems 4.1 (2014); 9-14.

[17]. P. Sajda, "Machine learning for detection and diagnosis of disease." Annu. Rev. Biomed. Eng. 8 (2006); 537-565.

[18]. UCI Machine Learning Repository. ILPD (Indian Liver Patient Dataset) Data Set. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[19]. Logistic Regression, retrieve from: HTTPS://WWW.SAEDSAYAD.COM/LOGISTIC_REGRESSION.HTM, LAST Accessed: 5 Octobor,2019

[20]. L. Breiman, Random Forests. Machine Learning, 45(1), (2001); 5–32. https://doi.org/10.1023/A:1010933404324

[21]. Decision Trees, retrieve from: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/, Last Accessed: 5 Octobor,2019

[22]. Support vector machine, retrieve from: http://www.statsoft.com/textbook/support-vector-machines, Last Accessed: 5 Octobor,2019

[23]. V. Vapnik, I. Guyon, T. H.-M, Learn, and undefined 1995. Support vector machines. statweb.stanford.edu (1995).

[24]. Zhang M, Zhou Z, "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition40.7:(2007); 2038-2048.

[25]. G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification (pp. 986–996). Springer, Berlin, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39964-3_62

[26]. Naive Bayes, retrieve from: https://www.geeksforgeeks.org/naive-bayes-classifiers/, Last Accessed: 5 Octobor,2019

[27]. S. M. Mahmud, et al. "Machine Learning Based Unified Framework for Diabetes Prediction."

Proceedings of the 2018 International Conference on Big Data Engineering and Technology. ACM (2018).

[28]. S. Safavian, D. Landgrebe, A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), (1991); 660-674.

[29]. A. Chervonenkis, Early history of support vector machines. In Empirical Inference (pp. 13-20). Springer, Berlin, Heidelberg (2013).

[30]. K.M. Leung, Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering (2007).

Cite this article as :

Dr. J Siva Prashanth, Vodapally Sriharsha, Salkapuram Sai Chaithanya Teja, Mohd Abdul Jabbar, "Comparison and Analysis of Liver Cancer Prediction Using ML ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 2, pp. 192-198, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRST52310216
Journal URL : https://ijsrst.com/IJSRST52310216