IMPROVING THE PREDICTION ACCURACY OF LIVER CANCER DETECTION USING MACHINE LEARNING TECHNIQUES

J Gold Beulah Patturose, K Agnes Prema Mary, C. Sheeba Thangapushpam, F V Jayasudha,

St.Joseph's Institute of Technology ,Chennai, Jayaraj Annapackiam CSI College of Engineering, Nazareth, Malla Reddy University, Hyderabad ,Sathyabama Institute of Science and Technology,Chennai

beulahpattu@gmail.com, nestamary@gmail.com, csheebathangapushpam@gmail.com,sudhafrance2@gmail.com

Abstract.

The liver is an essential organ that plays a crucial role in various bodily functions,digestion, waste elimination, and energy regulation. Liver diseases, particularly cancer, pose significant health risks, making early detection crucial for effective treatment. Traditional diagnostic methods are often costly and complex, prompting the adoption of machine learning (ML) techniques for classification. This study evaluates the performance of various ML algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), using the Indian Liver Patient Dataset (ILPD). To enhance predictive accuracy, the XGBoost boosting algorithm is applied. Experimental results demonstrate that SVM, when combined with boosting, achieves the highest accuracy of 92.5%. These findings underscore the potential of ML in improving the diagnosis of liver diseases.

Keywords Random forest ,Machine Learning, Decision tree, Liver Disease, Classification, Support vector machine ,Logistic regression,

1 INTRODUCTION

The largest bodily organ, the liver, carries out essential processes like metabolism, bile generation, and detoxification.. It has a unique ability to regenerate within 8 to 15 days if at least 25% of its tissue remains [1]. Cirrhosis, hepatitis, and liver cancer are among the liver illnesses that can result from genetics, infections, alcohol use, obesity, and diabetes. Jaundice, exhaustion, black urine, and stomach ache are typical symptoms. [2]. Around 50 million people worldwide suffer from liver diseases, causing 2 million deaths annually [2]. Early detection through blood tests, imaging, and biopsies is crucial but challenging [3]. Machine learning (ML) aids in diagnosing liver diseases by analyzing complex medical data efficiently [4]. Prediction accuracy is increased by methods like Decision Tree ,Logistic Regression.,Random Forest and Support Vector Machine, [4]. XGBoost further enhances performance by refining classification results [4]. ML-based liver disease prediction reduces diagnostic time and assists medical professionals in early intervention [4].

2. Literature Review

Data mining has been widely used in disease prediction. The authors of [5] applied Decision Tree, Naïve Bayes, and Rule-Based Algorithms for diabetes prediction, enhancing accuracy through feature extraction. Various classification techniques, including SVM, KNN, and LR, were used for liver disease prediction, with LR showing superior sensitivity.

The authors of [6] explored unsupervised classification methods such as Affinity Propagation, DBSCAN, and K-Means, using the Silhouette Coefficient for evaluation. K-Means performed best, and future research was suggested for heart, lung, and brain disease prediction.

In [7], Genetic Algorithms improved feature selection, enhancing SVM, Naïve Bayes, and KNN classification accuracy. A hybrid Decision Tree-Clustering approach further increased precision. The study in [8] combined Genetic Algorithms with K-Means, refining cluster centers and improving accuracy.

Outlier detection methods, such as Training-based KNN, KNN, and Local Outlier Factor, have been used to ensure better classification outcomes. In [10], six classifiers were tested, with Bagging Classifier achieving 69.3% accuracy. The Complement Naïve Bayesian (CNB) model further improved accuracy to 71.36%.

Rahman et al. [10] found LR to be the most accurate (75%) among RF, KNN, DT, SVM, and NB. In [11], RStudio-based C5.0 classifier achieved 75.19% accuracy on ILPD. Auxilia [10] introduced XGBoost with Genetic Algorithm-based SMOTE (GASMOTE), achieving 88% accuracy using RF.
Principal Component Analysis (PCA) [10] has also been applied to ILPD data, followed by classification using Naive Bayes, KNN, SVM, DT, and RF. Adaptive Boosting on SVM yielded the highest accuracy of 89.1%. These studies highlight the importance of data preparation for improved liver disease prediction.

3. METHODOLOGY AND CLASSIFICATION

The machine learning pipeline begins with data preprocessing, which includes handling categorical data (e.g., encoding features), balancing imbalanced classes (e.g., oversampling), and normalizing numerical values (e.g., scaling). The refined dataset is then partitioned into training (80%) and testing (20%) subsets. Classification models are trained on the former and validated on the latter to evaluate generalization capability.

3.1 Logistic Regression [LR]

A classification system called logistic regression predicts probabilities between 0 and 1 using the sigmoid function. It is mainly used for binary classification, assigning labels based on a threshold of 0.5. Despite its name, it is a classification method derived from linear regression. Logistic Regression has three types: binomial (two outcomes), multinomial (multiple unordered categories), and ordinal (ordered categories). Its ease of use, effectiveness, and interpretability make it popular. This makes it a well-liked option for a number of practical uses.

3.2 Decision Tree (DT)

A DT splits data hierarchically based on feature thresholds, creating a tree-like structure. Internal nodes represent decision rules (e.g., “Bilirubin > 1.2”), while leaf nodes denote final predictions. To minimize overfitting, criteria like entropy reduction guide splits, and post-pruning removes redundant branches.

3.3 Random Forest (RF)

Random Forest is a supervised learning algorithm that enhances accuracy by combining multiple decision trees. To enhance predictions, it employs regression averaging and majority voting for classification. This method reduces overfitting and increases model reliability.

3.4 Support Vector Machine (SVM)

Decision trees are supervised learning models that classify or predict outcomes by splitting data based on feature values. They consist of decision nodes for making choices and leaf nodes for final predictions. While easy to interpret, they can overfit, so pruning is often needed for better accuracy.

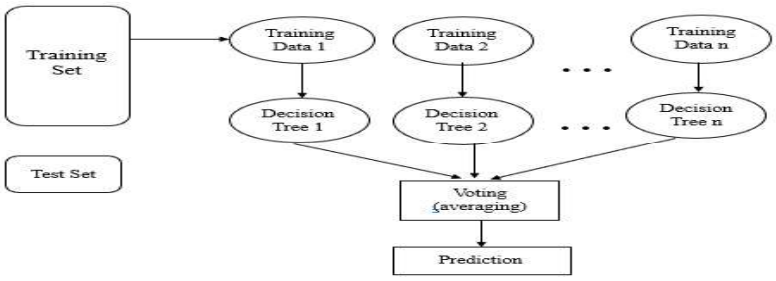

Figure 3.2 Steps followed by Random Forest

[32] Figure 3.2 shows how Random Forest works by employing random subsets of data and attributes to create several decision trees. This randomness helps reduce overfitting and enhances prediction accuracy. The final output is determined by averaging (for regression) or majority voting (for classification), ensuring more reliable results.

4 RESULTS AND DISCUSSION

4.1 Dataset Description

There are 142 women and 441 males in the dataset from northeast Andhra Pradesh, India, which includes 416 liver patient records and 167 non-liver patient records. The **class label "Dataset"** distinguishes liver patients from non-patients**. Google Colab [15]** is used for implementation, with **10-fold cross-validation** ensuring performance evaluation without overfitting. The dataset is split **80:20 for training and testing**, and **LR, DT, RF, and SVM** are compared for classification. These models are applied after data preprocessing, without feature selection.

4.2 Performance Evaluation

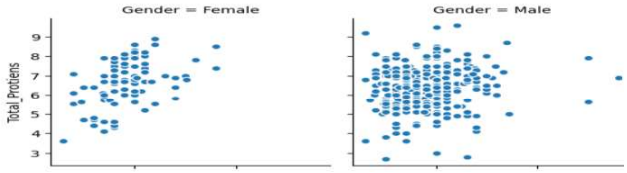

Figure 4.1 Total Protein Vs Gender

Figure 4.1 shows the effect of proteins according to gender. The amount of protein also differs according to the gender. The above figure gives the distribution of dataset among the male and female gender.

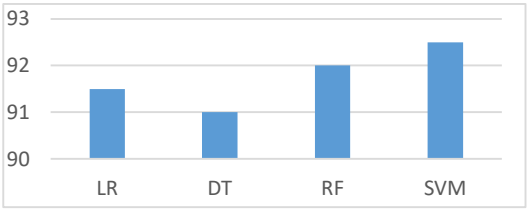

Figure 4.2 Performance analysis

To evaluate the effectiveness of the proposed models, multiple evaluation metrics were employed, with accuracy serving as the primary indicator of predictive performance. Accuracy measures the ratio of correctly classified instances to the total predictions made by the model. As depicted in Figure 4, which contrasts the performance of the four algorithms, the Support Vector Machine (SVM) emerged as the most effective classifier, surpassing the other models with the highest accuracy score. This outcome underscores SVM's robustness in handling the dataset's complexity and variability.

5. CONCLUSION

As a vital organ, the liver is integral to processing nutrients, regulating energy, and eliminating toxins. Early identification of liver disorders is critical to mitigating severe health outcomes, including liver cancer. Machine learning (ML) techniques offer a promising solution for enhancing the accuracy of liver disease detection. In this research, the efficacy of four ML algorithms, Logistic Regression (LR), Support Vector Machine (SVM), , Decision Tree (DT), and Random Forest (RF) was evaluated using the Indian Liver Patient Dataset (ILPD). The findings demonstrate that SVM, when combined with the XGBoost algorithm, achieved the highest classification accuracy of 92.5%, outperforming other models.

6. REFERENCES

1. *M. F. Rabbi, S. M. Mahedy Hasan, A. I. Champa, M. AsifZaman and M. K. Hasan, "Prediction of Liver Disorders using Machine Learning Algorithms: A Comparative Study," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 111-116, doi: 10.1109/ICAICT51780.2020.9333528*
2. *Revathi, M. & Raghuraman, G.. (2023). Kidney Stone Detection from CT Images using ALEXNET and Hybrid ALEXNET-RF Models. Journal of Circuits, Systems and Computers. 33. 10.1142/S021812662450107X.*
3. *N. M. Dipu, S. A. Shohan and K. M. A. Salam, "Deep Learning Based Brain Tumor Detection and Classification," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-6, doi: 10.1109/CONIT51480.2021.9498384.*
4. *J. K. Periasamy, B. S and J. P, "Comparison of VGG-19 and RESNET-50 Algorithms in Brain Tumor Detection," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126451.*
5. *V. Ramana and R. S. Kumar Boddu, "Performance Comparison of Classification Algorithms on Medical Datasets," 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2019, pp. 0140-0145.*
6. *Anagaw, Y.L. Chang, "A new complement naïve Bayesian approach for biomedical data classification," Journal of Ambient Intelligence and Humanized Computing, vol. 10, pp. 3889- 3897, 2019.*
7. *Rahman, A. K. M. & Shamrat, F.M. & Tasnim, Zarrin & Roy, Joy & Hossain, Syed, "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms," International Journal of Scientific & Technology Research, vol. 8, pp. 419-422, 2019.*
8. *Veena, G. & Sneha, D & Basavaraju, Deepti & Tanvi, Tripti, "Effective Analysis and Diagnosis of Liver Disorder," International Conference on Communication and Signal Processing, April 3-5, 2018, pp. 86-90.*
9. *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), Tirunelveli, 2018, pp. 45-50.*
10. *Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury, "Prediction of Liver Disease using Classification Algorithms", 4th International Conference on Computing Communication and Automation (ICCCA)(2018).*
11. *Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining B.V. Baiju Department of Information Technology Hindustan Institute of Technology and Science.*
12. *N.Ramkumar, S. Prakash, S. Ashok Kumar, K Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem" published in 2017 International Conference on Computer Communication and Informatics (ICCCI).*
13. *R. Saravanan and Pothula Sujatha , "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification" in Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018).*
14. *https://www.kaggle.com/datasets/uciml/indian-liver-patient-recordsPushpendra Kumar and Ramjeevan Singh Thakur, " Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN Method for LFT Imbalanced Data" in IEEE 6th International Conference on smart structures and systems ICSSS 2019*

Biographies

J Gold Beulah Patturose received bachelor degree in Electronics and Communication Engineering in 2010 and master degree in Computer Science and Engineering from, Anna University, Chennai, India, in 2013. She is currently pursuing the Ph.D degree from the Anna University ,Chennai, and also she is working as an Assistant Professor, St. Joseph's Institute of Technology, Chennai, Tamilnadu, India

Dr. K. Agnes Prema Mary is working as a Professor, Jayaraj Annapackiam CSI College of Engineering, Tamil Nadu, India .She received BE-EEE from, MS University, Chennai, India in 1998 ME-Power Electronics and Drive from, Anna University, Chennai, India, in 2007 and doctorate in the field of Electrical Engineering from Anna University ,Chennai

C. Sheeba Thangapushpam is working as a Assistant Professor Malla Reddy University, Hyderabad. She received BE-ECE from, Anna University, Chennai, India in 2012 ME-Embedded System from, Anna University, Chennai, India, in 2014. She is currently pursuing the doctorate in the field of Image & Signal Processing

Dr.F.V.Jayasudha is currently working as an Associate Professor in the Sathyabama Institute of Science & Technoogy, Chennai. She obtained her Bachelor degree in B.E – ECE from Aadhiparasakthi college of Engineering, Master's degree in M.E- Applied Electronics and doctorate in the field of image processing at Sathyabama Institute of Science & Technology