Machine Learning Model for Predicting Hepatocellular Carcinoma in Hepatitis C Patients

Arooj Fatima
Department of Science and Engineering
Southampton Solent University
Southampton, United Kingdom
aroojfatima497@outlook.com

Jarutas Andritsch
Department of Science and Engineering
Southampton Solent University
Southampton, United Kingdom
jarutas.andritsch@solent.ac.uk

***Abstract*— An estimated 58 million people suffer from chronic Hepatitis C virus around the world, while substantial evidence indicates that patients with Hepatitis C virus are at 17 times larger risk of developing Liver Cancer (Hepatocellular Carcinoma). Research has been carried out to predict hepatitis C and liver cancer at different stages in patients. In this research, we proposed the Classification model AdaBoost with Decision Tree as its base model to be trained and tested on patient dataset. The dataset contains records of clinical indicators and was acquired from the University of California Irvine Machine Learning Repository. The preparation of the dataset was done using balancing techniques i.e. SMOTE, it was encoded using Ordinal Encoding. The hyperparameters of AdaBoost model was tuned manually to find the most optimal combination. AdaBoost Classification Model achieved a 92.98% accuracy, and the AUROC of 0.97. The precision and recall differ for each class, "healthy" individuals were classified with a precision of 98% and a recall of 99% while patients with "Cirrhosis" (irreversible scarring of liver due to a tumor) were classified with 86% precision and 67% recall. The research concluded that machine learning has efficient applications in predicting diseases. Moreover, the clinical indicators mentioned in previous studies have proven to be vital in the prediction of HCC (liver cancer) however it is advised that, in future a larger dataset may be acquired to overcome any potential biases in the predictions. The current program successfully distinguishes between patients at different stages of HCC and HCV and can be further adapted to build decision systems to aid diagnosis.**

Keywords*— *Hepatitis C, Hepatocellular Carcinoma, Machine Learning, SMOTE, Classification Models, AdaBoost

I. INTRODUCTION

Hepatitis C is a virus (HCV), which causes the liver to be inflamed and infected. The World Health Organization [1] reported that 58 million people around the world have chronic Hepatitis C, and in 2019 an approximate 290,000 mortalities from Hepatitis C had also been diagnosed with Cirrhosis and Hepatocellular Carcinoma (HCC), the most common type of primary liver cancer. "Hepatocellular carcinoma risk increases to 17-folds in Hepatitis C Virus (HCV) infected patients compared to HCV-negative patients" [2]. Reference [3] found overwhelming epidemiological evidence indicating that patients with Hepatitis C virus are at a larger risk of developing Hepatocellular carcinoma (HCC). Most of those suffering from HCV belong to developing countries, around 2-5% of general population in the subcontinent of India and the Middle East is chronically infected from HCV [4]. While a staggering 80% of those diagnosed with HCC reside in low-income regions such as South-eastern Asia and sub-Saharan Africa. HCC has the shortest survival time among all types of cancers, and a prognosis is especially poorer in low-income countries, accredited to 2 main factors, a severe lack of appropriate resources and diagnosed when the tumour has advanced to its final stages [5].

Further research suggests that an early diagnosis of Hepatocellular carcinoma can in-fact increase the 5-year survival rate to more than 70%, while 60% diagnoses are carried out after the metastasis has occurred, resulting in the 5-year survival rate to decrease to less than 16% [6]. This timely diagnosis is especially a challenge in low-income regions such as several countries in sub-Saharan Africa, where the ratio between doctors and patients is 1:10,000, as opposed to the UK where there is 1 doctor for every 28 patients [7]. To achieve accuracy in diagnosis, HCC must first be understood. Hepatocellular carcinoma has 4 stages Hepatitis, Fibrosis, Cirrhosis, and Liver Failure [8].

The uprise in HCV in poorer demographics forces research to be carried out to raise awareness in its development into HCC, and find ways to minimise fatality rates. This research focuses on HCV in particular, as there is no effective vaccine against Hepatitis C [1], while Hepatitis B can be prevented via timely vaccinations [9]. Therefore HCV patient monitoring and prevention of HCC becomes vital.

Hepatitis C Virus causes the liver tissue to be damaged, to heal itself the Hepatic Stellate Cells are activated within the liver, along with excessive production of extracellular matrix, which causes scar tissues to be formed, also known as Liver Fibrosis [10]. Fibrosis then turns into Cirrhosis or permanent scarring, if left untreated for a longer period and becomes irreversible [8]. Patients with Cirrhosis have a greater risk of developing Liver Cancer as the scarred tissue restricts the flow of blood in the liver [11]. Liver cancer can be detected in Hepatitis C patients through several different ways, such as blood tests, ultrasound, CT or MRI scans and a liver cell biopsy [12]. To tackle the lack of resources and the dire need of an early diagnosis of HCC in HCV patients living in low-income countries, an AI based diagnosis system could be implemented to help classify HCV patients into different cancer stages.

The remainder of this paper is structured as follows: Section II reviews related work on utilizing machine learning to predict hepatocellular carcinoma and the clinical indicators used in diagnosing HCC. Section III details our methodology, including data collection and machine learning techniques. Section IV presents the results, and Section V concludes with a summary of findings and future research directions.

II. RELATED WORK

A wide range of research has been carried out in how machine learning can aid decision making, diagnosis and detection of several diseases, especially in HCV patients and the prognosis of HCC.

Reference [13] proposed a hybrid prediction model approach to classify Hepatitis C patients precisely into categories based on the progression of liver fibroses. The models combined were Support Vector Machine and Random Forest. The performance of the hybrid model was evaluated using confusion matrix, the proposed model achieved an accuracy of 41.541% and a recall rate of 40.556%, the precision was 41.23% and the F-measure of 42.332%. The model's performance has been enhanced using the Synthetic Minority Oversampling Technique (SMOTE). SMOTE helps overcome imbalanced data by selecting only essential features from the dataset and generating further occurrences to aid the minority class. The proposed hybrid model gained efficiency once provided with a more balanced data, the precision rose up to 98.0%, accuracy to 96.8%, Recall moved up to 99.1% and F-measure became 97.5%. This study emphasizes the use of balancing techniques as well as showing a distinctive difference in the existing machine learning models and the proposed Hybrid Predictive Model. Reference [14] carried out a study using the UC Irvine HCV dataset to emphasize the importance of balancing and data mining techniques, to classify individuals to HCV and HCC categories. Their method included training and testing 6 different classification models, such as Random Forest, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree and Logistic Regression. The results have been evaluated using AUROC, accuracy, Recall and Precision. The results were observed after the application of different techniques such as standardization and balancing. Random Forest outperformed all the models with an accuracy of 97.29%, AUROC of 0.998, while other models achieved AUROC between 0.896 to 0.972, and accuracy remained between 92.43% to 96.75%. This study contradicts the research carried out by [13] and proves that correct data mining techniques and simpler machine learning models can lead to fruitful results, instead of assembling models. Reference [15] investigated age, alpha-fetoprotein, albumin, total bilirubin, and alkaline phosphate in 4423 Chronic Hepatitis C related patients. They compared the evaluations of Decision Tree, multi-linear regression, reduce pruning error tree and the CART tree algorithm and concluded that the models with simple features achieve high accuracies such as Decision Tree, which achieved the highest accuracy of 95.6%. Reference [16] compared findings from 6 already existing HCC prediction models on over 2500 records of patients with Cirrhosis and cured HCV. The data was collected from 2 different cohorts, the Scotland HCV clinical database and Stratified medicine to Optimize Treatment of HCV. The patient values for 4 competing non-genetic HCC prediction models were calculated as well as 2 genetic models. Models included aMAP, VHA, THRI, ANRS CO12 CirVir, Gellert-Kristensen GRS and Dongiovanni GRS. The models were judged based on discrimination and calibration, patient values for 4 competing non-genetic HCC prediction models were calculated as well as 2 genetic models. The aMAP calculator performed the best, both in terms of discrimination and calibration, the Concordance Index in the Scottish cohort was 0.77, however it differed with each model, cohort, and age. A strength of this research lies in the accountability of non-HCC mortality in Cirrhosis patients, as they are at a higher risk of passing away from liver failure, so considering the deaths caused by factors other than HCC helps develop a better understanding of the biases in models' performances. While this study underscores the importance of predictive models in HCC stratification, at the same time it encourages further guidelines and thresholds to be put in place by healthcare providers to determine when HCC screening should be initiated based on individual risk assessments. As stated above, HCC contributes to further complications in the human body which become additional factors in the cause of death. Reference [17] developed and validated ensemble machine learning models to predict mortality in patients who developed bone metastases due to HCC. Artificial Neural Network, eXGBoosting Machine, Gradient Boosting Decision Tree, Support Vector Machine and Decision Tree are the models used to create the ensemble models. All 3 models attained a 70-75% accuracy in their predictions and AUROC rate of 0.779, 0.764, 0.778. A great strength and a gap in the previous literature this study fills are that it further classifies the patients into low and high mortality risk groups. The model performed exceptionally close to the observed risk of mortality in both groups by healthcare providers, which could prove extremely helpful in decision making process for both groups, such as whether a patient should be advised surgery if they have a high risk of passing away, within 3 months. There is a vast amount of literature supporting the hypothesis that HCV is a leading cause of HCC and mortality rates. Machine Learning has been proven immensely useful in making these deductions thus far, all the technologies used previously, and their processes of carrying out research have been pivotal to the present study.

To understand the implications of present study it is vital to first understand the core components and how they indicate or distinguish a healthy individual, to one suffering from HCV or HCC. These clinical indicators are proteins or enzymes and are found in either blood or the liver, and varying levels indicate positive HCC or HCV results. For example, higher levels of Bilirubin (BIL), Alkaline Phosphate (ALP), and Alanine Transferase (ALT) in liver indicate liver damage or a blockage i.e. tumor. However, a high level of Aspartate Aminotransferase (AST), in blood typically means an injury in the liver i.e. Cirrhosis. Lower levels of Albumin (ALB), in liver indicate severe liver damage i.e. inflammation or Cirrhosis [18].

III. METHODOLOGY

A. Dataset

The dataset needed for this research was acquired through the University of California Irvine Machine Learning Repository. It contained 13 different features, some recorded categorically the rest numerically, and contains missing values for some features, in total it holds 615 patients' records. The target variable which is the 'Category' column, contains 5 categories, which the patients were divided into, based on their diagnosis. Age and Sex were represented by their own columns, and the rest of 10 columns contain values of clinical indicators for each patient. As the original dataset contained only 615 rows of records, it could become a limitation when training models and testing them so to overcome that, a synthetic dataset based on the original dataset was generated using Conditional Tabular Generative Adversarial Network (CTGAN), which generates synthetic rows of tabular data, by learning the underlying distribution of the original dataset [19]. For efficiency and ease, the original dataset was pre-processed before being passed to CTGAN. As the categories and Sex were recorded as strings in the dataset, they were encoded before being passed to the model. The synthetic data generated contains 5000 new records and preserves all the columns from the original dataset. All the missing values in the original dataset were replaced with the mean value of that

column and it was specified to the model that columns ‘Category’ and ‘Sex’ were categorical features of the dataset to aid its understanding of the data.

B. Exploratory Data Analysis

Both datasets exhibited significant class imbalance, with only 1.1% of records labelled as 'suspected HIV' in the original dataset and 5.9% in the synthetic dataset, while a substantial 86.7% and 63.7% of records were labelled as healthy, respectively. Upon comparing each feature of both datasets, notable differences in patterns emerged, revealing that the synthetic dataset failed to accurately capture the relationships between variables and showed discrepancies in the generated values. The Correlation coefficient heatmap visualized the relationships between all variables, in the synthetic dataset it could be clearly seen that no features show a correlation, while Fig. 1 shows the real dataset has an array of features that strongly correlate to each other. For example, the “category” feature and “AST” (Aspartate Aminotransferase) showed a correlation of 0.65 (strong positive), meaning certain categories were strongly associated to higher values of “AST”. A moderate negative correlation could be seen between the variable “CHE” (Cholinesterase) and “Category”, meaning a decline in the values of “CHE” was seen as the categories progress to higher stages. While it could be seen in Fig. 1 that the strongest correlation, in the synthetic dataset was a weak positive correlation (0.15) between features “PROT” (total protein) and “CHOL” (lipid profile). One of the major limitations in the synthetic dataset was its outliers; as seen in Fig. 2, the values generated for “BIL” (Bilirubin) contained values in negative numbers, or excessively high positive values, which proposed a restriction in their usage for a healthcare research problem. SQL was used to extract the minimum and maximum values of each feature in the original dataset. Those values were utilized to set minimum and maximum thresholds for the synthetic dataset, all the data outside of those thresholds was dropped. Once the outliers had been removed from the Synthetic Dataset, another EDA was carried out to analyse the quality of the dataset post processing. The original dataset, as the literature suggested, contained data in varying values depending on categories or diagnosis, for example as Fig. 3 shows, the levels of “ALP” (alkaline phosphate) were the highest in patients diagnosed with Cirrhosis (Category 4). However, the synthetic dataset generated higher values for categories such as “Healthy” (Category 0) and “Suspected Blood Donor” (Category 1) and lower values for “Cirrhosis” (Category 4). This pattern could be seen in all of the features. As the literature suggests, HCV and HCC was diagnosed, based on the results of a combination of tests, therefore a correlation between those tests and stages of HCC and HCV were of utmost importance, whereas, the values generated by the synthetic dataset did not correlate to the categories of the dataset. Therefore, it was promptly identified and decided that the dataset failed to preserve the underlying patterns of the variables, such as an incline in one, and a decline in the other feature based on the category. Therefore the dataset had to be discarded, as the whole idea of a Synthetic dataset was larger quantities of similar values.

C. Pre-Processing

Several pre-processing techniques were implemented on the original dataset, and each yielded different results, the best results were achieved by the following techniques:

- Each unique value was mapped in both categorical columns i.e. diagnosis and sex and was replaced with a number starting from 0. This form of encoding was very similar to the concepts of Ordinal Encoding, it preserves the hierarchy in the categories and helps the models learn the patterns better.
- Reference [20] found that in comparison to more complex imputation methods such as conditional imputation, median imputation was not inferior. These findings led to the imputation method to be the median of the variables.
- As shown in the EDA, the dataset was highly imbalanced, meaning the difference in minority and majority class was quite significant, so to balance the dataset for this approach, current study used **SMOTE** (Synthetic Minority Over-Sampling Technique). SMOTE is an oversampling approach, where the dataset is balanced by creating new datapoints for the minority class synthetically. It uses KNN algorithm, each minority class sample is taken, and new synthetic examples are generated, joining all the k minority class nearest neighbour along the line segment [21].
- Scaling was omitted for this combination of pre-processing techniques, as experimentation with it in different settings showed no impact on results.

D. Evaluation Metrics

This research adopted the most common evaluation approaches from the current domain. Model predictions are categorized into four classes: True Positive (TP), where the model correctly predicts a positive instance; True Negative (TN), where the model correctly predicts a negative instance; False Positive (FP), where the model incorrectly predicts a positive instance; and False Negative (FN), where the model incorrectly predicts a negative instance. Accuracy, defined as the proportion of true predictions out of all attempts, includes both True Positive and True Negative predictions. Precision measures the proportion of correctly predicted positive values, while recall, also known as Sensitivity or True Positive Rate, represents the proportion of actual positives correctly identified by the model. The F1-Score combines Precision and Recall. Additionally, the Area Under the ROC Curve (AUROC) quantifies the overall performance across all classification thresholds, with values ranging from 0 (all predictions incorrect) to 1 (all predictions correct), as described by [22].

E. Model Selection

This research evaluated four models to identify the best performing model for using as the main prediction model of the research. Random Forest mitigates overfitting by averaging predictions from decision trees built on random training samples [23]. Logistic Regression organizes data into categories by applying a regression formula to establish the classification boundary [24]. The Extra Trees algorithm constructs unpruned decision trees with randomly split nodes using the entire training dataset, offering more randomness and reducing variance [25]. AdaBoost, an ensemble model, boosts weak classifiers by adjusting weights on training samples, known for its efficiency, feature selection, and minimal training time [26]. Each model's hyperparameters were manually fine-tuned to achieve optimal performance.

Fig. 1 Correlation Heatmaps of both datasets, showing a lack of correlation between features in the Synthetic Dataset (left) and Original Data (right)

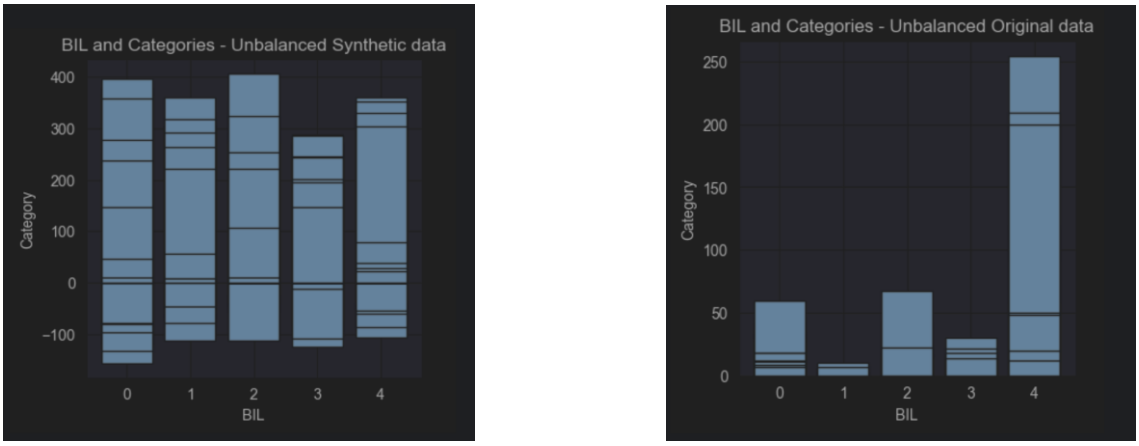

Fig. 2 Bilirubin and Categories compared in both datasets prior to outliers removal in Synthetic Dataset (left) and Original Data (right)

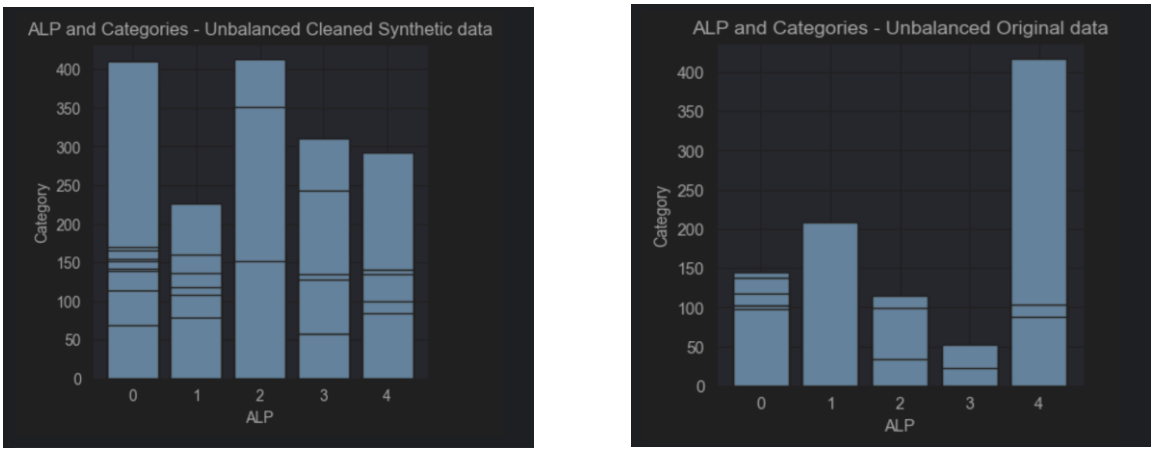

Fig. 3 Alkaline Phosphate and Categories compared in both datasets post outliers removal in Synthetic Dataset (left) and Original Data (right)

TABLE I. TABLE I. ACCURACY AND AUROC RESULTS

	Model			
	Random Forest	***Logistic Regression***	***Extra Trees model***	***AdaBoost Classifier***
Accuracy	0.9187	0.8211	0.8862	**0.9298**
AUROC	0.9092	0.9169	0.9263	**0.9749**

TABLE II. TABLE II PREDICTION RESULTS OF EACH CLASS

Class	**Model**											
	Random Forest			***Logistic Regression***			***Extra Trees model***			***AdaBoost Classifier***		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Healthy	0.94	1.00	0.97	0.98	0.89	0.93	0.90	1.00	0.95	**0.98**	0.99	0.99
Suspected HIV	1.00	0.33	0.50	1.00	0.67	0.80	1.00	0.67	0.80	**1.00**	0.33	0.50
HCV	1.00	0.56	0.71	0.29	0.44	0.35	1.00	0.44	0.62	**0.78**	0.78	0.78
Fibrosis	0.57	0.67	0.62	0.29	0.67	0.40	0.40	0.33	0.36	**0.56**	0.83	0.67
Cirrhosis	0.88	0.78	0.82	1.00	0.67	0.80	1.00	0.56	0.71	**0.86**	0.67	0.75

IV. RESULTS AND DISCUSSION

The research evaluated four prediction models, with the following results: Random Forest achieved an accuracy of 0.9187 and an AUROC of 0.9092. Logistic Regression recorded an accuracy of 0.8211 and an AUROC of 0.9169. The Extra Trees model attained an accuracy of 0.8862 and an AUROC of 0.9263. The AdaBoost Classifier outperformed all other models, with the highest accuracy of 0.9298 and an AUROC of 0.9749, as shown in Table I. The AdaBoost also excelled particularly in predicting and classifying categories such as “Healthy,” “HCV,” and “Cirrhosis.” Table II demonstrates the AdaBoost Classifier's effectiveness in distinguishing between classes and making correct predictions for most instances.

A probability test was carried out, which entails providing a trained classification model with an array of data representing each class, but unseen to the model before. A probability of the provided data belonging to each class is generated as a result. AdaBoost predicted each set as representative of its class with a probability of higher than 95%, confirming that the AdaBoost model was not affected by overfitting or underfitting, as it successfully identified the correct category for each category with significant margin. Feature importance analysis revealed that “BIL”(bilirubin) had the highest impact on the AdaBoost Classifier’s performance, while “ALB” was most influential for the Random Forest. “Sex” was the least contributing factor across all models.

All models performed the weakest for minority classes such as “Suspected HIV” and “Fibrosis,” highlighting the need for a larger and a balanced dataset to ensure accurate and early diagnoses for HCV and HCC. Consequently, the AdaBoost Classifier was identified as the best-performing model for the research project.

V. CONCLUSION

In conclusion, this study found that machine learning can play a crucial role in diagnosis of diseases and the implications of it in health sector for different demographics. A prediction system, or aid with initial diagnosis can lower the workload for health care officials as well as be advantageous for patients. An early diagnosis using this prediction system can be a cost effective and convenient first step for patients. A simple to use, time efficient and most importantly accurate system can help limit the spread of the disease as well as give patients a better at treatment. This study also concluded that a dataset acquired from a reliable source, however small in size was more serviceable than a synthetically generated dataset. Suitable pre-processing techniques play a part in determining the performance of classification models. Optimal choice and tuning of hyper-parameters enhanced the efficiency of AdaBoost classifier and positively impacted the accuracy rate of predictions.

In terms of further recommendations, it is vital to start with a larger dataset to ensure that this machine learning approach has the same implications for different demographics. Gathering the dataset of such scale can be a monumental task, but a significant one now, as the research in HCC diagnosis using machine learning cannot be furthered without it. Furthermore, this model can be used to create a Graphical User Interface (GUI) and supplied to marginally poor populations around the world, it can distinguish between categories that an HCV patient can be classified into. Overall, this research can be foundation to a greater system implied on a larger scale to aid healthcare operations as well as patients around the world.

REFERENCES

[1] World Health Organization, "Hepatitis C," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c#:~:text=Hepatitis%20C%20is%20an%20inflammation,including%20liver%20cirrhosis%20and%20cancer [Accessed: Aug. 4, 2024].

[2] J. Ng and J. Wu, “Hepatitis B- and Hepatitis C-Related Hepatocellular Carcinomas in the United States: Similarities and Differences,” *Hepatitis Monthly*, vol. 12, no. 10 HCC, Oct. 2012, doi: https://doi.org/10.5812/hepatmon.7635.

[3] A. K. Mitra, “Hepatitis C-related Hepatocellular Carcinoma: Prevalence Around the World, Factors Interacting, and Role of Genotypes,” *Epidemiologic reviews*, vol. 21, no. 2, pp. 180–187, Jan. 1999, doi: https://doi.org/10.1093/oxfordjournals.epirev.a017995.

[4] World Health Organization, "Guidelines for the screening, care and treatment of persons with chronic hepatitis C infection," [Online]. Available: https://iris.who.int/bitstream/handle/10665/246177/WHO-HIV-2016.06-eng.pdf?sequence=1. [Accessed: Aug. 4, 2024].

[5] M. C. Kew, “Hepatocellular carcinoma in developing countries: Prevention, diagnosis and treatment,” *World Journal of Hepatology*, vol. 4, no. 3, pp. 99–99, Jan. 2012, doi: https://doi.org/10.4254/wjh.v4.i3.99.

[6] N. Tsuchiya, Y. Sawada, I. Endo, K. Saito, Y. Uemura, and Tetsuya Nakatsura, "Biomarkers for the early diagnosis of hepatocellular carcinoma," *World Journal of Gastroenterology*, vol. 21, no. 37, pp. 10573–10573, Jan. 2015, doi: https://doi.org/10.3748/wjg.v21.i37.10573.

[7] C. Bruce-Lockhart and C. Campbell, "In charts: healthcare technology in low-income countries," *FinancialTimes*, May 17, 2020. https://www.ft.com/content/796a52e0-7334-11ea-ad98-044200cb277f (accessed Aug. 04, 2024).

[8] Cleveland Clinic, "Liver Disease: Signs & Symptoms, Causes, Stages, Treatment," [Online], Available: https://my.clevelandclinic.org/health/diseases/17179-liver-disease (accessed Aug. 04, 2024).

[9] NHS Choices, "Hepatitis B," 2024. https://www.nhs.uk/conditions/hepatitis-b/#:~:text=Vaccination%20is%20the%20best%20way,of%20them%20getting%20the%20infection. (accessed Sep. 27, 2024).

[10] M. Khatun and R. B. Ray, "Mechanisms Underlying Hepatitis C Virus-Associated Hepatic Fibrosis," *Cells*, vol. 8, no. 10, pp. 1249–1249, Oct. 2019, doi: https://doi.org/10.3390/cells8101249.

[11] National Cancer Institute, "Liver Cancer: Causes and Risk Factors," [Online]. Available: https://www.cancer.gov/types/liver/what-is-liver-cancer/causes-risk-factors#:~:text=Cirrhosis%3A%20The%20risk%20of%20developing,from%20working%20as%20it%20should. [Accessed: Aug. 4, 2024].

[12] NHS Choices, "Overview - Cirrhosis," 2024. https://www.nhs.uk/conditions/cirrhosis/ (accessed Aug. 04, 2024).

[13] U. K. Lilhore, et al. "Hybrid model for precise hepatitis-C classification using improved random forest and SVM method," *Scientific Reports*, vol. 13, no. 1, Aug. 2023, doi: https://doi.org/10.1038/s41598-023-36605-3.

[14] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intelligent Medicine*, vol. 2, no. 4, pp. 193–198, Nov. 2022, doi: https://doi.org/10.1016/j.imed.2021.12.003.

[15] S. Hashem *et al.*, "Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease," *Computer Methods and Programs in Biomedicine*, vol. 196, pp. 105551–105551, Nov. 2020, doi: https://doi.org/10.1016/j.cmpb.2020.105551.

[16] H. Innes *et al.*, "Performance of models to predict hepatocellular carcinoma risk among UK patients with cirrhosis and cured HCV infection," *JHEP Reports*, vol. 3, no. 6, pp. 100384–100384, Dec. 2021, doi: https://doi.org/10.1016/j.jhepr.2021.100384.

[17] Z. Long *et al.*, "Development and validation of an ensemble machine-learning model for predicting early mortality among patients with bone metastases of hepatocellular carcinoma," *Frontiers in Oncology*, vol. 13, Feb. 2023, doi: https://doi.org/10.3389/fonc.2023.1144039.

[18] "Monitoring your Hepatitis C," *Hep*, 2024. https://www.hepmag.com/basics/liver-health/hepatitis-c-lab-tests#:~:text=Alkaline%20phosphatase%20(ALP%20or%20Alk,and%20non%2Dliver%2Drelated (accessed Aug. 04, 2024).

[19] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veermachaneni, "Modeling Tabular Data using Conditional GAN." *Arxiv.org*, Oct, 2019, doi: https://doi.org/10.48550/arXiv.1907.00503 .

[20] Gijs F.N. Berkelmans *et al.*, "Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice," *Journal of Clinical Epidemiology*, vol. 145, pp. 70–80, May 2022, doi: https://doi.org/10.1016/j.jclinepi.2022.01.011 .

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research* vol. 16, pp. 321-357, June 2002, doi: https://doi.org/10.1613/jair.953

[22] J. Novakovic, A. Veljovic, S. Ilić, Ž. Papic, and Tomović Milica, "Evaluation of Classification Models in Machine Learning," *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, pp. 39-46, 2017.

[23] A. C. Kumar, J. A. John, M. Raja, and P. Vijaya, "Genetic factor analysis for an early diagnosis of autism through machine learning," *Elsevier eBooks*, pp. 69–84, Jan. 2023, doi: https://doi.org/10.1016/b978-0-323-98352-5.00001-x.

[24] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," Oct. 2019, doi: https://doi.org/10.1109/iccsnt47585.2019.8962457.

[25] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Machine Learning*, vol. 63, pp. 03-42, Mar. 2006, doi: https://doi.org/10.1007/s10994-006-6226-1

[26] S. S. Devi, V. K. Solanki, and R. H. Laskar, "Recent advances on big data analysis for malaria prediction and various diagnosis methodologies," *Elsevier eBooks*, pp. 153–184, Jan. 2020, doi: https://doi.org/10.1016/b978-0-12-818318-2.00006-4.