Hepatitis C Diagnosis Using Computational Intelligence Techniques

Zeynep Canbulat[1], Müjde Erol Genevois[1][0000-1111-5324-0612] and Michele Cedolin[2][0000-0003-2397-0010]

[1] Faculty of Engineering and Technology, Galatasaray University, Turkey
[2] College of Engineering and Technology, American University of the Middle East, Kuwait
merol@gsu.edu.tr

Abstract. The global impact of chronic Hepatitis C virus (HCV) infections, estimating around 58 million affected individuals worldwide with 1.5 million new cases annually. In 2019, approximately 290,000 deaths were attributed to Hepatitis C, as per the World Health Organization (WHO). Despite the availability of numerous laboratory tests for liver disease assessment, patients often show normal range laboratory values, and there are asymptomatic cases. The passage highlights the necessity for a sophisticated system to diagnose the disease accurately and efficiently, emphasizing the support of machine learning in this endeavor. It mentions the use of HCV data from the "UCI Machine Learning Repository" and the application of six computational intelligence algorithms, namely, logistic regression, random forest, k-NN, SVM, light-GBM, and XG-Boost. After data normalization, imputation, and oversampling processes light-GBM showed the best performance with 98,9% accuracy.

Keywords: Hepatitis C, Risk Analysis, Machine Learning, Health Data

1 Introduction

Machine learning methodologies hold significant promise for enhancing medical research and clinical care, particularly in the context of the growing utilization of electronic medical records by healthcare providers [1]. Healthcare facilities regularly record the health data of their patients in electronic health records (EHR) systems. These data encompass a heterogeneous array of elements, including but not limited to demographic information, prescription records, diagnostic records, vital signs, vaccination history, laboratory and radiology test results, medical concepts, clinical notes, procedural documentation, and treatment plans, all of which are systematically recorded during each patient encounter with a healthcare institution [2]. The escalating accumulation of medical data underscores the growing demand for proficient classifiers [3]. Classification algorithms initially undergo training using past patient data related to various diseases. Once trained, they make predictions regarding patients or potential illnesses which aid physicians in disease diagnosis. Numerous research endeavors have been conducted to date, focusing on biomedical datasets and the development of classification algorithms that form the foundation of decision support systems (DSS) [4]. The timely

identification of diseases significantly enhances the efficacy of treatment and ultimately preserves lives. Machine learning methodologies, a favored domain within the realm of information technology for early disease detection, yield valuable and productive outcomes [5], [6].

This research endeavor seeks to utilize machine learning algorithms, particularly in the context of hepatitis C, to predict and aid in the diagnosis of diseases. Globally, an estimated 58 million individuals are afflicted with chronic hepatitis C virus infection, with approximately 1.5 million new infections occurring annually. Among them, there are approximately 3.2 million adolescents and children enduring chronic hepatitis C virus infection. The World Health Organization (WHO) approximates that in 2019, roughly 290,000 fatalities resulted from hepatitis C, primarily attributed to cirrhosis and hepatocellular carcinoma (primary liver cancer). According to the web and editorial staff at Acıbadem Hospitals, it is estimated that approximately one out of every 100 individuals in Turkey is infected with the hepatitis C virus [7]. The remainder of this study is as follows. The second section introduces the relevant literature. The case study will be presented in the third section along with data and used performance measures. The final section shows the results and findings comparing the performance of the alternative models.

2 Literature Review

Computational intelligence (CI) algorithms are widely used techniques in medical diagnosis and risk prediction in health systems. Asri et al. [8] conducts a study to predict and diagnose the risk of breast cancer using k-NN, decision tree, Naive Bayes, and SVM. Akgül et al. [9] use logistic regression, k-NN, and SVM algorithms for the diagnosis of hypothyroidism. Coşar and Deniz [6], try to predict mortality risks of cardiac diseases using k-NN, the random forest and logistic regression. Oğuz et al. [10] predict stroke risk using thirteen machine learning models, where random forest achieved the highest accuracy. Dong et al. [11] predict the risk of diabetic kidney disease, firstly using Random Forest Recursive Feature Elimination (RFE), then applying seven machine learning techniques, where Light-GBM shows the highest performance.

Kavitha et al. [12] aim to predict Alzheimer's disease using random forest achieved with 5-fold cross-validation. Kaunang [13] focuses on diagnosing hepatitis C using random forest, SVM, naive Bayes, logistic regression, and neural networks. Delpino et al. [14] highlights that diabetes and chronic kidney diseases were the most frequently studied conditions, with support vector machines, random forests, and logistic regression being the most used machine learning models.

The most used algorithms for predicting health risks and diagnosing medical conditions include SVM, logistic regression, Naive Bayes, k-NN (k-Nearest Neighbors), and random forest. The algorithms that have demonstrated the highest performance include Light-GBM, random forest, logistic regression, k-NN, SVM, and MLP. To assess the

performance of these algorithms, various metrics are commonly employed, including precision, accuracy, AUC (Area Under the Curve), sensitivity, specificity, and F-score.

3 Case Study

3.1 Hepatitis C

Liver function tests typically include alanine transaminase (ALT) and aspartate transaminase (AST), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), bilirubin (BIL), prothrombin time (PT), international normalized ratio (INR), total protein and albumin (ALB). These tests can help determine an area of the liver that may be damaged and, depending on the pattern of elevations, establish a differential diagnosis [15]. In the diagnosis and risk assessment of hepatitis C, the most frequently utilized variables are ALT and AST, followed by ALB, ALP, GGT and BIL. White and red blood cells, hemoglobin, and platelet counts are also commonly used predictive features. However, Ahmed et al. [16] show that patients with advanced liver disease or cirrhosis often subtle changes in laboratory values, which are usually within normal ranges. In addition, laboratory tests can identify some, but not all, patients with significant liver damage. In addition, pregnancy, rapid bone growth or extrahepatic pathology may result in marked increases in ALP and may be mistakenly taken as evidence of liver disease and ALB is not specific for liver disease and can occur in any chronic inflammatory state or renal disease [17]. BLB can be high or low due to various diseases and pancreatitis may cause increases in GGT [18]. AST is present in a wide variety of other tissues such as heart, skeletal muscle, kidney, and brain [19].

In short, liver function tests are not disease specific, and many factors must be analyzed to diagnose a patient's hepatitis [20]. A sophisticated system is needed to help physicians diagnose diseases accurately and efficiently. Therefore, this study aims to diagnose hepatitis C and understand relationships between its variables using alternative CI techniques. In this study, hepatitis C will be diagnosed using "UCI Machine Learning Repository" data. First, pre-processing will be applied to the data. Next, normalization will be applied, and the oversampling method will be used. Variables for diagnosing hepatitis C will be used based on the CMP blood test so that hepatitis C can be determined during routine testing, but also based on liver function tests. Then, the XG-Boost, Light-GBM, MLP, SVM, logistic regression, random forest and KNN algorithms applied on the data with 5-fold cross-validation.

3.2 Dataset

The data are obtained from the "UCI Machine Learning Repository" website which contains various data that are publicly available[1]. Table 1. describes the data.

[1] https://archive.ics.uci.edu/dataset/571/hcv+data

Table 1. Data Characteristics and Description

Variable	Function	Type	Description	Unit	Missing Variable
Category	Objective	Categorical	"0=Blood donor", 0s= Suspicious blood donor", "1=Hepatitis C", "2=Fibrosis", "3=Cirrhosis"		No
ID	ID	Integer	Patient ID		No
Age	Feature	Integer	Age	Year	No
Sex	Feature	Binary	Sex		No
ALB	Feature	Continuous	Albumin		Yes
ALP	Feature	Continuous	Alkaline phosphatase		Yes
AST	Feature	Continuous	Aspartate aminotransferase		Yes
BIL	Feature	Continuous	Bilirubin		No
CHE	Feature	Continuous	Choline esterase		No
CHOL	Feature	Continuous	Cholesterol		Yes
CREA	Feature	Continuous	Creatinine		No
GGT	Feature	Continuous	Gamma - glutamyl transferase		No
PROT	Feature	Continuous	Protein		Yes
ALT	Feature	Continuous	Alanine aminotransferase		No

The distribution of blood test values among healthy, fibrosis, cirrhosis and hepatitis C patients are shown in Fig 1. where it might be observed that some indicators of hepatitis C are in the same interval which complicates the diagnosis of hepatitis C.

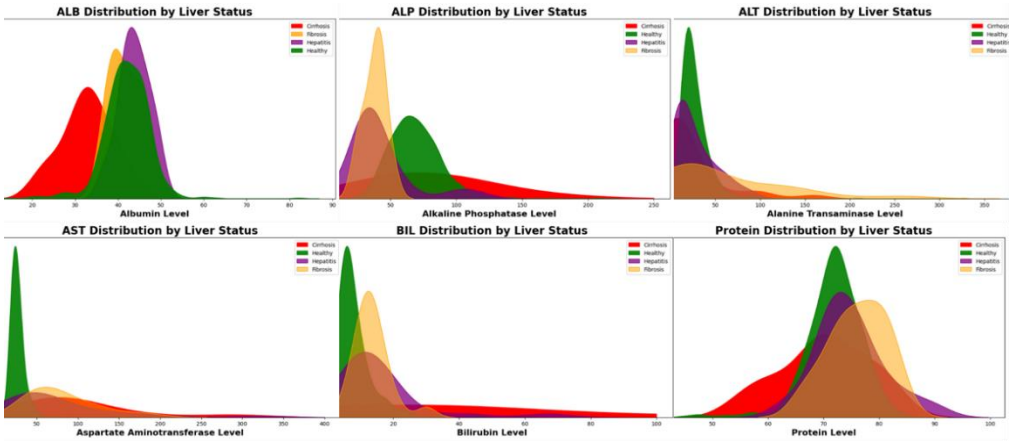


Fig. 1. Distribution of Blood Test Values Among Patients

3.3 Performance Measures

To evaluate the performance of the algorithms, sensitivity, accuracy, precision, and F-score will be measured. Accuracy measures the frequency of correct classifier predictions out of total predictions meanwhile recall explains how many real positive cases are predicted correctly out of all positive cases. Recall is a useful measure in cases where false negatives are of greater concern than false positives. Precision explains how many correctly predicted cases turned out to be positive. Precision is useful in cases where false positives are of greater concern than false negatives. In addition, F-score gives a combined idea of precision and sensitivity measures. It is highest when precision is equal to sensitivity.

4 Results and Findings

In the study, six different computational intelligence algorithms, namely logistic regression, SVM, kNN, LightGBM, XGBoost AND Random Forest are used with five-fold cross validation using python, Keras libraries. Table 2. Provides the outcomes of the study both for test data and oversampled test data, according to four alternative metrics.

Table 2. Model performances on both Data Sets

		Accuracy	Precision	Recall	F Score
Test Data	Logistic Regression	94,40%	87,00%	64,30%	74,00%
	SVM	96,50%	94,60%	76,50%	84,60%
	kNN	92,60%	91,10%	45,20%	60,30%
	LightGBM	97,20%	86,20%	92,20%	89,00%
	XGBoost	97,30%	88,20%	90,40%	89,20%
	Random Forest	96,90%	88,60%	86,10%	87,20%
Oversampled Test Data	Logistic Regression	92,60%	66,60%	82,60%	73,70%
	SVM	96,00%	80,30%	90,40%	84,90%
	kNN	95,80%	83,60%	82,60%	82,90%
	LightGBM	96,30%	79,30%	95,70%	86,60%
	XGBoost	95,60%	76,10%	93,90%	84,00%
	Random Forest	96,20%	80,10%	93,00%	86,00%

Fig.2. compares Accuracy and Precision for each model. The solid bars represent the metrics using the original dataset, while the semi-transparent bars represent the metrics using the oversampled dataset. The labels have been simplified to just the metric names and their oversampled counterparts (e.g., "Accuracy" and "Oversampled Accuracy"). Fig. 3. follows the same format for Recall and F Score, allowing for a direct comparison of how oversampling impacts these specific metrics across the various models.

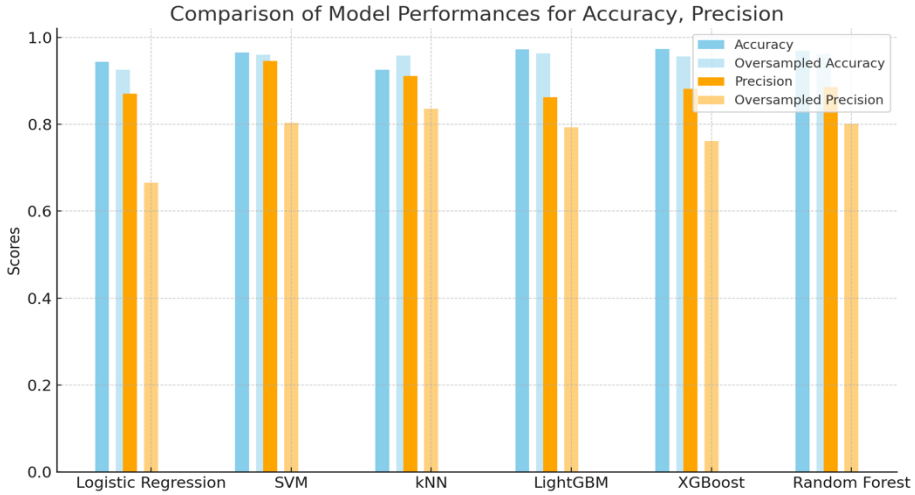


Fig. 2. Comparison of Model Performances for Accuracy and Precision

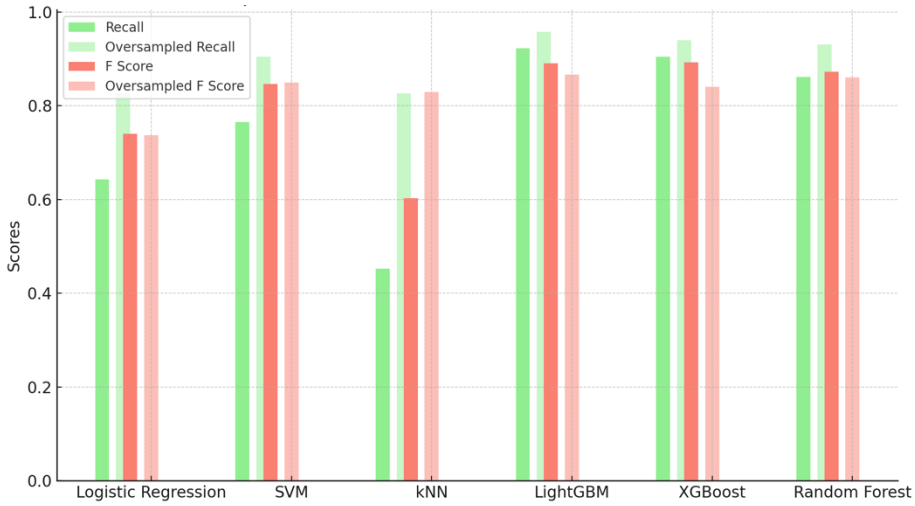


Fig. 3. Comparison of Model Performances for Accuracy and Precision

For some models, accuracy slightly decreases after oversampling (e.g., Logistic Regression, XGBoost), while for others, it remains relatively stable or slightly increases. This indicates that oversampling may not uniformly benefit all models in terms of overall accuracy. There's a noticeable decrease in precision for all models in the oversampled dataset. This suggests that while the models may be identifying more true positives, they're also increasing the number of false positives. All models show a significant improvement in recall with the oversampled dataset. This is expected, as oversampling aims to increase the representation of minority classes, making it easier for models to detect them. The F Scores are relatively stable or slightly improve for most models, indicating that the trade-off between precision and recall after oversampling

tends to balance out in terms of overall model performance. In summary, oversampling appears to significantly improve recall at the expense of precision. This trade-off results in a relatively stable or slightly improved F Score, suggesting that while models may become less precise, their ability to identify positive instances (especially from minority classes) improves, leading to potentially more balanced and fair predictions across different classes.

To determine the "best" model, we need to consider the trade-offs between the different performance metrics (Accuracy, Precision, Recall, and F Score) in both the original and oversampled datasets. For the test data set, XGBoost and LightGBM show high performance across all metrics, with XGBoost slightly leading in Accuracy and F Score. LightGBM, however, has a higher Recall, which could be crucial if identifying as many positive cases as possible is a priority. For the oversampled data set, LightGBM stands out with the highest Recall and a strong F Score, indicating excellent performance in identifying positive cases without sacrificing too much precision. SVM and Random Forest also show balanced performances across all metrics.

To conclude XGBoost and LightGBM are strong contenders for the best model with the original dataset, with XGBoost slightly leading in general performance but LightGBM excelling in Recall. With the oversampled dataset, LightGBM presents a strong case as the best model due to its excellent Recall and high F Score, making it highly effective at identifying positive cases while maintaining a good balance in overall performance. In Hepatitis C diagnosis, where missing a positive instance has serious consequences, a model with a higher Recall (like LightGBM with the oversampled dataset) must be preferred. Further research will focus on fuzzification of the data set according to the medical range of the values, and observing the impact of fuzzification on the prediction performance change.

Acknowledgment: This work has been financially supported by Galatasaray University Research Fund FBA-2024-1218.

References

1. Sidey-Gibbons, J., Sidey-Gibbons: Machine Learning in Medicine: A Practical Introduction. BMC Medical Research Methodology 19, 64 (2019).
2. Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, C., Cavedon, L.: The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. ACM Computing Surveys, 55(2), 1–36 (2022).
3. Badem, H.: Parkinson Hastaliğinin Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri İle Tanimlanmasi . *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 8 (2), 630-637 (2019).
4. Karakoyun, M. & Hacıbeyoğlu, M.: Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Siniflandirma Algoritmalarinin İstatistiksel Olarak Karşilaştirilmasi . Dokuz

Eylül Üniversitesi Mühendislik Fakültesi Fen Ve Mühendislik Dergisi, 16 (48), 30-42 (2014).

5. Bratić, B., Kurbalija, V., Ivanović, M. : Machine Learning for Predicting Cognitive Diseases: Methods, Data Sources and Risk Factors. J Med Syst 42, 243 (2018).
6. Coşar, M., Deniz, E.: Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. Avrupa Bilim Ve Teknoloji Dergisi, (ICAENS), 1112-1116 (2021).
7. Hepatit C Belirtileri, Tanı Ve Tedavi Yöntemleri, https://www.acibadem.com.tr/Ilgi-Alani/Hepatit-C, last accessed 2023/12/4.
8. Asri H., Mousannif H., Al Moatassime H., Noel T.: Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, Procedia Computer Science, 83, 1064-1069, (2016).
9. Akgül, G., Çelik, A. A., Ergül Aydın, Z., Kamışlı Öztürk, Z.: Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı. *Bilişim Teknolojileri Dergisi*, 13 (3), 255-268 (2020).
10. Oğuz, Ö., Bayir, S., & Badem, H.: Makine Öğrenmesi Yöntemlerinin Felç Riskinin Belirlenmesinde Performansı: Karşılaştırmalı Bir Çalışma. Computer Science, IDAP-2021: 5th International Artificial Intelligence and Data Processing Symposium. 274-287 (2021).
11. Dong, Z., Wang, Q., Ke, Y. et al. : Prediction Of 3-Year Risk of Diabetic Kidney Disease Using Machine Learning Based on Electronic Medical Records. J Transl Med 20, 143 (2022).
12. Kavitha C., Mani V., Srividhya S.R., Khalaf O.I. and Romero C.A.: Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Front. Public Health 10:853294 (2022).
13. Kaunang, F.: A Comparative Study on Hepatitis C Predictions Using Machine Learning Algorithms. 8ISC Proceedings: Technology, 33-42 (2022).
14. Delpino F. M., Costa, K., Farias, S.R., Chiavegatto Filho A.D.P., Arcêncio R.A., Nunes B.P.: Machine Learning for Predicting Chronic Diseases: A Systematic Review, Public Health, 205, 14-25, (2022).
15. Lala V., Zubair M., Minter D.A. : Liver Function Tests. Statpearls. Treasure Island (FL): Statpearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK482489/ last accessed 2023/12/27.
16. Ahmed, Z., Ahmed, U., Walayat, S., Ren, J., Martin, D. K., Moole, H., Koppe, S., Yong, S., Dhillon, S.: Liver Function Tests in Identifying Patients with Liver Disease. Clinical And Experimental Gastroenterology, 11, 301–307 (2018).
17. Johnson, P. J.: Role of the Standard 'Liver Function Tests' in Current Clinical Practice. Annals Of Clinical Biochemistry: International Journal of Laboratory Medicine, 26(6), 463–471 (1989).
18. Wolf P. L.: Biochemical Diagnosis of Liver Disease. Indian Journal of Clinical Biochemistry, 14(1), 59–90 (1999).
19. Thapa, B. R., Walia, A.: Liver Function Tests and Their Interpretation. Indian Journal of Pediatrics, 74(7), 663–671 (2007).
20. Polat K., Güneş S.: Hepatitis Disease Diagnosis Using a New Hybrid System Based on Feature Selection (FS) And Artificial Immune Recognition System with Fuzzy Resource Allocation, Digital Signal Processing, 16(6), 889-901 (2006).