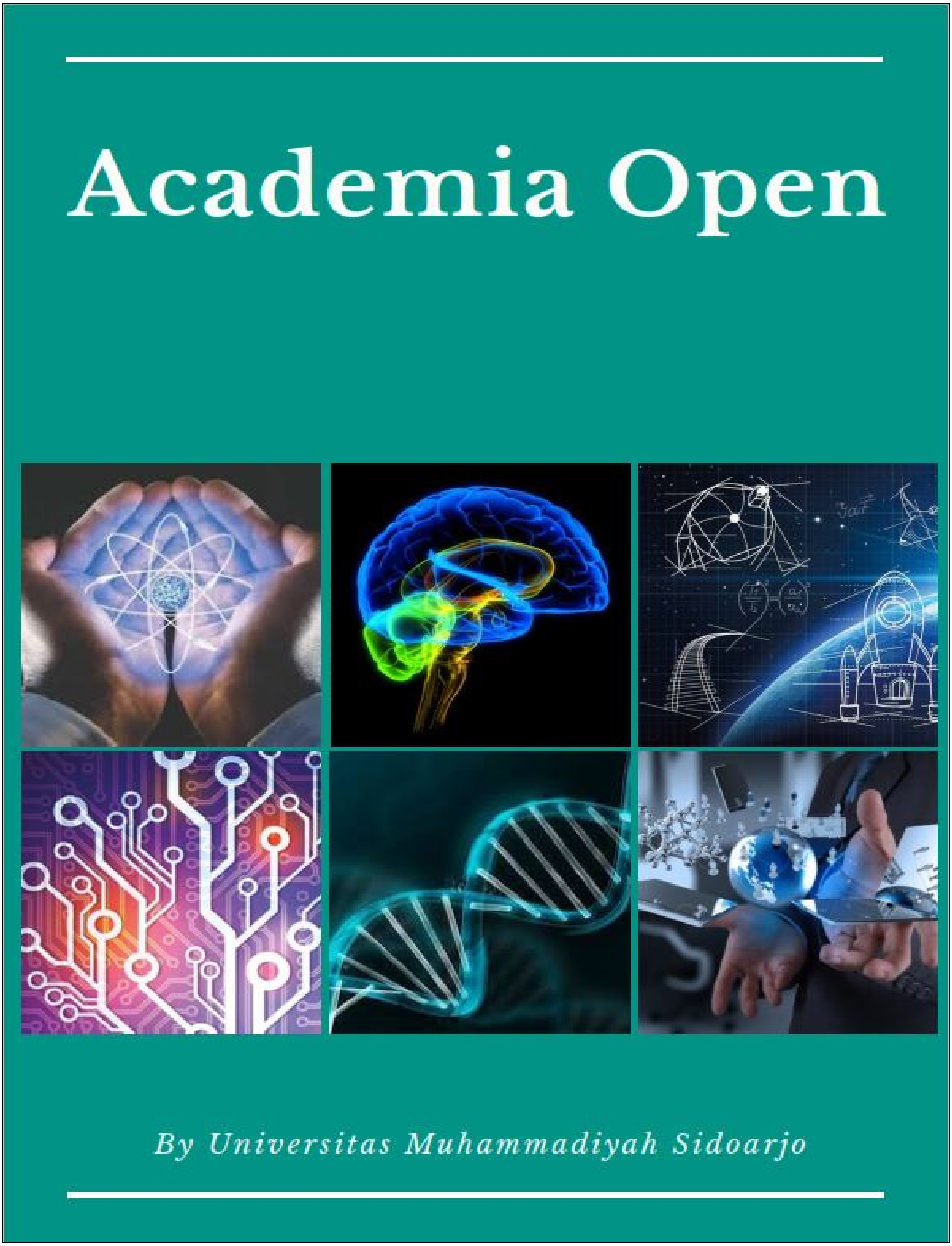
Academia Open
Vol 10 No 1 (2025): June (In Progress)
DOI: 10.21070/acopen.10.2025.10740 . Article type: (Computer Science)

Table Of Content

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo


Academia Open
By Universitas Muhammadiyah Sidoarjo

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open
Vol 10 No 1 (2025): June (In Progress)
DOI: 10.21070/acopen.10.2025.10740 . Article type: (Computer Science)

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Article information

Check this article update (crossmark)

Check for updates

Check this article impact [*]

Dimensions

LENS.ORG
Solving The Problem Of Problem Solving™

Semantic Scholar

Scilit

scite_

Google Scholar

wizdom.ai

Save this article to Mendeley

MENDELEY

[*] Time for indexing process is various, depends on indexing database platform

Academia Open
Vol 10 No 1 (2025): June (In Progress)
DOI: 10.21070/acopen.10.2025.10740 . Article type: (Computer Science)

Developing a Prediction Model to Identify Blood Types Most Susceptible to Viral Hepatitis Based on the CRISP-DM Methodology

Mengembangkan Model Prediksi untuk Mengidentifikasi Golongan Darah yang Paling Rentan terhadap Virus Hepatitis Berdasarkan Metodologi CRISP-DM

Esraa Hameed Kamel, ikamel@uowasit.edu.iq, (1)
Waist University, college of finearts ,waist , Iraq

[(1)] Corresponding author

Abstract

General Background: Viral hepatitis is a prevalent disease worldwide, with hepatitis B and C posing significant public health challenges. While most cases resolve naturally, chronic infections contribute to severe complications. **Specific Background:** Genetic predisposition, including blood type, has been hypothesized as a risk factor for viral hepatitis; however, its role remains unclear. **Knowledge Gap:** Limited studies have analyzed the association between ABO blood groups and susceptibility to hepatitis B and C using machine learning techniques. **Aims:** This study aims to determine the blood groups most susceptible to hepatitis B and C by applying advanced machine learning models. **Results:** Using a dataset of 500 patients and CRISP-DM methodology, the findings indicate that blood type B has the highest susceptibility (38% infection rate), while type O shows the lowest risk (15%). Statistical analysis (Chi-square, $p < 0.01$) confirms a significant correlation between blood group B and hepatitis infection. The XG-Boost model achieved the highest predictive accuracy (91%), identifying blood type B as the second most influential risk factor after age. **Novelty:** This study provides empirical evidence linking genetic factors, particularly blood type B, with hepatitis susceptibility using data-driven approaches. **Implications:** The findings highlight the importance of blood type screening in high-risk populations and the necessity of targeted prevention strategies.

Highlights:

1. Blood type may influence susceptibility to hepatitis B and C.
2. Blood type B shows highest risk; XG-Boost model achieves 91% accuracy.
3. Blood type screening aids early detection and targeted prevention strategies.

Keyword: Random Forest algorithm, Hepatitis B and C, KNN algorithm, Blood Groups, Decision Tree algorithm, support vector machine algorithm ,XG-Boost algorithm, neural network algorithm.

Published date: 2025-03-06 00:00:00

Introduction

Despite advancements in prevention and treatment, viral hepatitis remains a formidable global health threat with some 325 million people enduring chronic infection according to the World Health Organization, and over 1.1 million succumbing yearly either to cirrhosis or cancer caused by the disease. Differences in infection rates amongst individuals suggest a role for genetic factors like blood type in predisposing susceptibility. While progress has lifted some from the grip of this ailment, many worldwide remain in its grasp.[1]

Recent studies indicate that the Glycoantigens responsible for determining blood type represent the main influence on the ability of viruses to bind to host cells. For example, it was found that individuals with blood type (A) are 25% more probable to be infected with hepatitis (C) compared to blood type (O), while blood type (B) has an increased risk of infection with hepatitis (B). On the other hand, blood type (O) shows relative resistance due to the absence of these antigens.[2]

In this study, data from 500 patients from the Kaggle Dataset were analyzed, which included features such as age, gender, blood group, and liver enzymes (ALT, AST) to achieve the study goal of identifying the blood groups most associated with infection using machine learning algorithms. To improve early screening policies, especially in high-endemic areas, such as Africa and Southeast Asia, these results provide a clear practical framework[3]

Related work

In “Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt "A machine learning-based prediction framework is presented for predicting HCV among healthcare workers in Egypt. A database from the National Liver Center in Menoufia University is used. It includes 859 records and 12 features. To test the reliability of the proposed framework, two different scenarios are conducted, the first without feature selection and the second after feature selection based on sequential forward selection (SFS). Machine learning algorithms including naive Bayes, Random Forest (RF), K-nearest neighbor (KNN), and Logistic Regression are used. Then, the effect of parameter tuning on the learning techniques is evaluated. Experimental results show that SFS selection under the proposed framework achieves higher accuracy than without feature selection. In addition, the minimum learning time taken by the classifier is only 0.54 seconds under the full feature condition, and the classification accuracy is given as 94.06%. Finally, after hyperparameter tuning, the relevant classification accuracy is 94.88%. Only four features[4]

In “Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context” predicted the occurrence of liver ailment in the Indian populace, mostly in northeastern Andhra Pradesh, and this is performed through logistic regression and a voting classifier. To assess the model performance, a 5-Fold Cross Validation strategy was performed on a dataset of 584 patient records. The performance measures evaluated were Accuracy, Precision, Recall, and F1-Score, and the Accuracy Rates lie between 69.23% and 74.14%. While the model is relatively interesting, there is potential for improvement, which was demonstrated in the results, where precision and recall were not consistent. This currently advances knowledge further by additionally demonstrating the use of logistic regression with particular reference to the diagnosis of liver diseases. This encourages the incorporation of more comprehensive data and shows the critical importance that machine learning models play in enhancing diagnostic methods [5]**.**

In “Liver Disease Prediction and Classification using Machine Learning Techniques” Liver disease has become a global issue that requires more consideration. As a normal Kaggle dataset upon which we worked. Was collected in the northeastern part of Andhra Pradesh. There are 441 patient records for men and 142 medical records for women in the collection. The data consists of 10 features and 1 target to classify Hepatitis (1) or Hepatitis free (0). The evaluation results for logistic regression is 76.07 %, support vector machine is 74.36% , Gradient Boosting is 71.79% , Decision Tree is 70.09%, and random forest 80.34%, based on the accuracy values. The results indicated that after applying the PCA Technique for Feature Selection, RF achieved an accuracy of 80.37% which was significantly better than other previously applied methodologies [6].

In “A framework for identification and classification of liver diseases based on machine learning algorithms” Hepatocellular cancer (HCC) is the most frequently diagnosed form of liver disease are also frequently diagnosed with cirrhosis caused by hepatitis B, especially in Asian countries. The advent of artificial intelligence (AI) and machine learning Technique (ML) has significantly enhanced the feasibility of disease detection and classification in conjunction with clinical data. Based on the clinical experience and through the use of machine learning algorithms, performed analytical steps, including decision tree (DT), random forest (RF), logistic regression (LR), regularized regression (RR), and extreme gradient boosting (XG-Boost), the study aimed to identify the most relevant clinical parameters or risk factors for liver diseases in those particular 525 patients. Among the five machine learning approaches used in this study, the Random Forest (RF) classifier exhibited the highest Accuracy 76.2 %, Recall 84.3% , F1-score 77.5% , and AUC 99.9%. [7]

In “An Interpretable Machine Learning Approach for Hepatitis B Diagnosis” “Hepatitis B is a major public health concern worldwide. To detect this deadly virus, there has been significant effort to use machine learning techniques. The interpretability of our models improves the ability of humans to understand and trust the machine learning model. In this study, we applied Shapley additive explanations (SHAP), a game theory-based method, to explain and visualize the predictions of a machine learning model used to diagnose hepatitis B. Several models were built and achieved accuracy rates, including decision tree (75%), logistic regression (82%), support vector machines (75%), random forests (86%), adaptive boosting (Ada-Boost) (92%), and extreme gradient boosting (XG-Boost) (90%). At the same time, SHAP values indicated that bilirubin contributes significantly to the high mortality rate. Thus, elderly patients are more likely to die. Interpreting the results of the machine learning models demonstrated in this study can help health practitioners and health policymakers develop solutions to existing problems.".[8]

Methods

The Cross-Industry, Standard Process for Data Mining (CRISP-DM) is a data, mining methodology implemented in this research, as shown in Fig.1.

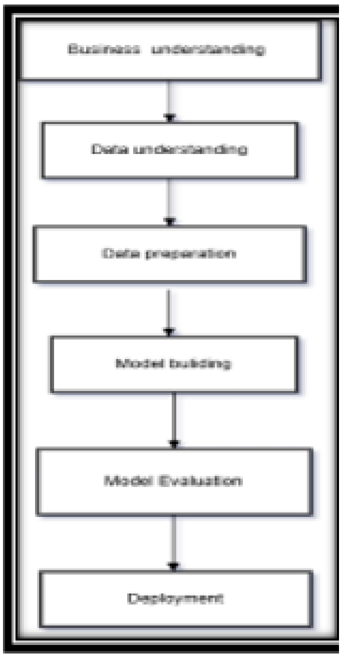


Figure 1. *The CRIPS-DM stage*

3.1 Business understanding: This is the first stage to transform real problems into a practical, executable technical framework. Objectives, knowledge gaps, success criteria, and requirements are identified and linked to data. At this stage[6], the high incidence of viral hepatitis with unexplained differences in infected individuals as well as the lack of studies that include analyzing the relationship between blood group (ABO) and viral hepatitis B and C using machine , learning techniques. The main question of our current study is whether blood group (ABO) is a risk factor for viral hepatitis. The performance, of learning algorithms in predicting this will be evaluated. The success criteria are the model's ability to analyze this relationship and its flexibility in handling big data. This research can be useful for hepatologists in making early diagnoses[9]. The sample size of 500 patients was determined using blood group, age, gender, diagnosis, and liver disorders (ALT and AST). Python language was used.

3.2 Data understanding: This is the second stage in the CRISP-DM framework[10], and it focuses on determining the nature of the available data. The Kaggle database (Hepatitis C and B dataset), in addition to the World Health, Organization (WHO) reports on the prevalence, of hepatitis, are the data sources for this research . The data were imported as CSV format, and then patterns and distributions were discovered. The aim of Table (1) was to understand the nature of the variables and thus determine the appropriate type of analysis for each variable, such as using the chi-square for the categorical variable.

Features	type	Range
Blood type	categorical	A, B, AB, O
Age	Numerical	15-85
ALT	Numerical	10–200 IU/L
AST	Numerical	8-180 IU/L
Gander	categorical	Male (1) Female

Diagnose	categorical	Not Infected (1) infected

Table 1. *Metadata Description*

Table (2) aimed to know the distribution of the sample among the different categories, and it was found that blood type (o) is the most common type. It also showed the imbalance in the data, as the number of infected people represented 30% of the total data. The table also showed that blood type (AB) is rare.

Features	Value	R atio
Blood type	A= 150 , B= 90 ,AB=60, O=200	30%,18%,12%,40%
Diagnose	Infected= 150Not infected=350	30% 70%
Gander	Male= 280 , female = 220	56% , 44%

Table 2. *Basic Distributions*

Table (3) aimed to determine the completeness of the data

Features	Ratio of missing value
ALT	3%
Blood type	1%

Table 3. *Missing value*

Table (4) showed a skewed distribution in the values due to the median not being equal to the mean, while the standard deviation value showed a dispersion in the data.

Features	Mean	Median	Standard Deviation
Age	45.2	47	12.3
ALT	48.3	45	22.1
AST	42.1	40	18.7

Table 4. *Basic statistic*

Table (5) showed a strong association between ALT and AST and that the elderly are more susceptible to the disease.

Feature 1	Feature 2	Pearson's Correlation Coefficient	Correlation type
ALT	AST	0.85+	Strong
Age	Diagnose	0.32+	Mid
Blood type	Diagnose	-	Categorical

Table 5. *Correlations*

3.3 Data preparation: This stage aims to transform raw data into model-ready data through a series of processes that address the issues identified in the data understanding stage[11]. Rows and columns with small missing data were removed <5% for the categorical feature (Blood Type) and replaced using the mean for the feature (ALT). There was a need to increase some sample's blood type (AB) due to their rarity. Encoding was implemented, and categorical variables such as blood type and gender were converted to numbers. Data were normalized, and numerical variables (such as Age, ALT) were made to a uniform scale. SMOTE (Synthetic Minority Oversampling Technique) was used to deal with unbalanced data[12]. Data was split (Train-Validation-Test Split) to ensure that the model was evaluated on data that it did not see during training.

3.4 Modeling: At this stage, predictive models are built using machine learning algorithms, and their performance is evaluated to reach the best model capable of predicting the risk of hepatitis based on blood types. [13] The data is divided into 70% training to build the model, 15% validation to adjust the parameters, and 15% testing for final evaluation. The algorithms are selected based on the nature of the data, its size, and the goal of the model. Random Forest algorithm was chosen for its ability to deal with non-linear interactions. The XG-Boost algorithm was,chosen for its effectiveness in unbalanced data. The SVM algorithm was chosen for its role in discovering complex patterns. A neural network algorithm was chosen for its effectiveness in dealing with complex data and non-linear interactions. It can automatically extract features. The decision tree algorithm is simple, and very easy to interpret. It does not require data calibration. Finally, K-Nearest Neighbors (KNN) algorithm is chosen for its simplicity and quick implementation. It is suitable for small or medium-sized data

3.5 Evaluation: This is the stage in which the model's performance is measured to ensure its ability to predict accurately and effectively.[14] The model's performance is evaluated based on several statistical measures, including accuracy, which represents the percentage of the correct predictions to the total predictions and is used as a general measure of performance. In the event of data imbalance, this measure is shaded. Precision, which represents the percentage of the correct positive predictions to the total positive predictions, reduces positive misdiagnosis. Recall, which represents the percentage of true positive cases that were correctly identified, is used to reduce negative misdiagnosis. AUC-ROC is used to analyses the model's capability to distinguish between the two classes.[15]

3.6 Deployment: This is the final stage of the project, where the research model is transformed into a practical application that can be used by beneficiaries (such as doctors or patients) to make informed decisions. Deployment involves creating an interactive interface and integrating the model into real systems. , thus making the model available for actual use in the early detection of high-risk groups.

Result and Discussion

Below is a detailed analysis of the execution of the various algorithms designed to predict the risk of hepatitis based on blood groups and demographic factors, as evidenced in table 6. The XG-Boost model stood out due to its effectiveness in handling imbalanced data. Its high accuracy stemmed from reducing errors across successive iterations. Its highest AUC-ROC score illustrates an excellent capability to differentiate between infected and uninfected individuals. Meanwhile, the Random Forest model attained an accuracy of 89% as it relies on clustering decision trees to decrease variance. This approach works well for nonlinear data but is less efficient than XG-Boost in managing complex interactions. The SVM model achieved a respectable accuracy rate of 88%. It relies on the RBF kernel to detect nonlinear patterns. However, careful tuning of the penalty factor (C) and the kernel is required. The neural network algorithm was 89% accurate because it can extract features automatically but needs more data to achieve better performance. Decision Tree accuracy was 84% a simple and easy-to-interpret model, but it is prone to overfitting KNN accuracy was 82% the lowest because it is sensitive to distances between points and is affected by poorly calibrated data

algorithm	Accuracy	precision	Recall	AUO-ROC
XG-BOOST	91%	87%	90%	0.93
Random forest	89%	85%	88%	0.91
Support vector machine	88%	84%	86%	0.89
Neural Network	89%	85%	87%	0.90
Decision Tree	84%	80%	82%	0.83
KNN	82%	78%	80%	0.81

Table 6. *Model evaluation*

Conclusion

Based on data analysis and evaluation of machine learning models, blood group B is the most susceptible group to viral hepatitis (B and C), as the results showed a relative increase in infection rates, as the infection rate among blood group B carriers reached about 38%, compared to group O, which recorded the lowest infection rate (15%). Statistical correlation tests (Chi-square) showed a significant relationship between blood group B and infection (p-value < 0.01).Genetic and biological reasons play a major role, as the presence of glycoproteins on the surface of red blood cells in group B facilitates the attachment of viruses to host cells, and the cellular immunity of this group is associated with a lower response of T-cells to hepatitis viruses compared to group (O) .The results also showed interaction with demographic factors, as it was found that patients with blood group B who were over 50 years old were 60% more susceptible to infection, compared to those under this age .The XG-Boost model achieved the highest predictive accuracy (91%), confirming that blood type B was the second most influential feature after age in determining risk (Feature Importance ≈ 0.18).

The study recommends including blood type testing in periodic screening programs for high-risk groups (such as the elderly in endemic areas). And the need to educate those with type B about the need to avoid risk factors such as sharing needles.

References

1. “Global progress report on HIV, viral hepatitis and sexually transmitted infections, 2021.” [Online]. Available: https://www.who.int/publications/i/item/9789240027077

2. L. Zhang, Y., & Fang, “ABO Blood Group and HCV Infection Risk: A Meta-Analysis of Asian Populations,” Wiley Online Libr., vol. 51, no. 9, pp. 789–800, 2021, [Online]. Available: https://doi.org/10.1111/hepr.13689
3. Kaggle Dataset, “Hepatitis C Patient Records.” [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset
4. E. Abad-Segura, M. D. González-Zamar, J. C. Infante-Moro, and G. R. García, “Sustainable management of digital transformation in higher education: Global research trends,” Sustain., vol. 12, no. 5, 2020, doi: 10.3390/su12052107.
5. I. Alwiah, U. Zaky, and A. W. Murdiyanto, “Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context,” Indones. J. Data Sci., vol. 5, no. 1, pp. 1–7, 2024, doi: 10.56705/ijodas.v5i1.121.
6. S. Tokala et al., “Liver Disease Prediction and Classification using Machine Learning Techniques,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 2, pp. 871–878, 2023, doi: 10.14569/IJACSA.2023.0140299.
7. H. Ding, M. Fawad, X. Xu, and B. Hu, “A framework for identification and classification of liver diseases based on machine learning algorithms,” Front. Oncol., vol. 12, no. October, pp. 1–7, 2022, doi: 10.3389/fonc.2022.1048348.
8. H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, “Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt,” Knowl. Inf. Syst., vol. 65, no. 6, pp. 2595–2617, 2023, doi: 10.1007/s10115-023-01851-4.
9. M. Badawy, N. Ramadan, and H. A. Hefny, “Healthcare predictive analytics using machine learning and deep learning techniques: a survey,” J. Electr. Syst. Inf. Technol., vol. 10, no. 1, 2023, doi: 10.1186/s43067-023-00108-y.
10. J. M. j. Herps, H. H. Van Mal, J. I. m. Halman, J. H. m. Martens, and R. H. m. Borsboom, “The process of selecting technology development projects: a practical framework,” Manag. Res. News, vol. 26, no. 8, pp. 1–15, 2003, doi: 10.1108/01409170310783619.
11. M. A. Jassim and S. N. Abdulwahid, “Data Mining preparation: Process, Techniques and Major Issues in Data Analysis,” IOP Conf. Ser. Mater. Sci. Eng., vol. 1090, no. 1, p. 012053, 2021, doi: 10.1088/1757-899x/1090/1/012053.
12. T. Wongvorachan and S. He, “A Comparison of Undersampling , Oversampling , and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining,” 2023.
13. M. Sumaiya, A. Mim, J. Nayeem, and S. Rana, “A Predictive Approach for Hepatitis Disease Diagnosis in Early Stage using Machine Learning Techniques,” no. January, 2024, doi: 10.2139/ssrn.4691067.
14. R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, “Application of machine learning algorithms for clinical predictive modeling: A data-mining approach in SCT,” Bone Marrow Transplant., vol. 49, no. 3, pp. 332–337, 2014, doi: 10.1038/bmt.2013.146.
15. D. Martin and W. Powers, “Evaluation : From precision , recall and F-measure to ROC , informedness , markedness & correlation EVALUATION : FROM PRECISION , RECALL AND F-MEASURE TO ROC , INFORMEDNESS , MARKEDNESS & CORRELATION,” no. January 2011, 2015, doi: 10.9735/2229-3981.