Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 611 – 621
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Early Detection of Hepatitis Disease Using Machine Learning Algorithms

[1*]Maya Gian Sister, [2]Yulia Nita, [3]Achmad Solichin
[1,2,3]Master of Computer Science, Faculty of Information Technology, Universitas Budi Luhur, Indonesia
Email: [1]2311601468@student.budiluhur.ac.id, [2]2311601435@student.budiluhur.ac.id,
[3]achmad.solichin@budiluhur.ac.id

Article Info
Article history:
Received Aug 03rd, 2025
Revised Sep 13th, 2025
Accepted Sep 25th, 2025

Keyword:
Diagnosis
Early Detection
Hepatitis
Machine Learning
Support Vector Machine

ABSTRACT
Hepatitis is an inflammation of the liver caused by viral infections, autoimmune disorders, or exposure to toxic substances. Hepatitis B and C are major public health concerns because they may progress to cirrhosis or liver cancer. In Indonesia, the transmission rate remains high, primarily through blood contact, unsterile needles, transfusions, and maternal delivery. Limited public awareness, coupled with the often asymptomatic nature of hepatitis, leads to delayed detection, which increases the risk of severe complications and mortality. Therefore, early detection is crucial to minimizing the disease burden. This study proposes a risk prediction model for hepatitis using non-laboratory clinical data and machine learning methods. Eight classification algorithms were compared Naïve Bayes, K-Nearest Neighbor (K-NN), Random Forest, Support Vector Machine (SVM), Decision Tree, AdaBoost, XGBoost, CatBoost, and LightGBM. Model performance was evaluated using K-fold cross-validation, with metrics including accuracy, precision, recall, F1-score, and AUC. The results show that the SVM with a linear kernel achieved the highest performance, with 87% accuracy and balanced F1-scores across all classes. The model successfully classified four categories: Acute Hepatitis, Chronic Hepatitis, Liver Abscess, and Parasitic/Viral Infections. These findings highlight the potential of machine learning to improve the early detection of hepatitis effectively and efficiently.
Copyright © 2025 Puzzle Research Data Technology

Corresponding Author:
Maya Gian Sister,
Faculty of Information Technology,
Universitas Budi Luhur,
Jl. Ciledug Raya, RT.10/RW.2, Petukangan Utara, Kec. Pesanggrahan, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12260, Indonesia
Email: 2311601468@student.budiluhur.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.38084

1. INTRODUCTION

Hepatitis is an inflammatory disease of the liver that may resolve spontaneously but carries the risk of progressing to fibrosis, cirrhosis, or hepatocellular carcinoma**.** The condition is primarily caused by hepatitis viruses A, B, C, D, and E, although non-infectious factors such as autoimmune disorders and exposure to toxic substances can also trigger it [1]. According to the World Health Organization (WHO) [2] (WHO, 2024), there are an estimated 296 million cases of hepatitis B and 58 million cases of hepatitis C worldwide, with more than three million new infections reported annually. Hepatitis A and E are transmitted via the fecal–oral route, whereas hepatitis B, C, and D spread through body fluids such as blood, sexual contact, contaminated needles, and perinatal transmission [3].

Most individuals with hepatitis remain asymptomatic, but some may experience symptoms such as fatigue, nausea, vomiting, fever, dark urine, or jaundice [4]. High-risk groups include infants born to mothers with active infection, hemodialysis patients, injection drug users, people living with Human

Immunodeficiency Virus (HIV), migrants from endemic regions, incarcerated individuals, and recipients of blood transfusions [5]. At the regional level, Indonesia bears the highest burden of acute hepatitis in the Association of Southeast Asian Nations (ASEAN), with a mortality rate of 2.14 per 100,000 population, higher than Cambodia (1.87) and substantially above eight other ASEAN countries, which report rates below 1 per 100,000 (Global Burden of Disease (GBD), 2019) [2]. Nationally, hepatitis is a major public health problem, causing approximately 60,000 deaths annually from hepatitis B and over 6,000 from hepatitis C [6].

The major challenge in hepatitis management lies in delayed diagnosis. More than 80% of new cases are detected only when severe liver damage has already occurred [7]. This delay increases the risk of clinical complications, restricts therapeutic options, and escalates the economic burden due to long-term treatment costs [8]. Furthermore, conventional diagnostic approaches remain dependent on laboratory tests that are costly, time-consuming, and often inaccessible in primary healthcare settings [9].

Machine Learning offers a promising alternative by enabling the analysis of large volumes of medical data, identifying hidden patterns, and providing faster and more accurate predictions [10]. Previous studies have shown that algorithms such as Support Vector Machine (SVM), Random Forest, and XGBoost can achieve accuracies exceeding 90% (Alizadehsani et al., 2019; Nasri et al., 2025). However, most existing research relies heavily on laboratory data, which limits applicability for early detection in primary healthcare services.

This study was conducted because hepatitis remains a global health threat with high mortality, particularly types B and C, which often progress asymptomatically until reaching cirrhosis or liver cancer [11]. Early detection is challenging, as most cases exhibit no symptoms, while predictive efforts are hindered by the complexity of medical data. Therefore, this research aims to develop a hepatitis risk prediction model based on non-laboratory clinical data by comparing various machine learning algorithms, including Naïve Bayes, K-Nearest Neighbor (K-NN), Random Forest, Decision Tree, SVM, AdaBoost, XGBoost, CatBoost, and LightGBM. Model evaluation is performed using 10-fold cross-validation with accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) as performance metrics. The contribution of this study is to provide an effective, efficient, and practical early prediction model to support primary healthcare services in accelerating detection, reducing complications, and facilitating preliminary community screening prior to medical consultation, while also offering theoretical contributions by enriching the literature on the application of machine learning in early disease detection.

2. RESEARCH METHOD

This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which comprises six stages, business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This framework was selected because it has been extensively applied in machine learning and data mining research, thereby ensuring a systematic workflow and producing reproducible results [12] [13].

2.1. Literature Review

Previous studies have examined various machine learning approaches for hepatitis prediction, applying different algorithms, attributes, and evaluation techniques (see Table 1).

In summary, prior research demonstrates that both traditional and advanced machine learning algorithms can achieve high accuracy in hepatitis prediction, particularly when supported by proper feature selection and data balancing techniques. These findings provide a strong foundation for the present study to further compare multiple algorithms using non-laboratory clinical data for early risk prediction.

2.2. Business Understanding

The business understanding phase was conducted through direct observation of the hospital information system, interviews with medical practitioners, and a comprehensive literature review. The primary objective at this stage was to compare the performance of eight classification algorithms for hepatitis diagnosis: Naïve Bayes, K-NN, Decision Tree, Random Forest, SVM, AdaBoost, Gradient Boosting, and CatBoost [13].

2.3. Data Understanding

In this phase, medical records of hepatitis patients from RSUD Haji Damanhuri Barabai were collected using purposive sampling. The dataset included demographic attributes, clinical notes, and diagnostic labels categorized into four classes Acute Hepatitis, Chronic Hepatitis, Liver Abscess, and Viral or Parasitic Infection. Descriptive analysis was carried out to examine class distribution, identify missing values, and assess potential data imbalance. The complete dataset before feature selection is presented in

Table 2. For brevity, only a portion of the records is shown in the table, while the full dataset was used in the analysis.

Tabel 1. Literature Review

No	Author(s) & Year	Method Used	Research Attributes	Research Findings
1	Ahmed et al. (2022) [14]	Random Forest, Decision Tree, and SVM algorithms with feature selection	19 attributes including age, gender, steroid use, antiviral use, fatigue, malaise, anorexia, hepatomegaly, liver firmness, palpable spleen, spider angioma, ascites, varices, bilirubin, alkaline phosphatase, SGOT, albumin, prothrombin time, and histology.	Random Forest achieved the highest accuracy of 96.1%, followed by Decision Tree at 94.3%, and SVM at 92.2%
2	Sharfina and Ramadhan (2023) [15]	Random Forest and Naïve Bayes with Synthetic Minority Oversampling Technique (SMOTE)	Age, gender, albumin, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, bilirubin, cholinesterase, cholesterol, creatinine, gamma-glutamyl transferase, and protein	Random Forest without SMOTE achieved 93% accuracy, which increased to 98% after SMOTE; Naïve Bayes achieved 88% without SMOTE and increased slightly to 89% after applying SMOTE
3	Damayanti and Testiana (2023) [16]	Naïve Bayes algorithm	Patient data categorized as blood donor, suspected donor, hepatitis, fibrosis, or cirrhosis, with attributes including age, gender, albumin, alkaline phosphatase, alanine transaminase, aspartate aminotransferase, bilirubin, cholinesterase, cholesterol, creatinine, gamma-glutamyl transferase, and protein	Naïve Bayes achieved an accuracy of 85.71% and was classified as "Good" based on the Area Under the Curve (AUC) standard
4	Putra et al. (2024) [17]	Naïve Bayes and K-NN algorithms	Age, gender, steroid use, antiviral use, fatigue, malaise, anorexia, hepatomegaly, liver firmness, spleen condition, ascites, varices, bilirubin, alkaline phosphatase, SGOT, albumin, prothrombin time, and histology	K-NN achieved the highest accuracy of 95.83% with 97% precision and 98% recall, whereas Naïve Bayes achieved 91.67% accuracy with 95% precision and recall
5	Diqi et al. (2024) [18]	Convolutional Neural Network (CNN) compared with traditional algorithms such as SVM, Decision Tree, K-NN, Gaussian Naïve Bayes, and Gradient Boosting	Liver enzymes (alanine transaminase and aspartate aminotransferase), bilirubin level, albumin, prothrombin time, and other physical conditions influencing patient prognosis	Convolutional Neural Network achieved perfect accuracy in classifying patients as alive or deceased, while SVM achieved 94% and Decision Tree only 75%

Tabel 2. Dataset Before Feature Selection

Service Date	Service Number	Medical Record No.	Gender	Age	Diagnosis Category	…	Weight Loss	Rash/Joint Pain	Chills
1/6/2023	2024/01/06/000887	134988	F	35	Acute Hepatitis	…	No	No	No
1/6/2023	2024/01/06/000785	186656	F	26	Acute Hepatitis	…	No	No	No
…	…	…	…	…	…	…	…	…	…
12/27/2024	2024/01/27/000708	185021	F	48	Abscess of liver	…	No	Yes	No
12/29/2024	2024/01/29/000766	186828	M	7	Abscess of liver	…	Yes	Yes	No

2.4. Data Preparation

The data preparation stage involved removing duplicate entries and handling missing values. The features were normalized using Min-Max Normalization:

$$x' = \frac{x - x_{max}}{x_{max} - x_{min}} \quad (1)$$

where x is the original value, x_{min} is the minimum, and x_{max} is the maximum of the feature. Categorical variables were transformed using one-hot encoding, feature selection was carried out with Recursive Feature Elimination (RFE), and class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [19].

2.5. Modeling

The modeling stage was conducted to evaluate and compare the performance of eight widely used classification algorithms in the context of hepatitis diagnosis. These algorithms were selected because they represent diverse learning paradigms, ranging from probabilistic approaches and distance-based classifiers to

decision tree ensembles and boosting techniques. By employing multiple algorithms, this study sought to provide a comprehensive perspective on how different machine learning methods perform when applied to the same medical dataset. The algorithms tested in this study included Naïve Bayes, K-NN, XGBoost, Random Forest, SVM, AdaBoost, Gradient Boosting, and CatBoost. Each algorithm was applied based on its fundamental mathematical formulation and learning principle, as described below [20].

Naïve Bayes applies Bayes' Theorem:

$$P(c|x) = \frac{\mathrm{P(X|C).P(C)}}{\mathrm{P(x)}} \tag{2}$$

Where P(C|X) is the posterior probability of class C given data X, P(X|C) is the likelihood, P(C) is the prior, and P(X) is the evidence. For continuous features, the Gaussian distribution is used:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma 2}} \tag{3}$$

with μ as the mean and σ as the standard deviation [21].

K-NN classifies a new instance based on the majority label of its k nearest neighbors using Euclidean distance:

$$\mathrm{d(x,y)} = \sum_{i=1}^{n} (\mathrm{xi} - \mathrm{yi})^2 \tag{4}$$

where n is the number of features [22].

Random Forest constructs an ensemble of decision trees, and the final prediction is determined by majority voting:

$$\mathrm{H(x)} = \mathrm{mode}\{\mathrm{h1(x)}, \mathrm{h2(x)}, \ldots, \mathrm{hk(x)}\} \tag{7}$$

where hi(x) is the prediction of the iii-th tree [23] [24].

SVM finds the optimal hyperplane with maximum margin:

$$\mathrm{w} \cdot \mathrm{x} + \mathrm{b} = 0 \tag{8}$$

where w is the weight vector, x the input vector, and b the bias. The margin is defined as:

$$\mathrm{Gain(S,A)} = \frac{2}{||\mathrm{w}||} \tag{9}$$

For non-linear data, kernel functions such as the Radial Basis Function (RBF) are employed:

$$\mathrm{K(xi, xj)} = \exp\left(-\gamma \,||\, \mathrm{xi} - \mathrm{xj}\, ||^2\right) \tag{10}$$

To provide probabilistic outputs, the decision function f(x) can be transformed using a logistic sigmoid function:

$$\mathrm{P(y = k \mid x)} = \frac{1}{1+e^{\mathrm{Af(x)+B}}} \tag{11}$$

where A and B are parameters estimated during calibration [25] [26].

AdaBoost adaptively combines weak learners as follows:

$$\mathrm{H(x)} = sign\left(\sum_{t=1}^{T} a_t\, h_t\,(\mathrm{x})\right) \tag{12}$$

where $h_t(x)$ is the weak classifier and α_t is its weight determined by accuracy [27] [28].

Gradient Boosting improves predictions iteratively by adding weak learners to correct residuals:

$$F_m(\mathrm{x}) = F_{m-1}(x) + \nu h_m\,(\mathrm{x}) \tag{13}$$

where ν is the learning rate [29].

CatBoost applies ordered boosting to handle categorical features effectively:

$$y^{\wedge}(t) = y^{\wedge}(t-1) + \eta f_t(x) \tag{14}$$

where η is the learning rate and $f_t(x)$ is the decision tree at iteration t. This method prevents target leakage and enhances generalization [30] [31].
The general form of the objective function in XGBoost is defined as:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{15}$$

where $l(yi,\hat{y}_i^{(t)})$ represents the loss function between the true value yi and the prediction $\hat{y}_i^{(t)}$, while Ω(fk) denotes the regularization term applied to each tree [32]. The complexity penalty function is given as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{16}$$

where T is the number of leaves, wj is the weight of leaf j, γ is the regularization parameter controlling the number of leaves, and λ is the penalty coefficient on leaf weights. This formulation ensures that the model balances fitting accuracy with model simplicity, thereby reducing overfitting and improving generalization [31].

2.6. Evaluation

The evaluation stage was conducted using stratified 5-fold cross-validation on 80% of the training data and independent testing on 20% of the test data [33]. The performance metrics employed included Accuracy, Precision, Recall, F1-Score, AUC, and the Confusion Matrix [13]. For multiclass classification, a one-vs-rest approach with macro averaging was applied. [34][35].

1. True Positive Rate (Recall) - Macro Averaging
   Recall measures the proportion of correctly identified positive instances from all actual positive cases.

$$TPR_{macro} = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{15}$$

2. Positive Predictive Value (Precision) – Macro Averaging
   Precision indicates the proportion of correctly predicted positive instances among all predicted positives.

$$PPV_{macro} = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} \tag{16}$$

3. F1-Score - Macro Averaging
   F1-Score represents the harmonic mean of Precision and Recall, providing a balanced measure that accounts for both false positives and false negatives.

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} \frac{2.\text{Precision}_i.Recall_i}{\text{Precision}_i.Recall_i} \tag{17}$$

2.7. Deployment

In the deployment stage, the optimized models were stored in reusable formats, such as Pickle or Joblib, so they could be applied to new data without retraining. The models produced classification outputs for Acute Hepatitis, Chronic Hepatitis, Liver Abscess, and Viral or Parasitic Infection, along with probability scores indicating the likelihood of each class. Although not intended to replace medical diagnosis, these predictions can serve as an early warning tool to support timely clinical examination.

3. RESULTS AND ANALYSIS

This section presents the research results, along with a comprehensive analysis and discussion. The findings are described using tables, figures, and graphs to facilitate a clear understanding of the outcomes. Furthermore, each result is critically interpreted and discussed to highlight its significance, relevance to previous studies, and contribution to the research objectives. Sub-sections are provided to ensure the analysis is structured, coherent, and systematically aligned with the research methodology.

3.1. Business Understanding

This research is motivated by the challenge of early hepatitis detection, which remains difficult due to the absence of clear initial symptoms, often leading to delayed clinical diagnoses. To understand the problem context, observations of the hospital information system were conducted, supported by interviews with medical practitioners and a literature review. The findings reveal that the complexity of medical data and the limitations of clinical examinations constitute the main obstacles. As a solution, this study proposes the application of machine learning to develop a predictive classification model. The model is expected to enhance the accuracy of early hepatitis diagnosis while also providing decision support for medical professionals and helping the community recognize potential risks at an earlier stage.

3.2. Data Understanding

The dataset consists of 561 patient records collected from RSUD Haji Damanhuri Barabai during the 2023–2024 period. It includes demographic information (age, gender), 17 clinical attributes, and diagnostic labels categorized into four classes: Acute Hepatitis, Chronic Hepatitis, Liver Abscess, and Viral/Parasitic Infection. Preliminary analysis indicates class imbalance, where Acute and Chronic Hepatitis cases are more dominant compared to the other two categories. This condition may reduce model performance, thus requiring data balancing strategies in subsequent stages.

3.3. Data Preparation

The data preparation stage was carried out through several essential steps. First, data cleaning was performed by removing duplicate and incomplete entries. Second, all attributes were transformed into numerical form to be processed by machine learning algorithms (see Table 3). Following this, all features were normalized to ensure consistent scales, and feature selection using RFE retained 17 relevant clinical attributes. Finally, to address class imbalance, the SMOTE was applied, resulting in a more proportional data distribution.

Table 3. Transformation of String Data into Numeric Form

Attribute	Data Type	Description
Gender	Numeric	1 = Male, 0 = Female
Age	Numeric	Patient's age in years
Diagnosis Category	Numeric	Acute Hepatitis = 1, Chronic Hepatitis = 2, Liver Abscess = 3, Viral/Parasitic Infection = 4
Fever	Numeric	1 = Yes, 0 = No
Fatigue	Numeric	1 = Yes, 0 = No
Loss of Appetite	Numeric	1 = Yes, 0 = No
Nausea and Vomiting	Numeric	1 = Yes, 0 = No
Upper Right Abdominal Pain	Numeric	1 = Yes, 0 = No
Dark Urine	Numeric	1 = Yes, 0 = No
Pale Stool	Numeric	1 = Yes, 0 = No
Jaundice	Numeric	1 = Yes, 0 = No
Itching	Numeric	1 = Yes, 0 = No
Edema/Ascites	Numeric	1 = Yes, 0 = No
Diarrhea/Digestive Disorder	Numeric	1 = Yes, 0 = No
Weight Loss	Numeric	1 = Yes, 0 = No
Rash/Joint Pain	Numeric	1 = Yes, 0 = No
Chills	Numeric	1 = Yes, 0 = No

3.4. Modeling

The modeling stage was conducted to compare the performance of eight classification algorithms, namely Naïve Bayes, K-NN, Decision Tree, Random Forest, SVM, AdaBoost, XGBoost, and LightGBM. The dataset was partitioned into 80% training data using stratified 5-fold cross-validation and 20% independent testing data. To address class imbalance, the SMOTE was applied.

As presented in Table 4, SVM achieved the highest performance, recording an accuracy of 86,61% and an AUC of 96,07%. Its capacity to construct an optimal hyperplane enabled robust classification within high-dimensional and unevenly distributed clinical datasets, thereby ensuring balanced predictive outcomes across diagnostic categories. In contrast, Decision Tree (81,25%) demonstrated susceptibility to overfitting the majority class, resulting in diminished recall for minority cases. Although Random Forest, XGBoost, and LightGBM attained relatively high AUC values (>94%) due to their ensemble mechanisms, their overall accuracies remained below that of SVM, indicating less effective classification of minority categories. Naïve Bayes (82,14%) showed moderate stability but was constrained by the independence assumption among features an unrealistic condition for clinical data where correlations are common (e.g., nausea, vomiting, and

fatigue). K-NN (83,04%) exhibited persistent sensitivity to class distribution, even after rebalancing with SMOTE.

Further improvements were observed after hyperparameter tuning, as shown in Table 5. Ensemble algorithms, particularly Random Forest, demonstrated enhanced performance with an accuracy of 91,51%. Nevertheless, despite surpassing SVM in accuracy, Random Forest and LightGBM displayed strong dependency on parameter configurations and the number of estimators, which reduces their robustness in varying settings. By contrast, SVM consistently achieved high performance even under simple parameterization (C = 0,1, linear kernel). Such consistency is particularly advantageous in primary healthcare contexts, where practical deployment requires models that are not only accurate but also computationally efficient and stable.

Table 4. Comparison of Algorithm Performance with SMOTE (Default Parameters)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
AdaBoost	83,04	84,17	70,43	73,70	93,61
Decision Tree	81,25	83,04	72,52	76,36	82,86
K-NN	83,04	81,75	77,64	79,27	90,36
LightGBM	81,25	80,21	70,91	74,27	94,46
Naïve Bayes	82,14	77,93	75,64	76,10	95,97
Random Forest	80,36	77,10	70,27	72,71	95,88
SVM	86,61	84,90	79,65	81,85	96,07
XGBoost	80,36	77,10	70,27	72,71	95,37

Table 5. Optimized Parameters and Best Accuracy

Model	Best Parameters	Best Accuracy (%)
Naïve Bayes	Default	79,46
K-NN	{'n_neighbors': 3}	83,75
SVM	{'C': 0.1, 'kernel': 'linear'}	88,59
Decision Tree	{'criterion': 'entropy', 'max_depth': 10}	85,44
Random Forest	{'max_depth': 10, 'n_estimators': 50}	91,51
AdaBoost	{'learning_rate': 0.5, 'n_estimators': 50}	85,56
XGBoost	{'learning_rate': 0.1, 'max_depth': 3, …}	91,02
LightGBM	{'learning_rate': 0.1, 'n_estimators': 100…}	91,27

These findings are consistent with evidence from previous studies, which have highlighted SVM as one of the most reliable algorithms for medical diagnosis and demonstrated its robustness in handling imbalanced clinical datasets [14]. Overall, this research reinforces the position of SVM as a strong candidate for predictive modeling of hepatitis, particularly when the objective is to provide accurate, stable, and easily deployable solutions in resource-constrained healthcare environments.

3.5. Evaluation

The evaluation was conducted using accuracy, precision, recall, F1-score, AUC, and the confusion matrix. The results presented in Table 6 indicate that SVM achieved an accuracy of 87% with a macro-F1 of 0,84, reflecting balanced performance across the four classes. For the majority of classes, such as Chronic Hepatitis, SVM obtained a precision of 0,92 and a recall of 0,90. In contrast, for minority classes such as Viral/Parasitic Infection, recall was lower (0,75), yet precision reached 1,00. This demonstrates that SVM adopts a conservative strategy in classifying underrepresented categories, where some cases may be missed, but positive predictions are almost always correct.

Table 6. Classification Report of SVM Model

Label	Precision	Recall	F1-Score	Support	AUC
Liver Abscess	0,71	0,77	0,74	13	0,94
Acute Hepatitis	0,83	0,87	0,85	39	0,92
Chronic Hepatitis	0,92	0,90	0,91	52	0,95
Viral/Parasitic Infection	1,00	0,75	0,86	8	0,96
Accuracy			0,87	112	

This advantage makes SVM more consistent compared to other algorithms. While Random Forest and XGBoost were able to achieve higher accuracy after parameter tuning, they tended to trade precision for recall in minority classes. In contrast, SVM maintained a better balance, producing a more reliable model for early screening contexts. These findings are consistent with earlier studies emphasizing the importance of handling minority class imbalance, as well as evidence highlighting SVM's robustness in maintaining predictive stability when applied to imbalanced medical datasets [15] [36].

3.6. Deployment

The deployment stage was carried out to implement the best-performing model, namely SVM, into a web-based prototype application. The application was developed using Python, supported by several libraries: scikit-learn for modeling, *imblearn* for data balancing with SMOTE, joblib for model storage, and *Streamlit* for building the user interface.

The application architecture was designed to enable users to perform interactive hepatitis diagnosis through three main menus:

1. Home, providing general information about the application.
2. Diagnosis, allowing both batch predictions (via Excel file upload) (see Figure 1) and manual predictions (through a form for individual patient symptoms) (see Figure 2).

Hasil Prediksi

	Gatal	Edema/Ascites	Diare/Gangguan Pencernaan	Berat Badan Turun	Ruam/Nyeri Sendi	Menggigil	Prediksi
0	Tidak	Ya	Tidak	Ya	Tidak	Tidak	Abses Hati
1	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Abses Hati
2	Ya	Ya	Tidak	Ya	Tidak	Tidak	Hepatitis Kronis
3	Tidak	Ya	Ya	Ya	Ya	Tidak	Infeksi Parasit atau Virus
4	Tidak	Tidak	Tidak	Ya	Tidak	Tidak	Abses Hati
5	Tidak	Tidak	Tidak	Tidak	Tidak	Ya	Hepatitis Akut
6	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Hepatitis Akut
7	Ya	Tidak	Ya	Tidak	Ya	Tidak	Hepatitis Kronis
8	Ya	Ya	Tidak	Ya	Ya	Tidak	Hepatitis Kronis
9	Tidak	Ya	Tidak	Tidak	Tidak	Tidak	Hepatitis Kronis

Figure 1. Prediction Results of Via Excel File Upload

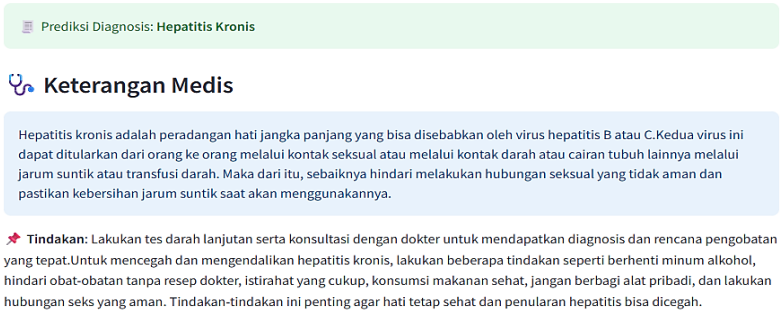

Figure 2. Diagnosis Prediction Results of Manual Predictions

3. Testing with New Data, offering flexibility to retrain the model using newly uploaded datasets (see Figure 3).

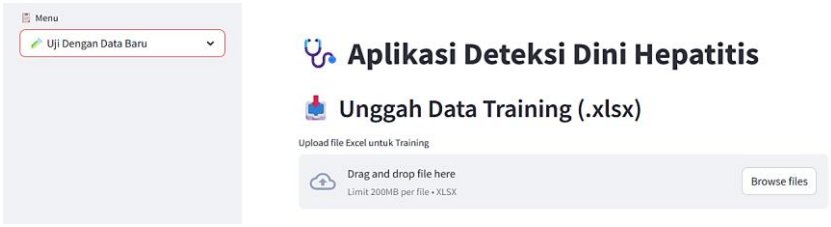

Figure 3. New Data Testing Module

Prediction results are presented in diagnostic tables that can be downloaded for documentation, while manual input provides not only predicted categories but also explanatory notes regarding the type of hepatitis, associated risk factors, and recommended actions. The system can be integrated into hospital information systems, primary healthcare centers, or private clinics to support early screening and decision-making. Its simple input requirements also make it suitable as a public health education tool for raising awareness of early hepatitis symptoms.

Despite its potential, the system's accuracy depends on data quality, and validation has so far been limited to a single hospital, requiring broader testing to ensure generalizability. Since the model relies on

non-laboratory clinical data, it cannot substitute formal medical examinations but should serve as a complementary decision-support tool. Overall, the deployment stage highlights the practical contribution of this research by providing a system that supports healthcare professionals while empowering the public through accessible early screening and health education.

3.7. Discussion and Comparative Analysis

The findings demonstrate that SVM achieved the highest performance, with an accuracy of 87% and a macro-F1 score of 0,84. This result can be attributed to SVM's capacity to manage high-dimensional clinical data while preserving balanced performance across classes, thereby ensuring greater consistency compared to other algorithms that required extensive parameter tuning to achieve optimal outcomes. These results are consistent with earlier studies that emphasize the robustness of SVM in medical datasets, yet they diverge from others that identify ensemble methods, such as Random Forest, as superior. Such contrasts highlight the critical role of dataset characteristics and preprocessing strategies, including the application of SMOTE, in shaping algorithmic performance.

The principal strength of this study lies in its utilization of non-laboratory clinical data, which offers greater accessibility and cost-effectiveness, alongside the development of a web-based application that demonstrates practical readiness for real-world deployment as an early screening tool. Nonetheless, certain limitations must be acknowledged the dataset was derived from a single healthcare institution, restricting generalizability, and the reliance on non-laboratory attributes precludes substitution for formal medical examinations. Overall, this research not only reaffirms the effectiveness of SVM but also advances novelty through the integration of non-laboratory attributes, contributing practical value by offering an accessible decision-support tool for early hepatitis detection in primary healthcare settings.

4. CONCLUSION

The objectives stated in the Introduction have been successfully achieved. Non-laboratory clinical factors demonstrated a significant contribution to improving hepatitis prediction, while the SVM exhibited the highest consistency with an accuracy of 87%. These findings confirm the premise that machine learning can serve as a reliable approach for early disease detection. For future work, the model may be further enhanced by incorporating larger and more diverse datasets, including laboratory and genetic factors, as well as by integrating the SVM prototype into hospital information systems or mobile health applications. Furthermore, subsequent studies may investigate the use of ensemble techniques or deep learning methods to further enhance predictive accuracy and robustness.

REFERENCES

[1] S. C. R. Nandipati, C. Xinying, and K. K. Wah, "Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques," *Appl. Model. Simul.*, vol. 4, pp. 89–100, 2020.

[2] W. H. O. WHO, "WHO sounds alarm on viral hepatitis infections claiming 3500 lives each day." [Online]. Available: https://www.who.int/news/item/09-04-2024-who-sounds-alarm-on-viral-hepatitis-infections-claiming-3500-lives-each-day

[3] P. Lusita, N. Indriani, H. Anggraini, and S. Handayani, "Faktor-Faktor yang Mempengaruhi Kejadian Hepatitis pada Ibu Hamil," 2021.

[4] J. Wu, H. Wang, Z. Xiang, C. Jiang, Y. Xu, and G. Zhai, "Role of viral hepatitis in pregnancy and its triggering mechanism," *J. Transl. Intern. Med.*, vol. 12, no. 4, 2024, doi: 10.2478/jtim-2024-0015.

[5] V. Harabor *et al.*, "Machine Learning Approaches for the Prediction of Hepatitis B and C Seropositivity," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, 2023, doi: 10.3390/ijerph20032380.

[6] R. Kemenkes, "Angka Hepatitis B dan C di Indonesia Turun." [Online]. Available: https://www.kemkes.go.id/eng/angka-hepatitis-b-dan-c-di-indonesia-turun

[7] D. Andriani, "Ini Cara Cegah Hepatitis Ala Dr dr Rino Alvani Gani, Sp.PD-KGEH," Bisnis.com. Accessed: Jul. 23, 2025. [Online]. Available: https://lifestyle.bisnis.com/read/20170831/106/685952/ini-cara-cegah-hepatitis-ala-dr-dr-rino-alvani-gani-sp.pd-kgeh-

[8] Kemenkes, "Hepatitis Akut Menular Lewat Saluran Cerna dan Saluran Pernafasan."

[9] D. Singh, D. Prashar, J. Singla, A. A. Khan, M. Al-Sarem, and N. A. Kurdi, "Intelligent Medical Diagnostic System for Hepatitis B," *Comput. Mater. Contin.*, vol. 73, no. 3, pp. 6047–6068, 2022, doi: 10.32604/cmc.2022.031255.

[10] Z. Xia, L. Qin, Z. Ning, and X. Zhang, "Deep learning time series prediction models in surveillance data of hepatitis incidence in China," *PLoS One*, vol. 17, no. 4 April, pp. 1–18, 2022, doi: 10.1371/journal.pone.0265660.

[11] A. Firdaus, "Menggugah kesadaran global atasi hepatitis yang kian mengancam," Antaranews.com. Accessed: Jul. 23, 2025. [Online]. Available: https://ambon.antaranews.com/berita/217908/menggugah-kesadaran-global-atasi-hepatitis-yang-kian-mengancam

[12] I. Cholissodin and A. A. Soebroto, *AI , Machine Learning & Deep Learning Book (Teori & Implementasi)*, no. July 2019. Malang, 2021. [Online]. Available: https://www.researchgate.net/publication/348003841

[13] A. Zein *et al.*, "Pengenalan Pembelajaran Mesin dan Deep Learning.," *J. Stud. Alquran dan Tafsir*, vol. 4, no. 1,

pp. 29–38, 2023, [Online]. Available: https://jurnal.pranataindonesia.ac.id/index.php/jik/article/download/96/49
[14] I. I. Ahmed, D. Y. Mohammed, and K. A. Zidan, "Diagnosis of hepatitis disease using machine learning techniques," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, pp. 1564–1572, 2022, doi: 10.11591/ijeecs.v26.i3.pp1564-1572.
[15] N. Sharfina and N. G. Ramadhan, "Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 8, no. 1, p. 33, 2023, doi: 10.31328/jointecs.v8i1.4456.
[16] A. Damayanti and G. Testiana, "Penerapan Data Mining untuk Prediksi Penyakit Hepatitis C Menggunakan Algoritma Naïve Bayes," *J. Manaj. Inform. Jayakarta*, vol. 3, no. 2, pp. 177–186, 2023, doi: 10.52362/jmijayakarta.v3i2.1098.
[17] A. D. Putra, D. Nurani, M. M. Dewi, and S. Alfie Nur Rahmi, "Supervised Machine Learning Model untuk Prediksi Penyakit Hepatitis," *Indones. J. Comput. Sci.*, vol. 13, no. 2, pp. 3329–3341, 2024, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135
[18] M. Diqi, M. E. Hiswati, and E. Damayanti, "Enhancing Hepatitis Patient Survival Detection : A Comparative Study of CNN and Traditional Machine Learning Algorithms," *CoreIT*, vol. 10, no. 1, pp. 21–31, 2024.
[19] M. Hussain *et al.*, "Rapid Detection System for Hepatitis B Surface Antigen (HBsAg) Based on Immunomagnetic Separation, Multi-Angle Dynamic Light Scattering and Support Vector Machine," *IEEE Access*, vol. 8, pp. 107373–107386, 2020, doi: 10.1109/ACCESS.2020.3000357.
[20] N. L. W. S. R. Ginantra *et al.*, *Data Mining dan Penerapan Algoritma*. 2021.
[21] Mustika *et al.*, *Data Mining dan Aplikasinya*. 2021. [Online]. Available: https://repository.penerbitwidina.com/uk/publications/351768/data-mining-dan-aplikasinya
[22] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. doi: 10.1016/C2009-0-61819-5.
[23] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Mach. Learn.*, no. January, 2012, doi: 10.1007/978-1-4419-9326-7.
[24] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7473 LNCS, pp. 246–252, 2012, doi: 10.1007/978-3-642-34062-8_32.
[25] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, 3rd ed. Waltham: Morgan Kaufmann, 2012.
[26] M. I. Hossain, "Support Vector Machine," *Mach. Learn.*, vol. 104, no. 14, pp. 33–36, 2023, doi: 10.18411/trnio-12-2023-769.
[27] C. Zhang and Y. Ma, *Ensemble machine learning: Methods and applications*. 2012. doi: 10.1007/9781441993267.
[28] A. J. Ferreira and M. A. T. Figueiredo, "Boosting algorithms: A review of methods, theory, and applications," *Ensemble Mach. Learn. Methods Appl.*, pp. 35–85, 2012, doi: 10.1007/9781441993267_2.
[29] Z. A. Ali, Z. H. Abduljabbar, H. A. Taher, A. B. Sallow, and S. M. Almufti, "Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review," *Acad. J. Nawroz Univ.*, vol. 12, no. 2, pp. 320–334, 2023.
[30] I. A. Febriansyah, A. Id Hadiana, and F. Rakhmat Umbara, "Prediksi Curah Hujan Menggunakan Metode Categorical Boosting (Catboost)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2930–2937, 2024, doi: 10.36040/jati.v7i4.7304.
[31] A. S. Zuhri, S. Kom, M. Kom, S. Kom, and M. Kom, "Comparison of Boosting Algorithms (LightGBM , CatBoost , and XGBoost) on Ship Ticket Sales Prediction," *Int. Conf. Artif. Intell. Navig. Eng. Aviat. Technol. ISSN*, vol. 1, no. 1, 2024.
[32] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi: 10.36456/jstat.vol15.no2.a5548.
[33] R. K. Dinata and N. Hasdyna, "Machine Learning Panduan Memahami Data Science, Supervised Learning, Unsupervised Learning dan Reinforcement Learning," 2020, *Unimal Press, Sulawesi*.
[34] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," *Technologies*, vol. 9, no. 4, 2021, doi: 10.3390/technologies9040081.
[35] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," *arXiv Prepr.*, pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756
[36] R. Alizadehsani *et al.*, "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Sci. Data*, vol. 6, no. 1, pp. 1–13, 2019, doi: 10.1038/s41597-019-0206-3.

BIBLIOGRAPHY OF AUTHORS

Maya Gian Sister, currently a Master's student in Computer Science at the Faculty of Information Technology, Universitas Budi Luhur. The research interests are directed toward data mining, machine learning, and their applications in healthcare and decision support systems. The ongoing academic focus is on developing predictive models and exploring computational approaches to enhance data-driven analysis.

Yulia Nita, currently a Master's student in Computer Science at the Faculty of Information Technology, Universitas Budi Luhur. The author has a strong interest in data mining, machine learning, and information systems, particularly their application in technology-based public health. Current studies are focused on developing predictive models and exploring computational approaches to support community health and informed decision-making.

Dr. Ir. Achmad Solichin, S.Kom., M.T.I., Dean of the Faculty of Information Technology at Universitas Budi Luhur. Areas of expertise include computer vision, image processing, artificial intelligence, and web-based system development. Research focuses on the application of intelligent systems in education, healthcare, and digital governance. Numerous publications have been presented in national and international journals, and recognition has been awarded as one of Indonesia's top scientists in the AD Scientific Index 2022. Academic activities also involve leadership roles and curriculum development in computer science and information technology.