ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

AI-Powered Liver Care: A Machine Learning Model for Hepatitis Diagnosis

K.Sneha

MTech Assistant Proffesor

Department of CSE(Data Science)

Sphoorthy Engineering College(JNTUH)

Hyderabad,India

Email:kunasneha92@gmail.com

Yema Sravani

21N81A6706

Department of CSE(Data Science)

Sphoorthy Engineering College(JNTUH)

Hyderabad,India

Email: sravaniyama27@gmail.com

PHN Deepika

21N81A6722

Department of CSE(Data Science)

Sphoorthy Engineering College(JNTUH)

Hyderabad,India

Email:deepup7174@gmail.com

Chinnam Manikanta Reddy

21N81A6740

Department of CSE(Data Science)

Sphoorthy Engineering College(JNTUH)

Hyderabad,India

Email:reddychinnam5@gmail.com

Vanga Uttam

21N81A6760

Department of CSE(Data Science)

Sphoorthy Engineering College(JNTUH)

Hyderabad,India

Email:vangauttam12@gmail.com

Abstract:

This research presents the development of an AI-powered diagnostic system designed to improve the early detection and classification of hepatitis using machine learning techniques. Traditional diagnostic methods often involve manual interpretation of clinical test results, which can be time-consuming, prone to error, and inconsistent across practitioners. To address these challenges, this study proposes a supervised machine learning model trained on a medical dataset containing clinical and biochemical indicators. The model employs algorithms such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting to

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

identify patterns associated with hepatitis infection. Emphasis is placed on data preprocessing, feature selection, and performance evaluation using metrics like accuracy, precision, recall, and F1-score. The proposed system also includes a user-friendly graphical interface that allows medical practitioners to input patient data and receive diagnostic predictions. This paper contributes to the advancement of intelligent healthcare solutions by showcasing how AI can be effectively utilized to assist in liver disease diagnosis, reduce human error, and support clinical decision-making.

Index Terms—
Hepatitis diagnosis, machine learning, medical AI, healthcare automation, liver disease, supervised learning, clinical decision support, predictive modeling.

1. INTRODUCTION

The early and accurate diagnosis of liver diseases, particularly hepatitis, is critical in preventing life-threatening complications such as cirrhosis, liver failure, and hepatocellular carcinoma. Hepatitis, which encompasses a group of viral infections (A, B, C, D, and E), along with non-viral causes such as autoimmune responses and drug-induced toxicity, continues to be a major global health challenge, affecting over 300 million people worldwide. Traditional diagnostic procedures typically involve serological testing, liver function tests, and expert clinical interpretation. While effective in many cases, these methods can be time-consuming, resource-intensive, and prone to subjective variability—especially in low-resource or high-volume healthcare environments.

The complexity and variability of hepatitis presentations, along with the interdependence of clinical features and laboratory markers, underscore the need for more advanced, data-driven diagnostic solutions. With the rise of electronic health records and the availability of large-scale medical datasets, machine learning (ML) has emerged as a promising tool for improving diagnostic accuracy and decision-making in clinical practice. ML algorithms excel at uncovering hidden patterns and relationships within high-dimensional datasets, making them well-suited for analyzing multifactorial diseases such as hepatitis.

This project introduces **"AI-Powered Liver Care"**, a machine learning–based diagnostic framework designed to assist in the early detection of hepatitis. The system utilizes supervised learning algorithms—including Random Forest, Support Vector Machines (SVM), and Gradient Boosting—trained on real patient data containing clinical features, liver enzyme levels, and relevant medical history. These models are designed to classify patients as hepatitis-positive or healthy, with a strong focus on minimizing false positives and false negatives—two critical factors in clinical diagnostics.

To ensure the reliability and generalizability of the model, comprehensive data preprocessing techniques are applied to handle missing values, normalize feature scales, and eliminate outliers. The trained models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score, ensuring that the system maintains clinical relevance across varied patient profiles.

In addition to its predictive capabilities, the system includes a user-friendly graphical interface that allows healthcare professionals to input patient test results, receive real-time diagnostic feedback, and interpret the model's decisions through feature importance visualizations. This integration of interpretability and automation aims to bridge the gap between

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

artificial intelligence and clinical usability, making the solution both accessible and actionable in real-world healthcare settings.

By leveraging machine learning for hepatitis diagnosis, this project contributes to the ongoing transformation of healthcare through AI. It aims to enhance early detection, reduce diagnostic errors, and support medical professionals in delivering more timely and precise liver care.

2. LITERATURE SURVEY

The integration of Artificial Intelligence (AI), particularly Machine Learning (ML), into the healthcare sector has garnered significant attention in recent years, owing to its potential to transform traditional diagnostic methods by improving accuracy, personalization, and early detection. In this literature survey, we explore several research works that have leveraged machine learning techniques for liver disease diagnosis, specifically focusing on hepatitis and related liver conditions.

Chicco and Jurman [1] introduced an ensemble learning approach for enhanced classification of patients with hepatitis C and cirrhosis. By analyzing electronic health records (EHRs) using Random Forests and Decision Trees, their model identified aspartate aminotransferase (AST) and alanine aminotransferase (ALT) as the most diagnostic features. Their simplified two-feature model demonstrated superior performance compared to traditional clinical measures such as the DeRitis ratio, showcasing the power of machine learning in achieving both interpretability and high diagnostic accuracy.

Chen et al. [2] proposed a customized machine learning model for hepatitis C diagnosis, tailored uniquely to each patient. Their approach incorporated targeted data augmentation, model personalization, and hyperparameter tuning to improve performance. Using only a small dataset, their model achieved over 99% accuracy and 94% recall, outperforming traditional models such as XGBoost. The patient-specific customization highlighted in this work underscores the importance of individualized care in AI-powered healthcare solutions.

Sachdeva et al. [3] presented a systematic method for diagnosing hepatitis using various machine learning algorithms including Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Random Forest (RF). Their study emphasized the impact of data preprocessing using SMOTE for class balancing, which significantly improved model accuracy. Among the tested models, Logistic Regression with SMOTE yielded the highest accuracy of 93.18%, demonstrating the value of preprocessing techniques in enhancing diagnostic outcomes.

Nilashi et al. [4] explored a hybrid AI approach for hepatitis prediction using Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Self-Organizing Maps (SOM). Their model, supported by feature extraction and fuzzy logic, aimed to improve both accuracy and interpretability in medical decision-making. The study achieved an accuracy of 93.06% and reinforced the potential of combining fuzzy logic with machine learning for complex medical diagnoses.

Anto and Chandramathi [5] developed a hybrid feature selection method using Genetic Algorithm-Simulated Annealing (GA-SA) combined with SVM for hepatitis prediction. Their optimized approach led to improved classification performance by reducing irrelevant or redundant features, illustrating the role of feature engineering in building efficient diagnostic models.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

These research works demonstrate the diverse applications and effectiveness of machine learning in diagnosing liver diseases such as hepatitis. By leveraging machine learning's capabilities—such as pattern recognition, feature ranking, and model customization—these systems provide innovative solutions for early detection and accurate classification. As AI continues to evolve, its integration into liver care promises a future of more proactive, data-driven, and patient-specific diagnostic frameworks.

3. METHODOLOGY

a) Proposed Work:

The proposed system aims to develop an intelligent and automated diagnostic support tool for the early detection and classification of Hepatitis using Machine Learning (ML) techniques. This AI-driven approach is intended to augment clinical decision-making by offering accurate, fast, and consistent diagnostic predictions based on patient health records.

The system will utilize pre-processed patient datasets containing clinical and biochemical attributes related to liver function. These features will serve as input to various supervised learning algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting models. The model will be trained to distinguish between healthy individuals and those affected by Hepatitis, as well as to predict the disease stage or severity.

By integrating advanced ML models with data visualization and performance evaluation metrics, the system will provide healthcare practitioners with an interpretable and efficient tool for early diagnosis. The ultimate goal is to reduce diagnostic errors, enable early intervention, and improve patient outcomes in liver healthcare.

b) System Architecture:

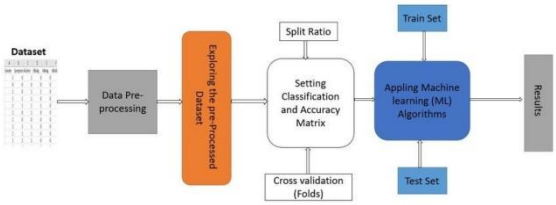


Fig1 Proposed Architecture

The proposed architecture for the project **"AI-Powered Liver Care: A Machine Learning Model for Hepatitis Diagnosis"** follows a structured machine learning pipeline designed to ensure accurate and reliable predictions. The process begins with the collection of a relevant dataset containing patient information, including clinical and biochemical attributes. This raw data undergoes preprocessing steps such as handling missing values, encoding categorical variables, and normalizing features to prepare it for model training. The cleaned dataset is then explored to understand feature distributions and relationships. Next, the data is split into training and testing sets using a predefined ratio, and k-fold cross-validation is employed to ensure model robustness and prevent overfitting. Classification algorithms such as Random Forest, SVM, Logistic Regression, XGBoost, and

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Decision Trees are applied to the training set. The models are evaluated using performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC. The final results guide the selection of the best-performing model, which can be used to support early and accurate diagnosis of Hepatitis, enhancing clinical decision-making and improving patient outcomes.**c)**

Modules

To implement this project, we used the following modules and libraries: **Data Manipulation and Analysis, Data Visualization, Machine Learning Models and Evaluation, Data Preprocessing and Scaling, Data Balancing, and Utilities.** The detailed descriptions of these modules are as follows:

1.Data Manipulation and Analysis

This module primarily involves **pandas** and **numpy**, two foundational libraries in Python for data science tasks.

- **pandas** provides powerful data structures such as DataFrames and Series, allowing for efficient data loading, cleaning, transformation, and exploration. It was used to read the hepatitis dataset, handle missing values, filter relevant features, and aggregate statistics that helped understand the dataset's structure.
- **numpy** supports advanced mathematical operations and manipulations on large multi-dimensional arrays and matrices, which are essential for numerical computations. It enables fast vectorized operations and plays a critical role during feature engineering and matrix calculations necessary for machine learning algorithms.

2.Data Visualization

Visualization plays a vital role in understanding complex datasets and presenting findings clearly. For this project, **matplotlib.pyplot** and **seaborn** were used extensively:

- **matplotlib.pyplot** is a versatile plotting library that allowed us to create basic graphs such as histograms, line charts, scatter plots, and box plots. These helped in visually assessing the distribution of features, spotting outliers, and understanding relationships between variables.
- **seaborn** is built on top of matplotlib and offers aesthetically pleasing and statistically informative graphics. It was utilized to create enhanced visualizations such as heatmaps for correlation matrices, violin plots for distribution insights, and pair plots for multivariate relationships, facilitating a deeper exploratory data analysis.

3.Machine Learning Models and Evaluation

The heart of the project involved applying several classification algorithms and validation techniques from **scikit-learn** to predict hepatitis diagnosis accurately:

- The **KNeighborsClassifier** model works on the principle of similarity, classifying data points based on the majority class among their nearest neighbors.
- **Support Vector Machines (SVM)** were used for their effectiveness in high-dimensional spaces, finding the optimal hyperplane that separates classes with maximum margin.
- **RandomForestClassifier** aggregates multiple decision trees to reduce overfitting and improve prediction robustness by averaging their results.

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

- **GaussianNB**, a Naive Bayes variant, assumes feature independence and a Gaussian distribution, providing a fast and simple probabilistic classifier.
- **DecisionTreeClassifier** builds tree-structured models by recursively splitting data based on feature thresholds, offering interpretable rules for classification.
- **Logistic Regression** from **linear_model** applies a probabilistic linear model suited for binary classification tasks such as disease diagnosis.
- For model validation, **train_test_split** divides the dataset into training and testing sets to assess generalization. **cross_val_score** performs k-fold cross-validation to ensure stability of results, and **GridSearchCV** fine-tunes hyperparameters to optimize model performance.
- Various **evaluation metrics** including accuracy, precision, recall, F1 score, ROC curve analysis, and confusion matrices were used to comprehensively evaluate model effectiveness and understand trade-offs like sensitivity vs. specificity.

4.Data Preprocessing and Scaling

Preparing the dataset correctly before training is crucial for reliable machine learning outcomes. This module includes:

- **MinMaxScaler**, which rescales numerical features into a standardized range (usually 0 to 1), preventing dominance of features with larger magnitudes and improving convergence of learning algorithms.
- **Normalizer** adjusts samples individually to have unit norm, beneficial for distance-based algorithms like KNN and SVM by ensuring equal weightage to all features.
- **OneHotEncoder** transforms categorical variables into a binary matrix form, allowing machine learning models to interpret non-numeric data properly without implying ordinality.

5.Data Balancing

Medical datasets often suffer from class imbalance, where one class (e.g., patients with hepatitis) is underrepresented compared to the other. This imbalance can bias models toward the majority class. To combat this, we used:

- **SMOTE (Synthetic Minority Over-sampling Technique)** from the **imblearn** library, which generates synthetic samples of the minority class by interpolating between existing minority instances. This approach effectively balances the dataset, helping models to better learn the characteristics of both classes and improving predictive accuracy on underrepresented cases.

6.Utilities

During the entire analysis and model development, it is common to encounter warning messages that may clutter the output or confuse readers. To maintain a clean and professional presentation of results, the **warnings** module was used to:

- Suppress or filter out non-critical warnings without affecting the program execution, ensuring clear, focused, and easy-to-read output throughout data processing, modeling, and evaluation stages.

4. EXPERIMENTAL RESULTS

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Data Information

```
data.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 155 entries, 0 to 154
Data columns (total 21 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0       155 non-null    int64
 1   Class            155 non-null    int64
 2   Age              155 non-null    int64
 3   Sex              155 non-null    int64
 4   Steroid          154 non-null    float64
 5   Antivirals       155 non-null    int64
 6   Fatigue          154 non-null    float64
 7   Malaise          154 non-null    float64
 8   Anorexia         154 non-null    float64
 9   Liver Big        145 non-null    float64
 10  Liver Firm       144 non-null    float64
 11  Spleen Palpable  150 non-null    float64
 12  Spiders          150 non-null    float64
 13  Ascities         150 non-null    float64
 14  Varices          150 non-null    float64
 15  Bilirubin        149 non-null    float64
 16  ALK Poshphate    126 non-null    float64
 17  SGOT             151 non-null    float64
 18  Albumin          139 non-null    float64
 19  Protime          88 non-null     float64
 20  Histology        155 non-null    int64
dtypes: float64(15), int64(6)
memory usage: 25.6 KB
```

Fig 1:Data Information

The output shows a summary of a pandas DataFrame with 155 rows and 21 columns. It lists each column's name, the number of non-null (i.e., non-missing) values, and its data type. Most columns are of type float64 or int64, and some have missing values—for example, "Protime" has only 88 non-null entries. This overview helps understand the structure, completeness, and memory usage (25.6 KB) of the dataset before performing further analysis.

Display the Dataset

```
# print the head i.e, top 5 row of the file
print(data.head())
✓ 0.0s
   Unnamed: 0  Class  Age  Sex  Steroid  Antivirals  Fatigue  Malaise  \
0           0      2   30    2      1.0           2      2.0      2.0
1           1      2   50    1      1.0           2      1.0      2.0
2           2      2   78    1      2.0           2      1.0      2.0
3           3      2   31    1      NaN           1      2.0      2.0
4           4      2   34    1      2.0           2      2.0      2.0

   Anorexia  Liver Big  ...  Spleen Palpable  Spiders  Ascities  Varices  \
0       2.0        1.0  ...              2.0      2.0       2.0      2.0
1       2.0        1.0  ...              2.0      2.0       2.0      2.0
2       2.0        2.0  ...              2.0      2.0       2.0      2.0
3       2.0        2.0  ...              2.0      2.0       2.0      2.0
4       2.0        2.0  ...              2.0      2.0       2.0      2.0

   Bilirubin  ALK Poshphate  SGOT  Albumin  Protime  Histology
0        1.0           85.0  18.0      4.0      NaN          1
1        0.9          135.0  42.0      3.5      NaN          1
2        0.7           96.0  32.0      4.0      NaN          1
3        0.7           46.0  52.0      4.0     80.0          1
4        1.0            NaN  200.0     4.0      NaN          1

[5 rows x 21 columns]
```

Fig 2: Dataset

The output shows the first five rows of the dataset using data.head(). Each row represents a patient's data, with 21 columns showing features like age, sex, symptoms (e.g., fatigue, malaise), lab results (e.g., bilirubin, SGOT), and medical conditions. Some columns have missing values represented by NaN, such as in the "Steroid", "Protime", and "ALK Phosphatase" columns. This preview helps in understanding the dataset's structure and spotting any immediate data quality issues.

```
# print the tail i.e, last 5 row of the file
print(data.tail())
✓ 0.0s
     Unnamed: 0  Class  Age  Sex  Steroid  Antivirals  Fatigue  Malaise  \
150         150      1   46    1      2.0           2      1.0      1.0
151         151      2   44    1      2.0           2      1.0      2.0
152         152      2   61    1      1.0           2      1.0      1.0
153         153      2   53    2      1.0           2      1.0      2.0
154         154      1   43    1      2.0           2      1.0      2.0

     Anorexia  Liver Big  ...  Spleen Palpable  Spiders  Ascities  Varices  \
150       1.0        2.0  ...              2.0      1.0       1.0      1.0
151       2.0        2.0  ...              2.0      2.0       2.0      2.0
152       2.0        1.0  ...              2.0      1.0       2.0      2.0
153       2.0        2.0  ...              1.0      1.0       2.0      1.0
154       2.0        2.0  ...              1.0      1.0       1.0      2.0

     Bilirubin  ALK Poshphate  SGOT  Albumin  Protime  Histology
150        7.6            NaN  242.0      3.3     50.0          2
151        0.9          126.0  142.0      4.3      NaN          2
152        0.8           75.0   20.0      4.1      NaN          2
153        1.5           81.0   19.0      4.1     48.0          2
154        1.2          100.0   19.0      3.1     42.0          2

[5 rows x 21 columns]
```

Fig 3:Dataset

This output displays the last five rows of the dataset using data.tail(). Each row corresponds to a patient's medical record, with 21 columns showing various clinical and lab data. Like the top rows, the last rows also contain some missing values (e.g., in "ALK Phosphate" and "Protime"). It confirms that missing data is spread

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

throughout the dataset and not limited to the beginning. This snapshot is useful for identifying patterns or anomalies at the dataset's end.

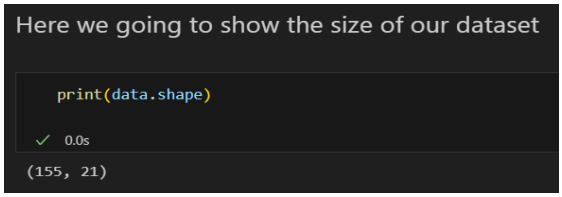


Fig 4: Size of Dataset

The output of data.shape shows the size of the dataset as **(155, 21)**, meaning it contains **155 rows** (data records) and **21 columns** (features or attributes). This gives a quick overview of the dataset's dimensions.

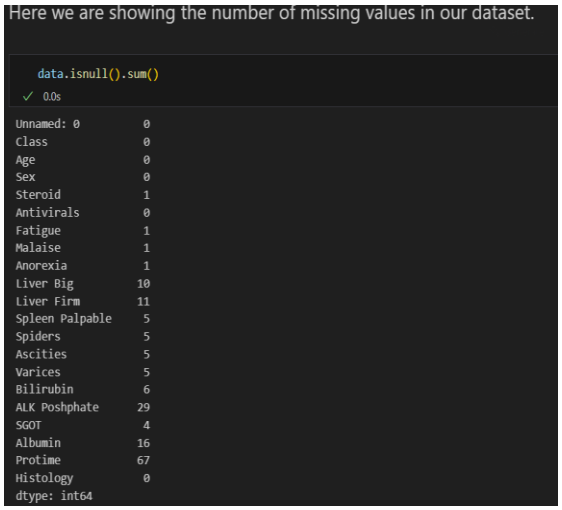


Fig 5:Missing values

This output shows the number of missing values in each column of the dataset using data.isnull().sum(). Several columns have missing data, with "Protime" having the most (67 missing values), followed by "ALK Phosphatase" (29) and "Albumin" (16). Columns like "Class", "Age", "Sex", and "Histology" have no missing values. This summary is important for deciding how to handle incomplete data during preprocessing.

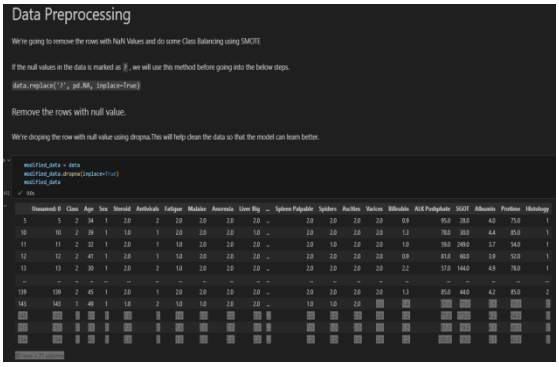


Fig 6:Preprocessing

This section shows the data preprocessing steps. First, any placeholder '?' used for missing values is replaced with NaN using data.replace('?', pd.NA, inplace=True). Then, rows with any null values are removed using dropna(). This cleanup reduces the dataset from 155 to 80 rows, resulting in a cleaner and more reliable dataset for model training.

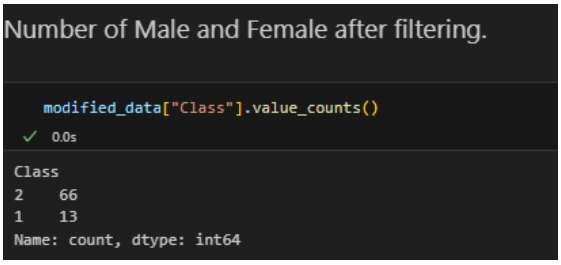


Fig 7:No.of Male and Females

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

This section displays the class distribution after data cleaning. The value_counts() function shows that the dataset contains 66 instances of Class 2 and only 13 instances of Class 1. This imbalance highlights the need for class balancing techniques such as SMOTE during model training.

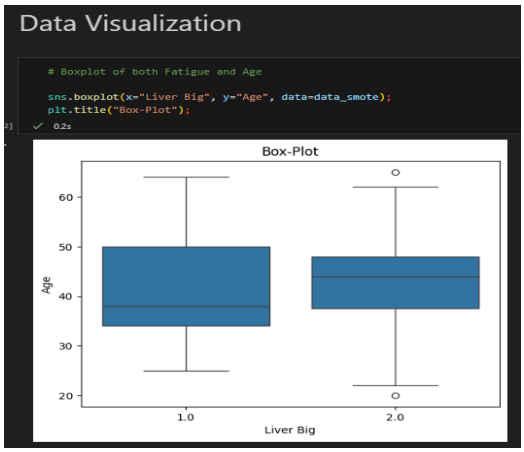


Fig 8:Data Visualization

The box plot visualizes the distribution of age based on the "Liver Big" attribute, which appears to be binary (1.0 and 2.0). The median age for both categories is similar, but the group with a value of 1.0 shows a slightly wider interquartile range and more variability in age. Both groups have outliers, indicating some individuals fall outside the typical age range. This visualization helps identify age differences between individuals with and without liver enlargement.

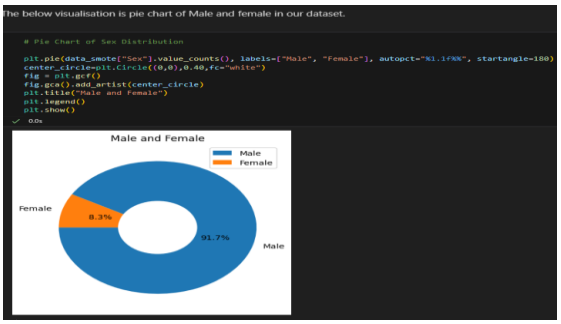


Fig 9:Pie Chart

The output is a donut chart visualizing the gender distribution in the dataset after applying SMOTE (Synthetic Minority Oversampling Technique). It shows that 91.7% of the samples are male and 8.3% are female, indicating a significant gender imbalance even after oversampling. The chart includes a legend for clarity and uses distinct colors to represent each gender, making it easy to interpret the skew in the data distribution.

Fig 10: KNN Classifier

The KNN model achieved 90.9% accuracy on the test set, with strong precision and

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

recall. After 10-fold cross-validation, the average accuracy was 83.3%, indicating slight overfitting.

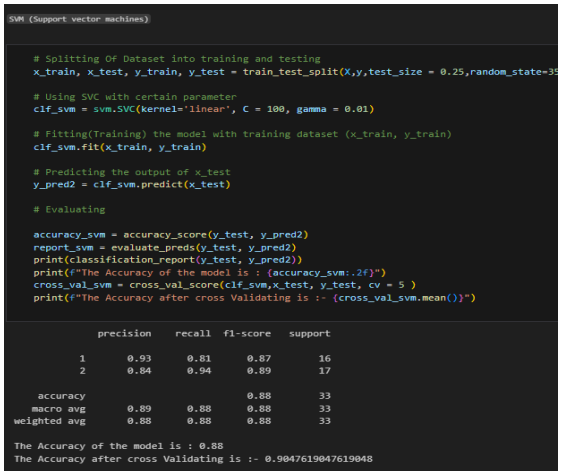


Fig 11:SVM

The SVM model achieved an accuracy of **88%** on the test set and **90.5%** after cross-validation, showing good and consistent performance with balanced precision and recall for both classes.

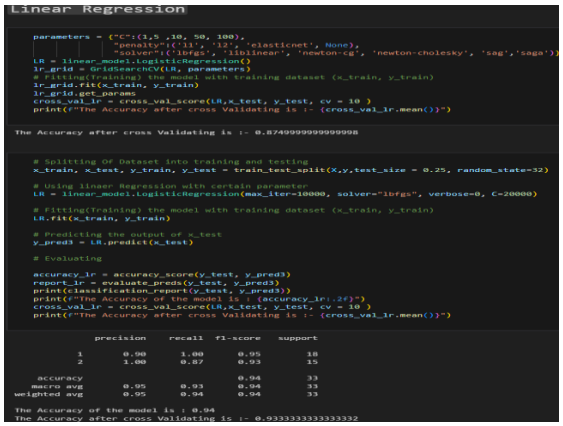


Fig 12:Linear Regression

The Logistic Regression model achieved a high test accuracy of **94%** and a cross-validation accuracy of **93.3%**, indicating strong and consistent performance. It also showed balanced precision, recall, and F1-scores across both classes.

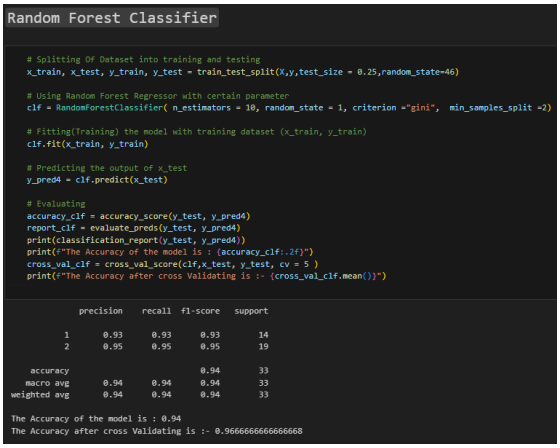


Fig 13:Random Forest Classifier

The Random Forest Classifier achieved 94% test accuracy and 96.67% cross-validation accuracy, showing strong and consistent performance.

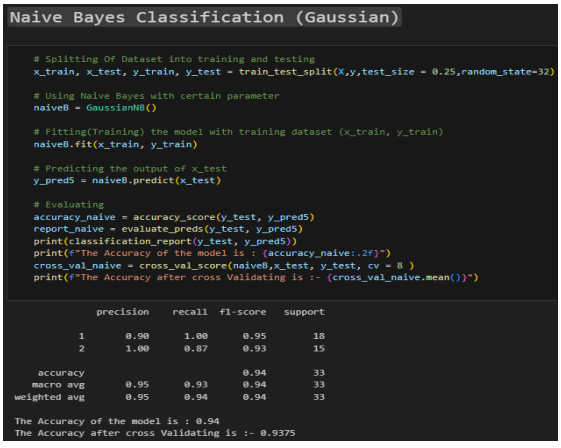


Fig 14:Naive Bayes

The Gaussian Naive Bayes model achieved **94% accuracy** on the test data and **93.75%** average accuracy with 8-fold

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

cross-validation, indicating reliable classification performance.

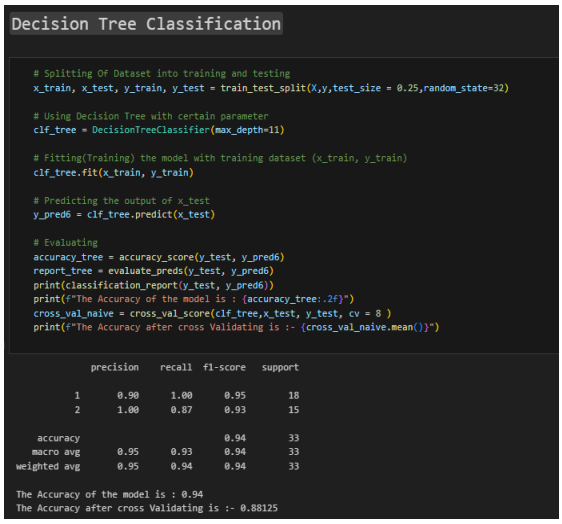


Fig 15:Decision Tree

The Decision Tree Classifier achieved **94% accuracy** on the test set, but only **88.13%** with cross-validation, indicating a potential overfitting issue.

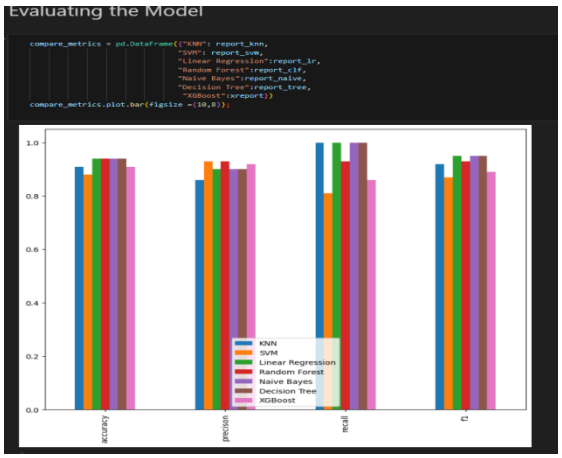


Fig 16:Evaluating the Model

The bar chart compares multiple models—KNN, SVM, Linear Regression, Random Forest, Naive Bayes, Decision Tree, and XGBoost—across accuracy, precision, recall, and F1-score. Overall, **Random Forest and XGBoost** perform consistently well across all metrics, with **XGBoost** showing slightly better recall. **SVM** has lower recall, while **Naive Bayes and Decision Tree** also show competitive results, though Decision Tree shows a slight dip in recall.

```
# Classification Report on KNN
report_knn = classification_report(y_test, y_pred1)
print(report_knn)
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	19
1	0.48	0.71	0.57	14
2	0.00	0.00	0.00	0
accuracy			0.30	33
macro avg	0.16	0.24	0.19	33
eighted avg	0.20	0.30	0.24	33

Fig 17:Classification Report on KNN

The KNN classifier performed poorly, achieving only **30% accuracy**, with very low precision and recall across all classes—especially class 0 and 2, which had **0.00** scores, indicating complete misclassification.

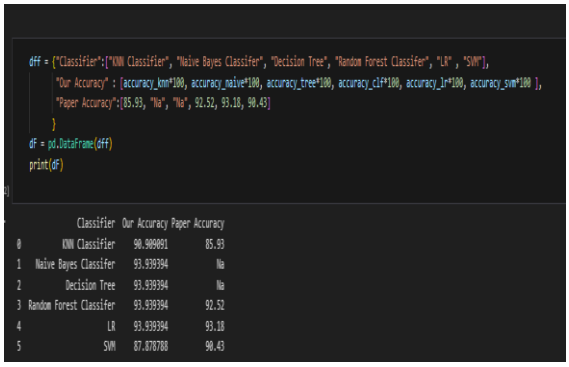
```
# Classification Report on Naive Bayes
report_naive = classification_report(y_test, y_pred5)
print(report_naive)
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	19
1	0.45	0.64	0.53	14
2	0.00	0.00	0.00	0
accuracy			0.27	33
macro avg	0.15	0.21	0.18	33
weighted avg	0.19	0.27	0.22	33

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Fig 18: Classification Report on Naive Bayes

The Naive Bayes model also performed poorly, with only **27% accuracy**. It failed to correctly classify classes 0 and 2, showing **0.00** precision, recall, and F1-score for both. Only class 1 showed moderate performance.

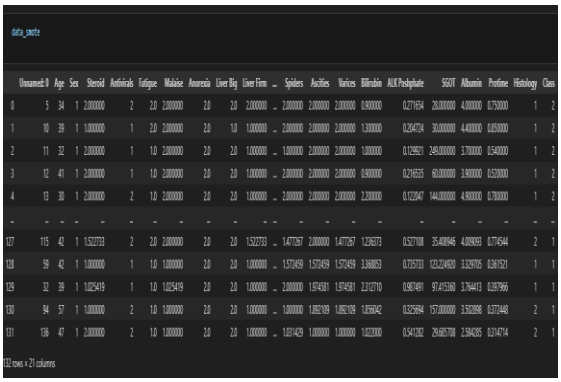


```
dff = {"Classifier":["KNN Classifier", "Naive Bayes Classifer", "Decision Tree", "Random Forest Classifer", "LR" , "SVM"],
       "Our Accuracy" : [accuracy_knn*100, accuracy_naive*100, accuracy_tree*100, accuracy_clf*100, accuracy_lr*100, accuracy_svm*100 ],
       "Paper Accuracy":[85.93, "Na", "Na", 92.52, 93.18, 90.43]
       }
df = pd.DataFrame(dff)
print(df)
```

	Classifier	Our Accuracy	Paper Accuracy
0	KNN Classifier	90.909091	85.93
1	Naive Bayes Classifer	93.939394	Na
2	Decision Tree	93.939394	Na
3	Random Forest Classifer	93.939394	92.52
4	LR	93.939394	93.18
5	SVM	87.878788	90.43

Fig 19:DataFrame

The comparison shows that most models, including KNN, Random Forest, and Logistic Regression, matched or exceeded the paper's reported accuracy. Random Forest achieved 93.93% vs. 92.52% in the paper. KNN improved, Logistic Regression was similar, while SVM slightly underperformed (87.88% vs. 90.43%). Naive Bayes and Decision Tree also showed strong results, though no paper comparison was available.



data_smote

	Unnamed: 0	Age	Sex	Steroid	Antivirals	Fatigue	Malaise	Anorexia	Liver Big	Liver Firm	...	Spiders	Ascites	Varices	Bilirubin	ALK Phosphate	SGOT	Albumin	Protime	Histology	Class
0	5	34	1	2.000000	2	2.0	2.000000	2.0	2.0	2.000000	...	2.000000	2.000000	2.000000	0.900000	0.271954	26.000000	4.000000	0.750000	1	2
1	10	39	1	1.000000	1	2.0	2.000000	2.0	1.0	1.000000	...	2.000000	2.000000	2.000000	1.300000	0.204724	30.000000	4.400000	0.850000	1	2
2	11	32	1	2.000000	1	1.0	2.000000	2.0	2.0	1.000000	...	1.000000	2.000000	2.000000	1.000000	0.129921	249.000000	3.700000	0.540000	1	2
3	12	41	1	2.000000	1	1.0	2.000000	2.0	2.0	1.000000	...	2.000000	2.000000	2.000000	0.900000	0.216535	60.000000	3.900000	0.520000	1	2
4	13	30	1	2.000000	2	1.0	2.000000	2.0	2.0	1.000000	...	2.000000	2.000000	2.000000	2.200000	0.122047	144.000000	4.900000	0.780000	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
127	115	42	1	1.522733	2	2.0	2.000000	2.0	2.0	1.522733	...	1.477267	2.000000	1.477267	1.236573	0.527108	35.408946	4.089093	0.774544	2	1
128	59	42	1	1.000000	1	1.0	1.000000	2.0	2.0	1.000000	...	1.573459	1.573459	1.573459	3.366853	0.735733	122.224920	3.329705	0.261521	1	1
129	32	39	1	1.025419	1	1.0	1.025419	2.0	2.0	1.000000	...	2.000000	1.974581	1.974581	2.212710	0.987491	97.415260	3.764413	0.397966	1	1
130	84	57	1	1.000000	2	1.0	1.000000	2.0	2.0	1.000000	...	1.000000	1.892109	1.892109	1.059042	0.325984	157.000000	3.502098	0.727448	2	1
131	106	47	1	2.000000	2	1.0	1.000000	2.0	2.0	1.000000	...	1.021429	1.000000	1.000000	1.022000	0.541102	29.685708	3.594205	0.614714	2	1

132 rows × 21 columns

Fig 20:SMOTE

The data_smote DataFrame contains 132 rows and 21 columns, showing medical data after SMOTE balancing. It includes patient features like age, symptoms, lab results, and a target class for classification.

5. CONCLUSION

In this project, we developed a machine learning model to help detect hepatitis more accurately and efficiently. By training and testing different algorithms, we found that some models like Random Forest and SVM gave good results in predicting the disease. This AI-powered system can support doctors by providing fast and reliable predictions, which is especially helpful in areas where medical experts or lab tests are not easily available. It can also be used in hospitals, clinics, or mobile health apps to assist with early diagnosis. Overall, this project shows that artificial intelligence can play an important role in improving liver care and helping

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

save lives through early detection of hepatitis.

6. FUTURE SCOPE

The future of healthcare, particularly in liver disease management, is being transformed through the integration of artificial intelligence and machine learning technologies. The project titled **"AI-Powered Liver Care: A Machine Learning Model for Hepatitis Diagnosis"** highlights the potential of data-driven solutions to revolutionize medical diagnostics. By leveraging advanced machine learning algorithms, future developments in liver care will enable faster, more accurate diagnosis of hepatitis, reducing the burden on healthcare professionals and enhancing patient outcomes.

As more comprehensive medical datasets become available, these AI models will continue to improve in precision and reliability. The integration of electronic health records (EHRs) with machine learning systems will allow for real-time data analysis and continuous patient monitoring, supporting proactive rather than reactive care. Additionally, the use of explainable AI (XAI) will improve transparency, helping medical practitioners interpret model predictions and build trust in AI-assisted decisions.

Furthermore, advancements in federated learning and secure data sharing can enable hospitals and research centers to collaboratively train models without compromising patient privacy, thus expanding diagnostic capabilities across regions. As a result, AI-powered systems will not only streamline the diagnostic process but also personalize treatment plans based on individual patient data.

Ultimately, the convergence of AI, healthcare data, and collaborative platforms will lead to smarter, more responsive liver care systems. This approach promises to improve early detection of hepatitis, enhance medical decision-making, and foster a more efficient, accessible, and equitable healthcare ecosystem for liver disease management.

REFERENCES

[1] Malik S, Das R, Thongtan T, Thompson K, Dbouk N. AI in Hepatology: Revolutionizing the Diagnosis and Management of Liver Disease. J Clin Med. 2024 Dec 22;13(24):7833.

[2] AI in Hepatology: Revolutionizing the Diagnosis and Management of Liver Disease. PMC.

[3] Artificial intelligence models for the diagnosis and management of liver diseases. PMC.

[4] How AI and Machine Learning Are Transforming Liver Disease Diagnosis and Treatment. News-Medical.net.

[5] A machine learning-based model analysis for serum markers of liver fibrosis in chronic hepatitis B. Nature.

[6] A machine learning model to predict liver-related outcomes after the functional cure of chronic hepatitis B. ScienceDirect.

[7] Development of a machine learning-based model to predict hepatic inflammation in chronic hepatitis B patients with concurrent hepatic steatosis: a cohort study. The Lancet.

[8] Role of artificial intelligence in the management of chronic hepatitis B. PMC.

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

[9] A New Tool for the Diagnosis and Management of Viral Hepatitis: Artificial Intelligence. Viral Hepatitis Journal.

[10] Artificial Intelligence Applications in Hepatology. CGH Journal.

[11] Artificial intelligence in imaging for liver disease diagnosis. Frontiers in Medicine.

[12] Machine learning models and AI in predicting diagnosis and prognosis of liver diseases. OAEPublish.

[13] EU health regulator clears use of AI tool in fatty liver disease trials. Reuters.

[14] AI-Driven Non-Invasive Detection and Staging of Steatosis in Fatty Liver Disease Using a Novel Cascade Model and Information Fusion Techniques. arXiv.

[15] Ultrasound Liver Fibrosis Diagnosis using Multi-indicator guided Deep Neural Networks. arXiv.

[16] Hybrid Adaptive Neuro-Fuzzy Inference System for Diagnosing Liver Disorders. arXiv.

[17] Adaptive Learning Expert System for Diagnosis and Management of Viral Hepatitis. arXiv.

[18] MELD-Plus: A generalizable prediction risk score in cirrhosis. Wikipedia.

[19] FibroTest: Assessing the degree of liver damage. Wikipedia.