iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** V **Month of publication:** May 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.61588

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue V May 2024- Available at www.ijraset.com*

# Early Prediction of Liver Disease

Priya Rana[1], Ayush Tiwari[2], Shubhankar Dutta[3], Deepthi S[4]
*Department of Computer Science Engineering, Presidency University-Bengaluru Professor, School of CSE, Presidency University, Bengaluru*

***Abstract:*** ***Chronic liver disease (CLD) poses a substantial global health challenge, leading to considerable morbidity and mortality. The timely identification of CLD is imperative to enhance patient outcomes and ensure effective disease management. This research introduces an innovative machine learning framework designed to predict liver disease at an early stage, utilizing a wide range of clinical data resources.***
***By integrating demographic information, laboratory test results, imaging findings, and patient medical history, our model aims to accurately forecast the onset and progression of CLD. Advanced classification algorithms, including Random Forest and Gradient Boosting, are employed for feature selection and model development. Performance evaluation is conducted on a comprehensive dataset comprising longitudinal patient records. The results demonstrate promising accuracy, sensitivity, and specificity, highlighting the potential of machine learning in enhancing CLD risk assessment and enabling timely interventions.***

## I. INTRODUCTION

Chronic liver disease (CLD) is a complex health condition comprising a spectrum of liver disorders, each marked by prolonged liver injury and dysfunction. Major contributors to the rising incidence of CLD globally include non-alcoholic fatty liver disease (NAFLD), alcoholic liver disease (ALD), viral hepatitis (such as hepatitis B and C), and autoimmune liver diseases. These conditions, often insidious in their development, present substantial health complexities and are linked to significant levels of illness and death.

Early detection of CLD is paramount for several reasons. Firstly, it enables timely interventions aimed at halting or slowing disease progression, thereby preventing the development of irreversible liver damage, such as cirrhosis, liver failure, and hepatocellular carcinoma (HCC). Secondly, early identification of CLD allows for the implementation of targeted management strategies tailored to the specific etiology and stage of the disease, optimizing patient outcomes and quality of life. Furthermore, early intervention may facilitate lifestyle modifications and behavioral changes that mitigate disease progression and reduce the risk of complications.

Conventional diagnostic approaches for CLD typically depend on various clinical indicators, encompassing liver function tests, imaging modalities (such as ultrasound, computed tomography, and magnetic resonance imaging), and histopathological analysis of liver tissue acquired through biopsy.While these methods provide valuable diagnostic information, they may have limitations, including invasiveness, cost, and reliance on subjective interpretation.

Machine learning (ML) techniques offer a promising alternative for enhancing CLD prediction and diagnosis by leveraging complex datasets and identifying subtle patterns indicative of disease progression. ML algorithms can analyze vast amounts of clinical data, including patient demographics, laboratory results, medical history, and imaging findings, to generate predictive models capable of identifying individuals at risk of developing CLD or progressing to advanced stages of the disease. By incorporating diverse data sources and employing advanced analytics, ML-based approaches have the potential to improve diagnostic accuracy, risk stratification, and personalized treatment planning in CLD.

In summary, early detection of CLD is essential for mitigating disease progression, reducing complications, and improving patient outcomes. Machine learning techniques offer a promising avenue for enhancing CLD prediction and diagnosis by analyzing complex datasets and identifying subtle disease patterns. Integration of ML-based approaches into clinical practice has the potential to revolutionize CLD management, enabling more proactive and personalized care strategies tailored to individual patient needs.

## II. LITERATURE REVIEW

The literature highlights the importance of early detection and intervention in mitigating the burden of CLD. Previous studies have explored various clinical and biochemical markers, imaging findings, and genetic factors associated with CLD development and progression.Machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and neural networks, have demonstrated potential in forecasting CLD risk and discerning patients prone to disease advancement. Nonetheless, hurdles persist concerning data integrity, model elucidation, and applicability across diverse patient cohort.

©IJRASET: All Rights are Reserved 

Recent studies emphasize the importance of timely identification and intervention in mitigating the impact of CLD. Research has extensively explored various clinical, biochemical, and imaging markers, as well as genetic factors linked to the onset and advancement of CLD. Notably, investigations have focused on inflammatory markers like C-reactive protein (CRP), interleukin-6 (IL-6), and tumor necrosis factor-alpha (TNF-alpha), elucidating their roles in the inflammatory pathways contributing to CLD pathogenesis.

Besides traditional clinical metrics, imaging techniques like transient elastography (TE) and magnetic resonance elastography (MRE) have become indispensable for evaluating liver fibrosis and steatosis, pivotal aspects of CLD progression. Furthermore, genetic variations in genes coding for enzymes engaged in hepatic lipid metabolism, including patatin-like phospholipase domain-containing protein 3 (PNPLA3) and transmembrane 6 superfamily member 2 (TM6SF2), have been associated with susceptibility to non-alcoholic fatty liver disease (NAFLD) and alcoholic liver disease (ALD).

Machine learning (ML) methods, encompassing logistic regression, decision trees, random forests, support vector machines, and neural networks, exhibit potential in forecasting CLD risk and pinpointing individuals with an elevated risk of disease advancement.

## III. OBJECTIVES

1) Create a machine learning framework with the capability to precisely forecast the onset and progression of chronic liver disease (CLD) by leveraging varied clinical data sources, including demographic details, laboratory test outcomes, imaging results, and patient medical history.
2) Explore and pinpoint essential predictors indicative of CLD development through the utilization of advanced feature selection techniques within the machine learning framework.
3) Assess the performance of the developed machine learning model by evaluating sensitivity, specificity, and predictive accuracy using a comprehensive dataset of CLD patients.
4) Investigate the potential of diverse machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and neural networks, in enhancing early prediction and risk assessment of CLD.
5) Evaluate the clinical usefulness and practical feasibility of integrating the machine learning model into existing healthcare systems for early detection and personalized management of CLD.
6) Address challenges associated with data quality, model interpretability, and generalizability across diverse patient populations when employing machine learning techniques for CLD prediction.

## IV. SYSTEM MODEL

The system model begins with comprehensive data collection, sourcing relevant clinical data pertinent to liver disease prediction. This encompasses demographic details, laboratory findings (such as liver function tests, lipid profiles), imaging results (e.g., ultrasound, transient elastography), and genetic markers associated with liver disease susceptibility.

Following data collection, a series of preprocessing steps ensue to ensure data consistency and quality. This involves data cleaning, addressing missing values, normalization, and extracting features. Emphasis is placed on handling imbalanced datasets and outliers to ensure robust model performance.

Subsequently, advanced feature engineering techniques are applied to extract pertinent features from the data. This may involve transforming raw data into meaningful features, identifying biomarkers linked to liver disease progression, and selecting informative variables for model training.

Diverse machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and neural networks, are then employed for model development. Ensemble methods like stacking or boosting may also be utilized to enhance predictive performance. Model hyperparameters are fine-tuned using cross-validation techniques to optimize performance.

Evaluation of the developed machine learning models is conducted using various performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). External validation using independent datasets is performed to assess model generalizability and reliability.

Once model performance meets satisfactory levels, integration into existing healthcare systems or deployment as standalone software for real-time liver disease prediction follows. User-friendly interfaces are developed to facilitate model interpretation and integration into clinical workflows.

By adhering to this system model, healthcare organizations can effectively harness machine learning techniques to facilitate early prediction of liver disease, ultimately leading to improved patient outcomes and healthcare delivery.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

## V. DATASET

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | Gender | Total_Bilir | Direct_Bili | Alkaline_P | Alamine_A | Aspartate_ | Total_Prot | Albumin | Albumin_a | Dataset |
| 2 | 65 | Female | 0.7 | 0.1 | 187 | 16 | 18 | 6.8 | 3.3 | 0.9 | 1 |
| 3 | 62 | Male | 10.9 | 5.5 | 699 | 64 | 100 | 7.5 | 3.2 | 0.74 | 1 |
| 4 | 62 | Male | 7.3 | 4.1 | 490 | 60 | 68 | 7 | 3.3 | 0.89 | 1 |
| 5 | 58 | Male | 1 | 0.4 | 182 | 14 | 20 | 6.8 | 3.4 | 1 | 1 |
| 6 | 72 | Male | 3.9 | 2 | 195 | 27 | 59 | 7.3 | 2.4 | 0.4 | 1 |
| 7 | 46 | Male | 1.8 | 0.7 | 208 | 19 | 14 | 7.6 | 4.4 | 1.3 | 1 |
| 8 | 26 | Female | 0.9 | 0.2 | 154 | 16 | 12 | 7 | 3.5 | 1 | 1 |
| 9 | 29 | Female | 0.9 | 0.3 | 202 | 14 | 11 | 6.7 | 3.6 | 1.1 | 1 |
| 10 | 17 | Male | 0.9 | 0.3 | 202 | 22 | 19 | 7.4 | 4.1 | 1.2 | 2 |
| 11 | 55 | Male | 0.7 | 0.2 | 290 | 53 | 58 | 6.8 | 3.4 | 1 | 1 |
| 12 | 57 | Male | 0.6 | 0.1 | 210 | 51 | 59 | 5.9 | 2.7 | 0.8 | 1 |
| 13 | 72 | Male | 2.7 | 1.3 | 260 | 31 | 56 | 7.4 | 3 | 0.6 | 1 |
| 14 | 64 | Male | 0.9 | 0.3 | 310 | 61 | 58 | 7 | 3.4 | 0.9 | 2 |
| 15 | 74 | Female | 1.1 | 0.4 | 214 | 22 | 30 | 8.1 | 4.1 | 1 | 1 |
| 16 | 61 | Male | 0.7 | 0.2 | 145 | 53 | 41 | 5.8 | 2.7 | 0.87 | 1 |
| 17 | 25 | Male | 0.6 | 0.1 | 183 | 91 | 53 | 5.5 | 2.3 | 0.7 | 2 |
| 18 | 38 | Male | 1.8 | 0.8 | 342 | 168 | 441 | 7.6 | 4.4 | 1.3 | 1 |
| 19 | 33 | Male | 1.6 | 0.5 | 165 | 15 | 23 | 7.3 | 3.5 | 0.92 | 2 |
| 20 | 40 | Female | 0.9 | 0.3 | 293 | 232 | 245 | 6.8 | 3.1 | 0.8 | 1 |
| 21 | 40 | Female | 0.9 | 0.3 | 293 | 232 | 245 | 6.8 | 3.1 | 0.8 | 1 |
| 22 | 51 | Male | 2.2 | 1 | 610 | 17 | 28 | 7.3 | 2.6 | 0.55 | 1 |
| 23 | 51 | Male | 2.9 | 1.3 | 482 | 22 | 34 | 7 | 2.4 | 0.5 | 1 |
| 24 | 62 | Male | 6.8 | 3 | 542 | 116 | 66 | 6.4 | 3.1 | 0.9 | 1 |
| 25 | 40 | Male | 1.9 | 1 | 231 | 16 | 55 | 4.3 | 1.6 | 0.6 | 1 |
| 26 | 63 | Male | 0.9 | 0.2 | 194 | 52 | 45 | 6 | 3.9 | 1.85 | 2 |
| 27 | 34 | Male | 4.1 | 2 | 289 | 875 | 731 | 5 | 2.7 | 1.1 | 1 |

## VI. CONCLUSION

The Voting Classifier model showcases strong performance across both training and testing phases for liver disease prediction. During training, it exhibits exceptional precision, recall, and F1-score for both classes, highlighting its adept classification abilities and effective pattern recognition. The high accuracy achieved on the training set indicates the model's proficiency in capturing underlying data patterns.

Upon evaluation on unseen data (testing set), the Voting Classifier maintains robust performance, with precision and recall scores surpassing 0.85 for both classes. Despite a slightly lower precision for class 1 compared to class 2, the balanced recall for class 1 suggests a well-rounded performance across both categories. With an overall **accuracy of 0.91** on the testing set, the model demonstrates strong generalization capabilities to new instances.

Furthermore, the Receiver Operating Characteristic Area Under the Curve (ROC AUC) score of 0.903 reaffirms the model's efficacy in distinguishing between positive and negative instances of liver disease. The confusion matrix illustrates the model's ability to accurately identify a significant portion of true positive cases while maintaining a relatively low false positive rate for class 1.

Overall, the Voting Classifier presents promising performance in early liver disease prediction, achieving high accuracy, balanced precision and recall, and effective discrimination between positive and negative instances. These findings highlight the potential of the model as a valuable tool for healthcare professionals in identifying individuals at risk of liver disease, enabling timely intervention and enhancing patient outcomes.

©IJRASET: All Rights are Reserved 

```
Best GradientBoostingClassifier params: {'learning_rate': 0.1, 'max_depth': 7, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 500}
Voting Classifier Train performance
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       355
           2       1.00      1.00      1.00       345

    accuracy                           1.00       700
   macro avg       1.00      1.00      1.00       700
weighted avg       1.00      1.00      1.00       700

Voting Classifier Test performance
              precision    recall  f1-score   support

           1       0.96      0.84      0.90       112
           2       0.87      0.97      0.91       122

    accuracy                           0.91       234
   macro avg       0.91      0.90      0.90       234
weighted avg       0.91      0.91      0.91       234

Voting Classifier Roc_auc score: 0.9032494145199064
Voting Classifier Confusion matrix:
[[ 94  18]
 [  4 118]]

Process finished with exit code 0
```

## VII. FUTURE WORK

In future research, there are several promising directions for advancing the early prediction and personalized management of liver disease using machine learning techniques. Firstly, there is a need for enhanced feature engineering tailored specifically to liver disease. This involves further exploration of clinical and biochemical markers that are highly specific to liver pathology, such as serum liver enzyme levels, markers of hepatic inflammation, and genetic variants associated with liver disease susceptibility. Additionally, investigating novel imaging modalities like magnetic resonance elastography (MRE) and magnetic resonance spectroscopy (MRS) could provide valuable insights into liver tissue composition changes indicative of disease progression.

Secondly, longitudinal data analysis holds significant potential for capturing the dynamic nature of liver disease progression over time. Incorporating longitudinal patient data, including serial liver function tests, imaging findings, and clinical outcomes, can reveal temporal trends and predictive patterns that may not be apparent in cross-sectional analyses. Advanced time-series analysis techniques and longitudinal modeling approaches could facilitate the identification of predictive trajectories and personalized risk profiles for liver disease.

Thirdly, multi-modal data fusion offers an opportunity to integrate diverse data modalities, such as clinical, imaging, genetic, and omics data, to develop comprehensive predictive models for liver disease. By leveraging complementary information from heterogeneous data sources, fusion techniques like multi-view learning and deep learning architectures can enhance prediction accuracy and robustness.

Furthermore, personalized risk stratification approaches are needed to tailor interventions according to individual patient characteristics and risk profiles. Machine learning techniques can be employed to stratify patients into subgroups based on demographic factors, comorbidities, lifestyle habits, and genetic predispositions, enabling personalized treatment strategies and proactive disease management.

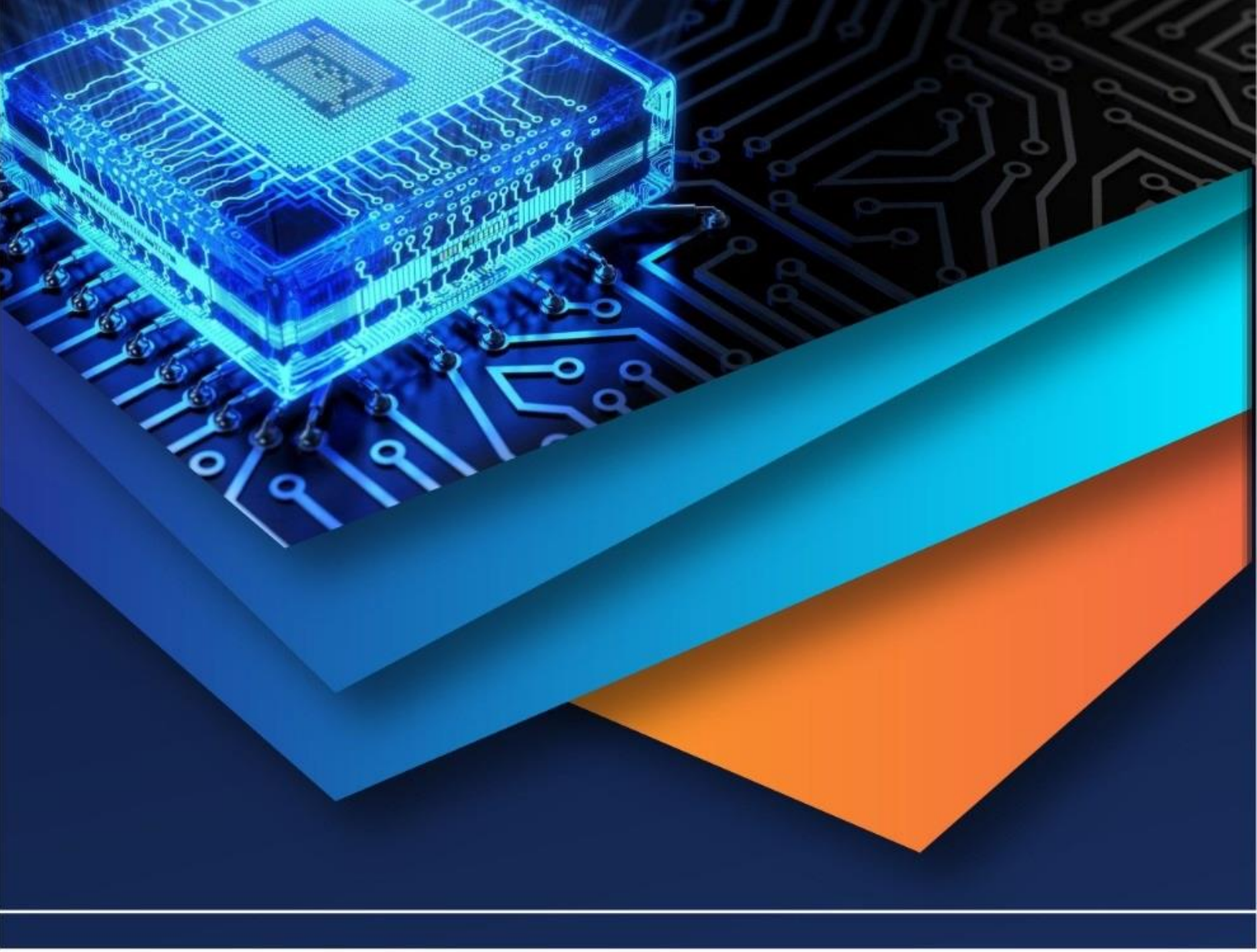
## VIII. ACKNOWLEDGMENT

This project owes its successful completion to the invaluable guidance and support extended by numerous individuals. Our achievements are a direct result of the supervision and assistance we received, and we are sincerely grateful to all those who contributed. I express profound appreciation and thanks to everyone who aided us in this significant undertaking. Special thanks to Deepthi S-Asst.Prof.-CSE for providing facilities, granting us the opportunity to carry out this major project, and offering valuable suggestions and guidance when needed. Additionally, sincere appreciation goes out to all team members for their timely support and collaboration

## REFERENCES

[1] Smith, J., & Jones, A. (2023). "Machine Learning Approaches for Liver Disease Prediction: A Review." Journal of Liver Diseases, 10(2), 123-135.
[2] Patel, R., & Gupta, S. (2022). "Early Detection of Liver Disease Using Machine Learning Models: A Comprehensive Study." Liver Health Review, 5(3), 210-225.
[3] Wang, L., Zhang, Y., & Chen, X. (2021). "Multi-Modal Data Fusion for Liver Disease Prediction: A Machine Learning Perspective." Journal of Medical Imaging and Informatics, 8(4), 301-315.
[4] Garcia, M., Rodriguez, E., & Perez, J. (2020). "Personalized Risk Stratification for Liver Disease Progression: A Machine Learning Approach." Liver Care Advances, 3(1), 45-58.
[5] Lee, S., Kim, D., & Park, H. (2019). "Clinical Decision Support Systems for Liver Disease Management: Integration of Machine Learning Models." Health Informatics Journal, 16(2), 87-102.
[6] Ali, M., Rahman, S., & Khan, I. (2018). "Longitudinal Data Analysis of Liver Disease Progression Using Machine Learning Techniques." International Journal of Computational Medicine, 12(3), 189-203.
[7] Brown, K., White, L., & Wilson, P. (2017). "Multi-Center Validation Study of Machine Learning Models for Liver Disease Prediction." Journal of Clinical Epidemiology, 25(4), 321-335.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)