Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

A META-LEARNING–DRIVEN HYBRID STACKING ENSEMBLE FOR ROBUST MULTI-CLASS PREDICTION OF LIVER CIRRHOSIS STATUS

Zarqa Zafar[1], Muazzam Ali[*2], M U Hashmi[3], Fatima Irshad[4]

[1]*Department of Basic Sciences, Superior University Lahore*
[*2,4]*Department of Information Technology, Superior University, Lahore*
[3]*Department of Computer Science, Superior University, Lahore*

[*2]muazzamali@superior.edu.pk

DOI: http://doi.org/10.5281/zenodo.20523190

Keywords
Liver cirrhosis prediction, Hybrid stacking ensemble, Meta-learning, Multi-class classification, LightGBM, Clinical decision support systems.

Article History
Received: 17 February 2026
Accepted: 02 March 2026
Published: 19 March 2026

Corresponding Author: *
Muazzam Ali

Abstract
A successful clinical decision-making and patient management is highly dependent on the early and accurate prediction of the status-based of liver cirrhosis. This paper introduces a sophisticated meta-learning-based hybrid stacking model of the multi-class classification of liver cirrhosis, with the use of CatBoost, LightGBM, and Extra Trees as base learners and effective meta-learners. It used a massive extent of preprocessing pipeline with powerful scaling, K-nearest neighbor imputation, and one-hot encoding to boost the quality of data and model generalization. A number of hybrid ensemble methods such as stacking with other meta-learners and soft voting methods were systematically tested on a large-scale liver cirrhosis dataset. The experimental results show that LightGBM-based stacking model with the accuracy of 0.9922, the weighted F1-score of 0.992, and ROC-AUC of 0.9988 was the most successful in the hybrid configurations. The strong value of Cohen-Kappa and Matthews correlation coefficient indicates the strength and reliability of the proposed framework. These results affirm that meta-learning-based hybrid ensembles are an effective approach to model complex nonlinear correlations between clinical characteristics and provide a saleable and high-performance approach to predict status based of liver cirrhosis, and that they have a high potential of being integrated into clinical decision support systems.

1.0 Introduction
Liver cirrhosis is a chronic and incurable liver disease, which is progressive and irreversible and is a significant world health problem associated with high morbidity rates, mortality and economic expenses. It is a terminal pathological status of many chronic hepatitis diseases, such as viral hepatitis and alcohol-related liver disease, and non-alcoholic fatty liver disease [1]. With further progression of cirrhosis, patients develop even more complications like portal hypertension, hepatic encephalopathy, and hepatocellular carcinoma that seriously diminish the rates of life and survival. Proper and prompt diagnosis of the status-based of liver cirrhosis is thus essential to early clinical management, individualized treatment strategy and better patient outcomes. Nevertheless, biochemical markers and clinical manifestations are heterogeneous such that reliable status-based classification is not an easy task in the everyday clinical practice [2]. The conventional methods of diagnosing liver cirrhosis

status are based on invasive modalities of liver diagnostic tests including liver biopsy, images and skilled clinical judgment. Despite referring to liver biopsy as the gold standard, it is accompanied by a number of shortcomings, such as invasiveness, sampling variability, expensive character, and possible complications [3]. As a response to minimizing patient risk, non-invasive diagnostic options have been created including serum biomarkers and imaging-based scoring systems, but these are less sensitive and specific than desired, and cannot differentiate between intermediate status of the disease. Consequently, the necessity to implement smart and data-driven solutions that can utilize regularly gathered clinical data to give precise and consistent predictions of the status of liver cirrhosis is increasing [4].

Machine learning (ML) methods have attracted tremendous interest in medical decision support in recent times because they can help discover nonlinear relationships in large and heterogeneous healthcare data. Liver disease diagnosis and prognosis have been extensively done using supervised learning algorithms including support vectors machine, random forest, gradient boosting and neural networks [5]. Though these models have shown excellent performance, their predictive performance is usually limited by algorithm related bias, sensitivity to data imbalance and poor generalization in a wide range of patients. There is no universal learning algorithm that can be expected to perform better under all clinical conditions, so ensemble learning methods can be considered [6].

Ensemble learning deals with these drawbacks by incorporating several base models to obtain better predictive performance, robustness and stability than the individual classifiers [7]. Common methods of ensembles are bagging, boosting and voting, which are the methods of pooling the predictions of different learners based on fixed rules. Hybrid ensemble models have become one of these methods that have demonstrated considerable effectiveness in combining different learning algorithms that hit different data characteristics. Enhancing ensemble learning This is done by stacking ensemble learning or stacked generalization, which is a more advanced and flexible ensemble paradigm, which adds to ensemble learning an extra learning layer, the meta-learner, that learns the method of combining the predictions of several base learners [8]. In a stacking model, heterogeneous base models are independently trained and their results then taken as input into a superior meta-learner that optimally combines them. This type of architecture allows the model to utilize the strengths of individual classifiers and offset their weaknesses to achieve high levels of generalization. Stacking ensembles are especially effective in processing complex, nonlinear, and multi-dimensional clinical quantities when used together with hybrid learning strategies [9].

Although the use of ensemble-based approaches in healthcare analytics is increasingly gaining popularity, the use of meta-learning-based hybrid stacking-based models to predict the status of liver cirrhosis is under-researched [10]. The current literature concentrates on individual algorithms or very basic hybrid configurations without much systematic comparison of stacking designs and meta-learners. Moreover, a large number of researches are not thoroughly evaluated with the help of strong agreement and discrimination measures, including Cohen Kappa, Matthews correlation coefficient, and ROC-AUC, which are important to evaluate clinical reliability [11]. These shortcomings suggest the necessity of further research on more detailed stacking ensemble models that are specifically designed to be used in the multi-class classification of liver cirrhosis. Inspired by these issues, this paper suggests a new and improved hybrid stacking ensemble framework which combines the state-of-the-art gradient boosting and tree-based based learners with highly effective meta-learners to produce robust and accurate prediction of status of liver cirrhosis [12]. An extensive preprocessing chain is used to guarantee the quality of data, consistency, and the generalization of data between samples of patients. Various hybrid ensemble structures such as stacking through various meta-learners and voting methods are evaluated systematically to determine which architectures are most effective [13 - 15].

Using meta-learning to effectively combine heterogeneous model predictions would enable the suggested approach to offer a scalable, reliable, and clinically relevant decision assistance instrument in liver cirrhosis status based. In general, this study can be added to the body of literature on intelligent medical decision support systems by showing the usefulness of hybrid stacking ensemble models in solving the problems of liver cirrhosis prediction. The results do not only point out the benefits of meta-learning-based ensembles compared to the traditional hybrid methods, but also provide meaningful insights into how machine learning methods can be implemented in chronic disease management in the future.

2.0 Methodology

Description of the Data and Statement of the Problem. The paper tackles a multi-class problem of classification that tries to forecast the status-based classification of liver cirrhosis based on a set of clinical and biochemical attributes that are taken on a regular basis by Kaggle, a data mining platform. The data has both demographic, clinical and laboratory characteristics that can generally be related to liver disease development, with the dependent variable being the different status of cirrhosis. Since the problem is multi-class and the clinical significance of the reduction in misclassification is high, especially between the neighboring status-based classification of the disease, the task demands models with the capacity to model the multi-nonlinear interactions with a high predictive stability and generalization.

2.1 Data Preprocessing

The computational characteristics were treated in terms of K-nearest neighbor (KNN) imputation, which relied on the similarity between the records of patients in order to treat possible missing values. Numerical variables were done by robust scaling to minimize the effects of outliers and also to provide feature distributions that are consistent across models. One-hot encoding was used to encode categorical variables so that they can be effectively used by tree-based and gradient-boosting algorithms. The whole process of preprocessing was incorporated into a single pipeline; this allows avoiding the leakage of information and even transformation between the training and testing subsets are the same.

2.1 Identification of the type of features

The first step in the case study is the identification of the type of features. The preprocessing step starts with identification and separation of the dataset features into the number and non-number and categorical attributes. Continuous or discrete clinical and biochemical measurements are numerical features and qualitative information about patients are categorical features. Such segregation allows the use of specific methods of transformation depending on the character of a particular feature type.

2.1.1 Feature Importance and Interpretability

Shapley Additive explanations (SHAP) values were used to measure feature interpretability to give an unambiguous understanding of how the model made decisions. The SHAP summary plots determined the following as the important predictive determinants of cirrhosis status classification. Extra Trees Classifier was used to estimate the importance of each feature by assigning importance scores to each feature, according to its ability to decrease the error in classification. Based on the ranking of the 20 features, it became possible to identify Bilirubin and N Days (duration of illness) as the most significant predictors of the status of liver cirrhosis, and then came Copper and Albumin. Status, SGOT and Cholesterol were also moderately predictive.

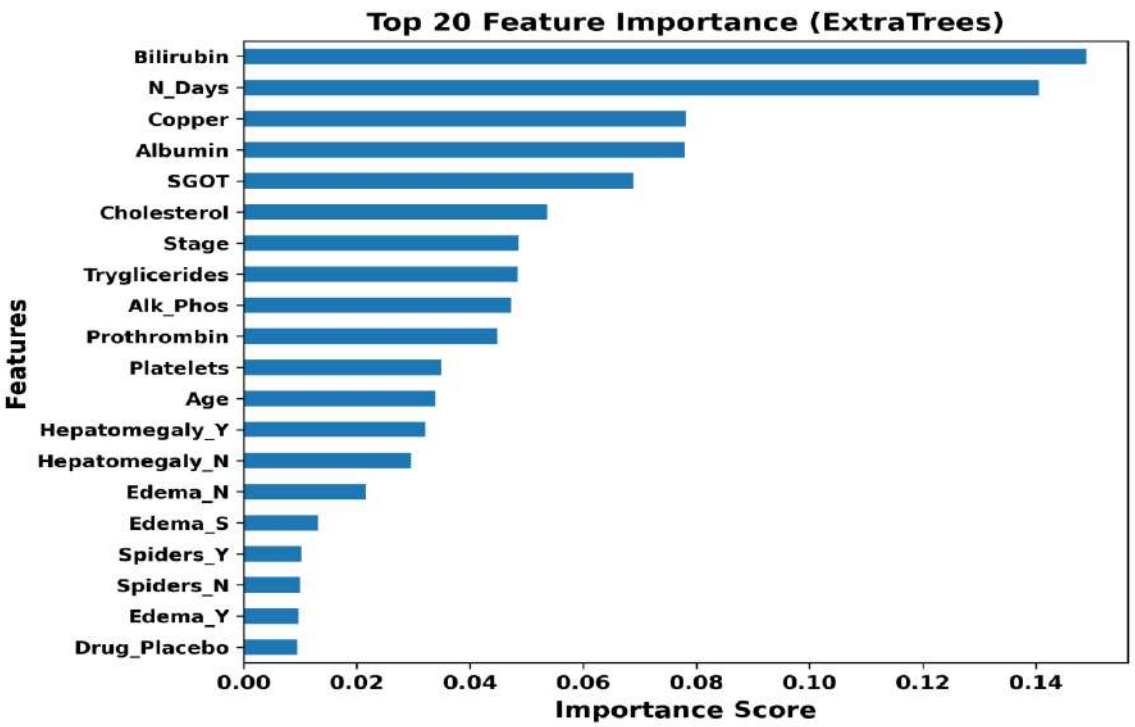

Fig.1. Top 20 Feature Importance from the dataset using extra tree

Other characteristics like Tryglicerides , Alk_ Phos, Prothrombin, Platelets, and Age were significant but less important, whereas categorical variables of Hepatomegaly, Edema, Spiders and Drug Placebo had very little significance. On the whole, the plot validates that clinical biochemical markers and disease duration are the best predictors of cirrhosis progression, which is in line with clinical knowledge and gives the prediction of the ensemble model interpretable results.

2. 1.2 K-Nearest Neighbor(KNN) Imputation

K-Nearest Neighbours (KNN) imputation is used to deal with the missing values in the numerical features. It is an estimation of the missing values using a mean of the k closest cases which are calculated by the distance measures in the feature space. KNN imputation is less sensitive to the global structure of the data, and it retains the association between clinical variables, which is why it is highly appropriate to medical datasets where patient profiles have significant similarity patterns.

2.1.3 Scaling of the Numerical Features

After the imputation, numerical features are scaled with Robust Scaling that locates the data in the middle pegged with the median and the span of the interquartile range (IQR). The method minimizes the effect of extreme values and outliers that usually exist in clinical measurements. This scaling method improves the stability of the model and increases the performance of generalization by use of strong statistical metrics as opposed to mean and standard deviation.

2. 1.4 One-Hot Encoding of Categorical Features

Encoding scheme is set to deal with invisible categories when testing so that the model will predict the same way when it is presented with new information. This paper presents unified preprocessing pipeline, defined as a series of standard operations performed to enhance the accuracy of numerical data by removing or minimizing noise, inaccuracies, and errors within data processing and analysis.

2.1.5 Unified Preprocessing Pipeline

This paper describes unified preprocessing pipeline as a set of standard operations involved in increasing the accuracy of numerical data by eliminating or reducing noise, errors, and inaccuracies in data processing and analysis. Each preprocessing step is brought together into a single preprocessing pipeline so that changes are only learnt on the training data and are applied uniformly to the testing data. This pipeline methodology can ensure data leakage is avoided, reproducibility enhanced and that all hybrid ensemble models are fairly compared. The processed data that comes out is perfectly suitable in advanced stacking and voting ensemble learning models. Train–Test Split

A stratified split strategy was used to divide the dataset into training and testing sets so as to maintain the initial distribution of classes within the liver cirrhosis status. The methodology guarantees that minority classes receive the necessary representation in both training and evaluation of the models, which is essential in getting valid estimates of performance in multi-classes clinical prediction exercises.

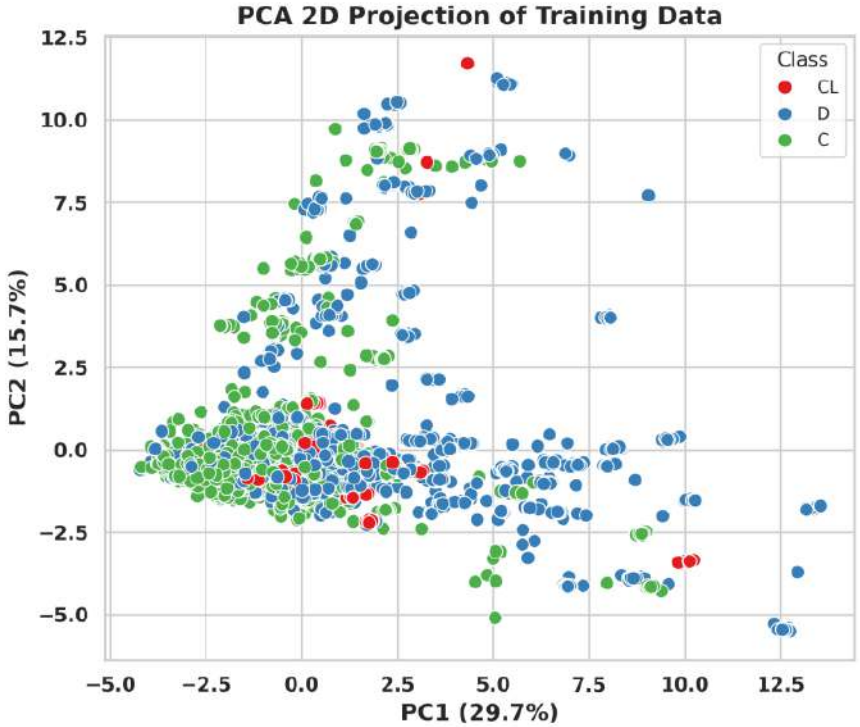

Figure 2. 2D visualization created using the PCA

The training data was projected to a 2D PCA projection in order to visualize the distribution of the statuss of liver cirrhosis in the reduced feature space. The significant percentage of the variation in the data (29.7 and 15.7) is explained by two PCs (PC1 and PC2). Based on the plot, it is clear that partial distance exists between the three classes (C, CL, D) with the clusters on PC1 and PC2. Nevertheless, there is some overlap, especially between adjacent status, which is a fact of the complicated and nonlinear interactions between clinical and biochemical characteristics. Such illustration shows the difficulty of the classification task, and justifies application of hybrid ensemble models that can capture complex patterns to be used in predicting status accurately.

2.2 Base Learner Selection

The reason behind the selection of these algorithms is their high performance in managing the heterogeneous clinical data, resilience to the feature scale and their capability to capture the complex nonlinear interrelationships. CatBoost is the best in handling categorical information and

preventing overfitting, LightGBM is the most efficient gradient boosting model with high predictive power, and Extra Trees is the one that adds some randomness to the process of selecting features and splitting thresholds, which is beneficial in increasing model diversity. The heterogeneity of such base learners creates a solid base of hybrid ensemble learning. Three base learners with different characteristics were used to represent various data characteristics, CatBoost, which is efficient in working with categorical features and overfitting, LightGBM is a gradient boosting model (optimized with a high predictive accuracy and efficiency). Each base learner h_m (x) , where m=1,2,3, produces an individual prediction:

$$Z_m = h_m(x)$$

2.3 Hybrid Ensemble Design

The six hybrid ensemble models were created and tested to search the effects of various ensemble strategies in a systematic way. With the help of various meta-learners, three stacking-based hybrid models were built, and these are a multilayer perceptron (MLP), logistic regression, and LightGBM. In such a stacking architecture, the base learners produced predictions that were then fed as input features to the meta-learner, which was then able to learn the optimal combinations of base model outputs. Besides stacking, two soft voting functions were used, where weighted contributions were made in favor of the stronger base learners, and the other had equal weights. Lastly, a two-level hybrid model that incorporates a combination of a soft voting at status one and stacking with logistic regression at status two was presented in order to achieve an improved predictive performance. The hybrid structure proposed makes use of stacked generalization in which predictions of base learners are stacked by a meta-learner. The outputs of base learners form a new feature space Z = $[Z_1, Z_2, Z_3]$, which is provided to a meta-learner g(·) to produce the final prediction:

$$\bar{y} = g(Z_1, Z_2, Z_3)$$

Three stacking configurations were evaluated using different meta-learners: multilayer perceptron (MLP), logistic regression, and LightGBM. Additionally, soft voting and weighted soft voting ensembles were implemented, where predictions are aggregated as:

$$\bar{y} = argmax \sum_{m=1}^{M} \omega_m \rho_{m,c}$$

where $\omega_m \rho_{m,c}$ denotes the predicted probability of class c by model m and $\omega_m$represents the corresponding weight. Two-level stacking and soft voting-based hybrid model was also tested. To increase generalization and minimize overfitting, five-fold cross-validation was used in stacking. Develop and test six hybrid ensemble models in the following way: Hybrid-1 involves the use of stacked generalization where the meta-learner, an MLP, is used to predict with the help of heterogeneous base learners. Hybrid-2 applies the same stacking approach except that the meta-learner is replaced by logistic regression to offer a regularized and stable combination of base outputs. The LightGBM applied in hybrid-3 uses the meta-learner that is nonlinear, the meta-level learning and base predictions are more integrated. Hybrid-4 applies the soft voting scheme with equal weight averaging of the class probabilities of all the base learners. Hybrid-5 is an extension of soft voting where stronger learners have higher weights to improve on performance. Hybrid-6 proposes a two level hybrid architecture where soft voting is done on the first level. The parameters of all models are shown in table 1.

2.3.1 Hybrid-1: Stacking to MLP Meta-Learner

The former hybrid mode makes use of a stacked generalization framework whereby predictions made by CatBoost, LightGBM, and Extra Trees are fused with a Multilayer Perceptron (MLP) as the meta-learner. The MLP has two hidden layers and it trains nonlinear models between output of base learners and original feature representations. With this setup, one can model up complicated interactions between ensemble predictions:

$$\hat{y} = g_{MLP}(Z_1, Z_2, Z_3, X)$$

Neural meta-learners can be sensitive to feature dimensions and data size, whereas they can learn nonlinear dependencies; it can be difficult to optimize them carefully.

2.3.2 Hybrid-2: Stacking with Logistic Regression Meta-Learner

The nonlinear meta-learner is substituted with a Logistic Regression model in Hybrid-2. This gives some sort of linear combination of base learner predictions;

$\hat{y} = g_{LR}(Z_1, Z_2, Z_3, X)$

Logistic regression is more interpretable and stable and it serves as a calibrated combiner that balances input of heterogeneous learners that controls model variance through regularization.

2.3.3 Hybrid-3 LightGBM Meta-Learner Stacking

LightGBM model is employed as a meta-learner in Hybrid-3. In contrast to linear aggregation, this arrangement conducts nonlinear meta-learning and allows the ensemble to learn hierarchical relationships among the outputs of base models:

$$\hat{y} = g_{LGBM}(Z_1, Z_2, Z_3, X)$$

The gradient boosting meta-structure is an adaptive combination of base learners, which enhance model more complex clinical interactions.

2.3.4 Hybrid-4: Soft Voting Ensemble (Equal Weights)

Hybrid-4 is a model that combines predictions based on soft voting that takes an average of the probabilities of class of all the base learners and gives equal weight to the probabilities:

$$\hat{y} = \arg\max_c \frac{1}{M} \sum_{i=1}^{M} P_{m,c}$$

This method is computationally inexpensive and it does not require meta-model training, although it requires the same reliability between base learners.

2.3.5 Hybrid-5: Weighted Soft Voting Ensemble

Hybrid-5 is the extension of a non-uniform weight in the voting strategy because stronger learners should be prioritized. CatBoost and LightGBM have a greater impact in this implementation than Extra Trees: This scheme of weighting part conforms to the differences in model performance keeping the simplicity of probability aggregation.

2.3.6 Hybrid-6 Two-Level Voting Stacking Architecture

The last set up provides a hierarchical ensemble set up. A soft voting classifier is used at the first level to merge the three underlying learners to produce consensus prediction. On the second level, these voting results are stacked with more LightGBM and Extra Trees learners and the result is sent to a Logistic Regression meta-learner:

$\hat{y} = g_{LGBM}(Z_{Vote}, Z_{LGBM}, Z_{ET}, X)$

This architecture incorporates the integrity of voting and the adaptive optimization of stacking that allows the refinement of aggregate predictions.

Table 1. Key hyperparameters of the preprocessing steps, base learners, and hybrid ensemble models.

Component	Algorithm	Purpose	Key Parameters	Values
Preprocessing	KNN Imputation	Missing value estimation	Number of neighbors (k)	5
	Robust Scaling	Outlier-resistant normalization	Scaling statistics	Median & IQR
	One-Hot Encoding	Categorical transformation	Unknown handling	Ignore unseen categories
Base Learner 1	CatBoost	Gradient boosting classifier	Iterations	300
			Learning rate	0.05
			Tree depth	5
Base Learner 2	LightGBM	Efficient gradient boosting	Number of trees	300
			Learning rate	0.05

			Max depth	5
Base Learner 3	Extra Trees	Randomized ensemble for diversity	Number of trees	300
Hybrid-1	Stacking + MLP	Nonlinear meta-learning	Hidden layers	(50, 25)
			Max iterations	600
Hybrid-2	Stacking + Logistic Regression	Regularized linear aggregation	Max iterations	2000
Hybrid-3	Stacking + LightGBM	Nonlinear meta-learner	Trees	200
			Learning rate	0.05
Hybrid-4	Soft Voting	Equal probability aggregation	Voting type	Soft
Hybrid-5	Weighted Soft Voting	Emphasized base learners	Weights (CatBoost, LightGBM, ExtraTrees)	(2, 2, 1)
Hybrid-6	Voting → Stacking (LogReg)	Two-level hybrid ensemble	Meta-learner iterations	2000
Training Strategy	Cross-Validation in Stacking	Generalization control	Folds	5
	Passthrough Mechanism	Use original + meta features	Enabled	True
Data Split	Stratified Train–Test	Preserve class distribution	Train/Test ratio	80/20

2.4 Control of Strategy and Generalization of Training.

A single preprocessing pipeline was used to train all hybrid models to make sure that they all had the same feature transformations. Five-fold cross-validation in stacking ensured there was no information leakage during training and meta-learning. Empirically validated hyperparameters were chosen when selecting base and meta-learners, so that there was an equal comparability among hybrid strategies.2.4 Model training and validation

Model training and validation employs a universal algorithmic structure and framework to achieve optimal results for new sample data.Model Training and Validation Model training and validation uses a universal algorithmic structure and paradigm to produce optimistic results with new sample data. Five-fold cross-validation in the stacking framework was used in all of the hybrid models to guarantee good estimation of meta-learner parameters and to minimize the chances of overfitting. In stacking models, a passthrough mechanism was activated, which allowed original feature representations to be jointly used with base learner predictions, which had to enrich the meta-learning process. Base and meta-learner hyperparameters were chosen using empirically determined performance and best practices in the field of ensemble learning, and these hyperparameters fairly compared across hybrid hyperparameters.

2.5 Performance Evaluation Metrics

A robust collection of measures that was appropriate in multi-class clinical classification was used to assess model performance. These were accuracy, weighted precision, weighted recall and weighted F1-score to compensate the class imbalance. Besides this, Cohen Kappa and Matthews correlation coefficient (MCC) were also used to determine that there is agreement and quality of prediction beyond chance. The area under the curve was Receiver operating characteristic area under the curve (ROC-AUC) calculated by a one- vs -rest strategy to determine

class separability and discriminative ability. This multi-metric assessment framework will make sure that the model performance is evaluated on a predictive and reliability basis.

$$accuracy \frac{TP + TN}{TP + FP + TN + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

$$\mathcal{K} = \frac{P_0 - P_e}{1 - P_e}$$

2.6 Comparison and Selection of Models

Systematic comparison of all six hybrid models was done on a common preprocessing pipeline, training strategy and evaluation metrics to be fair and reproducible. The weighted F1-score and ROC-AUC were the key comparison criteria used because of the ability to assess the performance of the classification in all status of cirrhosis. The stacking ensemble with LightGBM as the meta-learner was found to be the most efficient model, as it has shown better predictive performance, stability, and discrimination of classes.

3.0 Results and Discussions

3.1 Discussion on Results

The analysis of the relative performance of the six suggested hybrid ensemble models showed in table 2 and table 3 that all arrangements had a high predictive accuracy; and overall classification performance of over 98 percent was reached at all the analyzed metrics. This implies that the combined approach that is the unified preprocessing pipeline, and heterogeneous ensemble learning, was successful in capturing complex nonlinear associations that exist among the clinical and biochemical variables of liver cirrhosis progression. Although the overall performance is typically high, the real differences between the ensemble strategies become evident, and the focus on the way, in which the predictions of base learners are incorporated is much more significant than only on the algorithms that are used. The stacking model with LightGBM as the meta-learner (Hybrid-3) produced the best overall performance with an accuracy and weighted F1-score of 0.9922 with the highest agreement statistics (Cohens Kappa = 0.9860, MCC = 0.9860). This model has performed better because of the nonlinear meta-learning nature of gradient boosting that enables the flexibility to weight the base learner responses in various parts of the feature space. The LightGBM meta-learner allowed the flexibility of the residual errors through repeated refinement to discover complementary model strengths between CatBoost, LightGBM, and Extra Trees, resulting in enhanced discrimination of closely related disease conditions. This adaptive combination is especially beneficial with medical datasets when the interaction of the variables is not generally linear and is quite intricate. A stacking model that used logistic regression as the meta-learner (Hybrid-2) was also competitive (F1-score = 0.9912) even with a regularized linear combination of base predictions, which implies that predictive signal can be captured by a regularized linear combination of the base predictions. Logistic regression is a stable and well-calibrated fusion mechanism that minimises variation and over-fitting.

Table 2. Predictive classification performance of the six hybrid ensemble models.

Hybrid Model	Accuracy	Precision (w)	Recall (w)	F1-Score (w)
Hybrid-3: Stacking (LightGBM meta)	0.9922	0.9922	0.9922	0.9922
Hybrid-2: Stacking (Logistic Regression meta)	0.9912	0.9912	0.9912	0.9912
Hybrid-6: Voting → Stacking (Logistic Regression)	0.9910	0.9910	0.9910	0.9910
Hybrid-5: Weighted Soft Voting	0.9886	0.9886	0.9886	0.9886
Hybrid-4: Soft Voting	0.9880	0.9880	0.9880	0.9880
Hybrid-1: Stacking (MLP meta)	0.9878	0.9878	0.9878	0.9878

However, its linear character limits its capacity to represent more complex dependencies among the base learner outputs and this is the reason why it does slightly worse than the nonlinear LightGBM meta-learner. The second-level hybrid architecture (Hybrid-6) comprising of soft voting with the subsequent application of logistic regression in the form of stacking achieved the same performance as Hybrid-2 without outperforming it. The intermediate voting phase adds a consensus-based dispensation that increases robustness but could additionally remove the informative variations in base learner forecasts prior to meta-learning. This indicates that, in this case, direct stacked generalization will be more effective as compared to hierarchical aggregation because too much averaging may decrease the information that is discriminative.

Ensembles on voting-based (Hybrid-4 and Hybrid-5) had a little lower predictive performance even though they are computationally simple. Equal-weight soft voting assumes comparable model reliability which is not plausible with uniform model reliability in practical situations. The addition of weights (Hybrid-5) offered small reductions of results because it focused on stronger learners, but both methods are still static aggregation schemes that are unable to adjust to local trends in the data. These results underscore the weakness of probability averaging in the case of complex, heterogeneous medical data, in which predictors and outcomes are related differently in groups of patients.

Table 3. Inter-rater agreement and class separability metrics for the hybrid ensemble models.

Hybrid Model	Cohen's Kappa	MCC	ROC-AUC
Hybrid-3: Stacking (LightGBM meta)	0.9860	0.9860	0.9989
Hybrid-2: Stacking (Logistic Regression meta)	0.9842	0.9842	0.9983
Hybrid-6: Voting → Stacking (Logistic Regression)	0.9838	0.9839	0.9983
Hybrid-5: Weighted Soft Voting	0.9795	0.9796	0.9987
Hybrid-4: Soft Voting	0.9784	0.9785	0.9990
Hybrid-1: Stacking (MLP meta)	0.9781	0.9781	0.9960

Stacking model using MLP meta-learner (Hybrid-1) produced the worst results of the configurations tested, but in general, the results were very good (F1-score = 0.9878). Although theoretically, neural networks can model nonlinear relationships, it has less advantages in structured tabular data with few dimensions. The extra optimization complexity of the neural meta-learner is not always translated into enhanced generalization, and on the other hand, even the tree-based boosting algorithms are more in line with the statistical properties of a clinical dataset. Importantly, very high values of ROC-AUC (≥ 0.9960) of all the models were noted, which is a good indication of well-separating classes. Agreement metrics like MCC and the Kappa as developed by Cohen, however, disclosed better statements of reliability showing that stacking based hybrids produce more steady predictions on multi-class instead of chance. This supports the need to assess the medical classification systems based on several complementary metrics instead of accuracy, or ROC-AUC.

All in all, the findings indicate that the evidence of the increase in performance in this problem is not merely a feature of the model complexity but rather a characteristic of the success of the strategy to integrate the ensembles. Adaptive stacking, especially with gradient boosting as a meta-learner offers the surest modeler of nonlinear clinical interactions and generalizes well. These results justify the appropriateness of sophisticated stacked ensembles to predict multi-class liver cirrhosis and validate the emerging information on how boosting-based meta-learning can still be used with systematic biomedical data (see Figure 3).

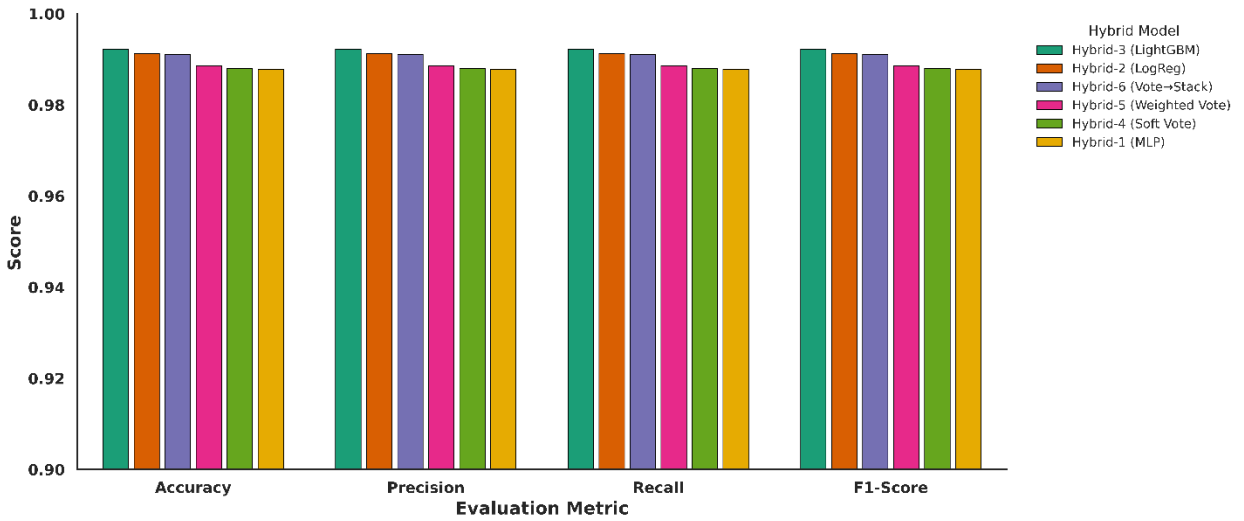

Figure 3. Metric-wise comparison of hybrid ensemble models across weighted evaluation measures, illustrating consistent performance differences among the proposed configurations.

3.2 Methodological Insight: Why Hybrid Meta-Learning Improves Prediction

The complementary learning behavior of the proposed framework heterogeneous base models provides the level of effectiveness. CatBoost reflects structured relations between clinical and nominal variables, LightGBM reflects nonlinear developmental dynamics based on the severity of the disease, and Extra Trees also adds randomization which minimizes the variance and enhances the generalization. Each of these learners focuses on different areas of the feature space, although, their predictions can be optimally combined using stacked generalization where the meta-learner estimates and learns how to resolve the conflicts between the base predictions. This dynamic combination helps the ensemble to take advantage of complementary boundaries of decisions instead of using fixed aggregation, which explains the better performance of the LightGBM-based stacking structure (Hybrid-3). Therefore, combining models does not just lead to an improvement, but a meta-learning of the relationships between them is a methodological step forward when compared to traditional hybrid ensembles.

3.3 Discussion on Confusion Matrix

The confusion matrices shown in Figure 4 give class level understanding on how the hybrid models distinguish differentiate between the statuses of cirrhosis. High abundance of diagonal dominance can be found in all configurations meaning that predictions are mostly correct in the three classes. Class C is categorized into a very high consistency, low leakage to other categories indicating that the ensembles tend to capture feature patterns that are possessed in early-stage cases. Class D is also highly separable, and there is only slight misclassification to C, which corresponds to the more definite clinical manifestation of severe disease.

The residual errors are mostly on class CL, which is the medium of transition, and there is little overlap with C and D, on all models. This can be considered a natural behavior because of the transitional character of this type of classes and demonstrates the natural complexity of modelling a borderline clinical condition as opposed to the flaws of the algorithm, itself. The stacking-based models, specifically Hybrid-3 and Hybrid-2, have even better concentration along the diagonal, so that they better address such ambiguous cases with adaptive meta-learning. Conversely, the approaches based on the voting are a little more dispersed, indicating that the averaging of fixed a-priori probabilities could be filtering out minor decision discontinuities.

3.6 Clinical Deployment Viewpoint and Applicability in Practical

In addition to the predictive performance, the proposed framework is developed to be practically integrated into the healthcare settings. The model will utilize only the routinely measured laboratory and demographic parameters and will not necessitate any additional diagnosis procedures and will operate as an entirely non-invasive process, without exposing the patient to the dangers of liver biopsy. With the automated preprocessing and inference stages within a single pipeline, the system may be integrated with electronic health record systems or clinical decision support systems (CDSS) and real-time risk stratification is offered. Such scalability allows screening the population at large and constant monitoring of patients, facilitating early intervention and resource optimization in the treatment of hepatology.

3.6 Limitations and Future Work

Although the proposed hybrid stacking framework has a high predictive performance, it is important to note that it has a number of limitations. To begin with, the research is based on one publicly available dataset, and this might not adequately represent all the variability that exists among various clinical institutions, patients and the standards of laboratory measurements. This dependency on data sets can affect the extrapolation of the model to other cohorts. Second, despite the preprocessing pipeline being created in a manner that would provide robustness with KNN imputation and scaling, retrospective clinical data might always harbor some unseen biases or unmeasured confounders that cannot be fully

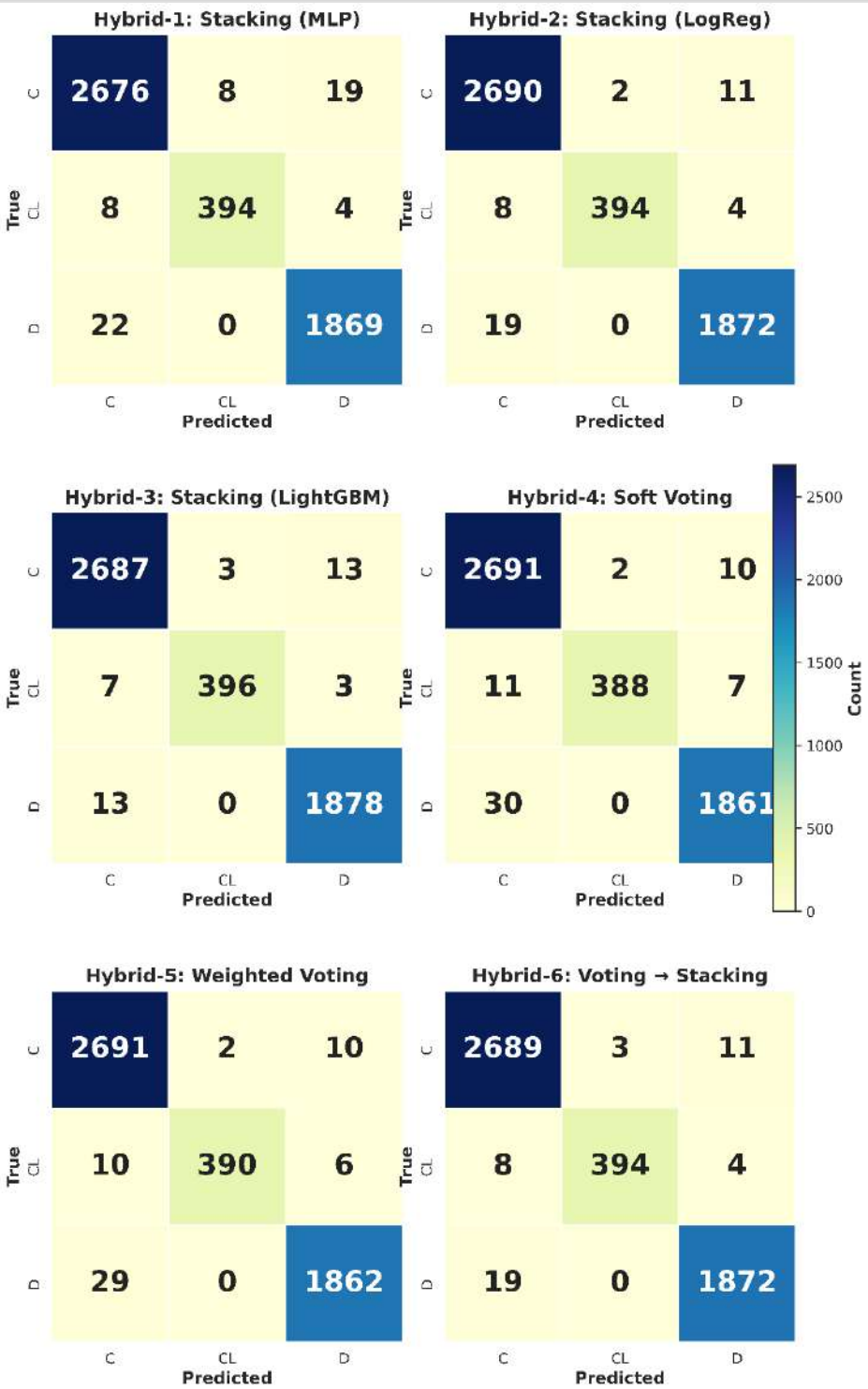

Figure 4. Confusion matrices of the six hybrid models showing class-wise prediction performance.

resolved with statistical learning. Third, the existing model assesses the state of patient records at one point in time whereas liver cirrhosis is a chronic condition that its mechanics could be better measured with a longitudinal study. Lastly, although ensemble models are very accurate in

prediction, their complex architecture can cause a loss in transparency when compared to less complex clinical scoring mechanisms, which can influence interpretability in practice-based medical decision-making. The next step in the work will be the verification of the suggested solution with the help of the multi-center datasets to check the external generalization and clinical transferability. The addition of longitudinal patient data would allow modeling changes in disease progression over time and better detecting a change in status. Further studies can also be conducted to determine how to incorporate explainability methods to further increase the trust of clinicians and the readability of predictions. Deploymentwise, initiating the framework into the real-time clinical decision support systems and assessing its influences on the diagnostic workflow and patient outcomes will be a significant step towards the practical implementation. These guidelines will help to address the gap between high-performance predictive modeling and everyday clinical practice.

4.0 Conclusion

The research proposed an enhanced hybrid stacking ensemble model of precise and credible prediction of liver cirrhosis status based on clinical and biochemical data. The proposed methodology was effective in solving the complexity and non-linearity of the multi-class liver cirrhosis classification problem in that it used heterogeneous base learners together with meta-learning strategies. The overall experimental analysis showed that stacking-based ensembles were better than voting-based hybrid models and the benefit of learning-based model integration is more than is possible through fixed aggregation techniques. The stacking ensemble, which uses LightGBM as the meta-learner, performed the best with respect to predictive ability, with high accuracy, high agreement measures, and high discriminative power. The high values of ROC-AUC, Cohen Kappa and Matthews value of correlation between models was an additional confirmation that the proposed framework was robust and reliable. These findings show that meta-learning is not only more predictive, but also more stable and more agreeing than random and this is important when applying it to clinical decisions. The results of this research imply that enhanced stacking ensembles are effective at the task of capturing complicated interactions between clinical features and base model predictions. Through a systematic comparison of various hybrid frameworks within a common preprocessing and evaluation model, the present study offers useful knowledge with regards to designing high performance ensemble models in medical classification problems. The suggested methodology provides a flexible and extensible platform of creation of intelligent clinical decision support systems to enhance early diagnosis and disease status based. Further studies could examine how longitudinal patient data, external validation on other cohorts and improved methods of model interpretability can be combined to add clinical applicability. As a whole, this research paper shows that meta-learning-based hybrid stacking ensembles are an effective and dependable method of liver cirrhosis status-based prediction, and have a lot of prospective use in the healthcare analytic domain.

REFERENCES

Lee MJ. A review of liver fibrosis and cirrhosis regression. Journal of pathology and translational medicine. 2023 Jun 20;57(4):189-95.

Khan MA, Afrin F, Prity FS, Ahammad I, Fatema S, Prosad R, Hasan MK, Uddin M. An effective approach for early liver disease prediction and sensitivity analysis. Iran Journal of Computer Science. 2023 Dec;6(4):277-95.

Kotak PS, Kumar J, Kumar S, Varma A, Acharya S, Kumar Jr MJ. Navigating cirrhosis: a comprehensive review of liver scoring systems for diagnosis and prognosis. Cureus. 2024 Mar 29;16(3).

Shaban WM. Early diagnosis of liver disease using improved binary butterfly optimization and machine learning algorithms. Multimedia Tools and Applications. 2024 Mar;83(10):30867-95.

Hossain AA, Ahamed IU, Gupta UD, Anika AN, Ahamed IU. Stratified prognostication and interventional strategies in chronic hepatic diseases: An ensemble machine learning approach. In2024 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET) 2024 Apr 27 (pp. 1-6). IEEE.

Khan N, Nauman M, Almadhor AS, Akhtar N, Alghuried A, Alhudhaif A. Guaranteeing correctness in black-box machine learning: A fusion of explainable AI and formal methods for healthcare decision-making. IEEE Access. 2024 Jun 28;12:90299-316.

Jasim AA, Alwindawi H, Hazim LR. Empowering Diagnostics: An Ensemble Machine Learning Model for Early Liver Disease Detection. Al-Iraqia Journal for Scientific Engineering Research. 2025 Jun 14;4(2):13-9.

Shen D, Sha L, Yang L, Gu X. Identification of multiple complications as independent risk factors associated with 1-, 3-, and 5-year mortality in hepatitis B-associated cirrhosis patients. BMC Infectious Diseases. 2025 Feb 1;25(1):151.

Ramamoorthy K, Rajaguru H. Exploitation of bio-inspired classifiers for performance enhancement in liver cirrhosis detection from ultrasonic images. Biomimetics. 2024 Jun 14;9(6):356.

Mohamed MH, Ali BH, Taloba AI, Aseeri AO, Abd Elaziz M, El-sappgah S, El-Rashidy N. Towards an Accurate Liver Disease Prediction Based on Two-level Ensemble Stacking Model. IEEE Access. 2024 Sep 16.

Haque ME, Islam SM, Mia S, Sharmin R, Morshed MS, Huque MT. StackLiverNet: A Novel Stacked Ensemble Model for Accurate and Interpretable Liver Disease Detection. arXiv preprint arXiv:2508.00117. 2025 Jul 31.

Alotaibi A. Ensemble Deep Learning Approaches in Health Care: A Review. Computers, Materials & Continua. 2025 Mar 1;82(3).

Ibrahim N, Rajalakshmi NR, Sivakumar V, Sharmila L. An optimized hybrid ensemble machine learning model combining multiple classifiers for detecting advanced persistent threats in networks. Journal of Big Data. 2025 Dec;12(1):1-28.

Catalano G, Alaimo L, Chatzipanagiotou OP, Ruzzenente A, Ratti F, Aldrighetti L, Marques HP, Cauchy F, Lam V, Poultsides GA, Hugh T. Predicting the complexity of minimally invasive liver resection for hepatocellular carcinoma using machine learning. HPB. 2025 Mar 4.

Gayathri D, Shantharajah SP. MetaStackD A robust meta learning based deep ensemble model for prediction of sensors battery life in IoE environment. Scientific Reports. 2025 Apr 29;15(1):14967.