Revolutionizing livercare: predicting liver cirrhosis

A Project Report Submitted in partial fulfillment of the requirements Of

ARTIFICIALINTELLIGENCEANDMACHINELEARNING

Internship at

SMARTBRIDGE in collaboration **APSCHE**

By

1. **Pamula Sumalatha (22F01A04I6)**
2. **Nune Venkata Siva Bhavya(22F01A04H9)**
3. **Narra Nagapushpa (22F01A04H2)**
4. **Narra Nikhitha(22F01A04H3)**

TeamID:

LTVIP2025TMID40792

ST.ANN'S COLLEGE OF ENGINEERING AND TECHNOLOGY

1. Introduction

We leverage machine learning techniques to predict outcomes for patients with liver cirrhosis based on clinical data. By analysing a dataset comprising various features related to liver disease and patientcharacteristics,thisstudyidentifieskeypredictorsofcirrhosisoutcomesanddevelopspredictive modelsthatcanaidcliniciansin decisionmakingandmanagementof thecondition.

Our report explains our process for data pre-processing, exploratoryanalysis, model development,and evaluation, yieldinguseful models to predict cirrhosis outcomes to assist clinicians. Our modelling approach will also be applicable to larger datasets on liver health, so our pilot study reported herecanbe scaledup to improveperformancefurther.

Livercirrhosisisasignificantglobalhealthburden,characterizedbyprogressivedeteriorationof liver function due to long-standing liver injury. The disease progression is often silent and can culminate in life-threatening complications without timely intervention. Early and accurate prediction of cirrhosis outcomes can significantly enhance clinical decision-making and patient care. We identify which medical markers are predictive of liver cirrhosis outcomes so these can be gathered and used with our modeltopredictcareneedsanddiseaseoutcomes.

Theobjectivesofthisprojectwere:

a) Exploratorydataanalysistouncoverpatternsandinsightswithinthelivercirrhosisdataset.
b) Developandevaluate machinelearningmodels,specificallyfocusingonGradientBoostingandRandomForest,to predict liver cirrhosis outcomes.
c) Identifykeyfeaturesinfluencingcirrhosisprognosisandassessthemodels'predictiveperforma ncethroughvarious metrics.
d) Provideactionablerecommendationsforcliniciansbasedontheanalysisfindingsandsuggest whatfuturereal-world data should be gathered.

2. DataDescriptionandPre-processing

Our data set is synthetic and based on the MayoClinicLiverCirrhosisStudy(1974-84)The data features include many medical measures and demographic variables for hypothetical patients with liver cirrhosis.Each record represents a patient's clinical profile and outcome . Our training data is complete and consists of data entries for 7905 patients across 19 medical features with their liver cirrhosis outcomes . Outcomes were measured after Ndays under observation, with three possible outcomes:***'C'- patientisstillalive,'CL'-patienthadlivertransplantandisstillalive,*** or***'D'-patientdied*.**Thedataisimbalancedwith*62.8% C,3.5% CL, 33.7%D*, making CLaminority class.A separate hold out test dataset consists of 5271 entries with outcomes removed . Our task is top redictthe probability of each outcome class for each entry in this test dataset,and score ourmodel's predictions against their truevalues.We developed our model entirely on the training dataset so that our final score against the test dataset will fairlymeasure how ourmodel might perform onnewpatient data inaclinical setting.

Data Features Description: Each medical data feature indicates the patient's condition and are fed into our model training

andanalysedbyourtrainedmodeltopredictlivercirrhosisoutcomesandtherebyassisteffectiveman agementofthe
disease.The dataset includes the following key features:Age, Sex, Bilirubin, Albumin, Cholesterol, Drug, Ascites, Hepatomegaly, Spiders, Edema, Copper, Alk Phos (Alkaline Phosphatase), SGOT, Triglycerides, Platelets, Prothrombin Time, and Stage. The target variable, "Status," categorizes patient outcomes into C,CL,D referring to Alive with Cirrhosis (‘C’), Alive with Transplant (‘CL’), or Dead(‘D’).Transplant recipients who nonetheless die soonafterwards are assigned D, so CL is a label forsuccessful transplantrecipients.AppendixDexplainseachfeatureanditsmedicalsignificanceto liver cirrhosis.

Pre-processingSteps

1. **Datacleaning:**Initially, we confirmed our synthetic dataset was completed with no missing values or inconsistent data entries.Later,we simulatedreal world data through adeletion and imputation study that simulated missing datavalues and addressed these through appropriate imputation strategies, ensuring acompleted ataset for analysis.

2. **Featureencoding:**Categoricalvariablessuchas Sex,Drug,Ascites,Hepatomegaly,Spiders,andEdemawereencoded usingone-hotencodingtoconvertthemintoanumericalformatsuitableformachinelearningalgorithms.
3. **Feature scaling:** Continuous features were standardised to have zero mean and unit variance. This scaling ensures that features with larger magnitudes do not unduly influence the model's performance.
4. **Datasplitting:**Initially,wedividedthedatasetintotrainingandvalidationsetsfollowingan80/20split.Stratified samplingwasemployedtomaintaintheproportionofeachoutcomeclassinbothsubsets,solvingthe riskofclass imbalancebeingexacerbatedinthesplit.Later,wewillusecross-validationtechniquestoovercome possibleoverfittingto biases introduced by a given choice of 80/20 split.

Through these pre-processing steps, the dataset was transformed into a clean, well-structured formatconducive to effective machine learning model development. These steps are crucial for ensuring that subsequent analyses are based on reliable and consistent data.

3. ExploratoryDataAnalysis

Prior to model development, we conducted an exploratory data analysis (EDA) to gain insights into the dataset's characteristics and uncover any inherent patterns or anomalies. This analysis helped us understandofthedata'sstructure,distributions,andrelationshipsbetweenfeatures.

Keyfindings:

- **Descriptivestatistics:**Summarystatisticsprovidedanoverviewofthecentraltendency,dispersion,andshapeofthe dataset's distributions. Notably, variables such as Bilirubin and Albumin exhibited skewed distributions, indicating the presence of outliers or non-normal distribution patterns.
- **Correlation analysis:** A correlation matrix was generated to identify potential relationships between features. This analysis helped in understanding which variables might have collinear relationships, aiding in the feature selection process to avoid multicollinearity in thepredictivemodels.
- **Classdistribution:**Thedistributionofthetargetvariable,'Status',wasexamined,revealinga classimbalance witha higher prevalence of the 'Alive Cirrhosis (C)' outcome compared to 'Dead (D)' and'Alive Transplant (CL)'. This insight guided the choice of stratified sampling and evaluation metrics sensitive toclassimbalance.
- **Visualexplorations:**Histograms,boxplots,andscatterplotswereusedtovisualisethedistributionsofindividual featuresandtheirinteractions.Forinstance,boxplots highlightedthepresenceofoutliers infeatures likeBilirubin, suggesting potential issues with extreme values influencing model performance.
- **Featureexplorations:**Throughscatterplotsandpairplots,relationshipsbetweenkeyfeatureswereexplored.Notable observationsincludedthepotentialrelationshipsbetweenBilirubinlevelsandliverdiseaseoutcomes,reinforcing clinicalunderstanding that higher bilirubin levels might be associated with worse prognosis.

Theexploratory data analysis guided oursubsequent datapre-processing and model development stages. We are confident our feature handling is careful enough and encouraged that our feature variables will predictlivercirrhosisoutcomes,layingasolidfoundationforourpredictivemodelling.

4. Imputationmethods

For our studywe chose median, K Nearest Neighbors (KNN) andmultiple imputationof chained equations (MICE) as ourmethods. We used random deletions to remove 10%of the data values. Ourdata values were chosenby choosingrandomlyacross the columnsand rows. Thiswasdone to simulate the randomincompleteness of real-worlddata where the diagnosticsuite is not always fully available. KNN estimates the values of the missing data based on the similarity to is neighbors whilethe median imputation takes the median of the remaining values of the attribute in the dataset and estimates it as the values for the missing data. MICE replaces the missing values with multipleimputations usinga regressionmodel iterativelyuntil theconvergencecriteria is met. Mean-Squared Error (MSE) was chosen as a metric to compare the results of the differentimputationsand MICE turnedout to be the best amongst the three methodswithan average MSEof 0.000901. We also usedXGBoost to compare the imputation methodsand MICE again turnedout to be the best amongst thethree methods with a log loss of 0.510.

5. Tacklingimbalancedclassdata

Ourdatasetisimbalancedwithonly3%CLcomparedto63%C and34%D. So,weaddressthis class imbalanceto avoid biastowards overrepresented classes C and D and neglect of the minority classCL.Weconsideredtwotechniques,‘weightadjustment’andanoversamplingtechniquecalled

‘SMOTE’. Weight adjustment simply gives more weight to CL data in inverse proportion to its under-representation in the dataset. Whereas our oversampling technique SMOTE tried to balance the data by inferringadditional CLdataentriesbasedontheexistingdata (AppendixE). SMOTEofferedeasy gains for the non-optimizedmodels. Although in the case of the optimized model, we found that class weight adjustment could converge on a slightly better log loss in the end. The result for the fully optimizedgradientboostingmodel+classweightingwasavalidationloglossof0.46comparedto 0.53withSMOTE.

6. Methodology

This section outlines the systematic approach employed in developing and evaluating predictivemodels for liver cirrhosisoutcomes. The methodology encompasses model selection, training,evaluation, and validation processes, underpinned by arigorous data analysis framework. For modelselection, given the nature of the outcome variable (categorical with three levels), Gradient Boosting and Random Forest classifiers were chosen for their robustness and efficacy in handling multi-classclassification problems. These models are well-suited for dealing with imbalanced datasets and offer interpretableinsightsintofeatureimportance.NeuralNetworkswerealsoexplored.

For training, the dataset was partitioned into training (80%) and validation (20%) sets using stratifiedsampling to maintainoutcome class proportions, addressing the issue of class imbalance. Thispartitioning ensures that the models are evaluated onunseen data, providing a fair estimate of their generalisation performance. Training data is then oversampled with SMOTE. The Gradient BoostingClassifier was trained using a range of hyperparameters to optimise performance, with a focus onlearningrate, number of estimators, andtree depth. Hyperparameter tuningwas conductedusing grid searchwithcross-validationtoidentifytheoptimalconfiguration.LikeGradientBoosting,the RandomForestmodelunderwenthyperparametertuningwithafocus on thenumber of estimators, maxdepth,andminsamplessplit.Theclassweightparameterwasadjustedtoaddressclass imbalance,enhancingthemodel'ssensitivitytominorityclasses.

For evaluation, we favoured LogLoss in our development but also measured accuracy, precision, recall, and F1-scoremetrics forourmodels,offeringa comprehensiveviewof theirperformance, particularly on the minority transplant CL. Given theimbalancednature of the dataset, particular emphasis was placed on recall and F1-score to ensure that the models effectively identifytheminorityclass CL.

Resultsafterfine-tuning(Log-lossonvalidationset):

1. SMOTE+RandomForest : 0.489
2. SMOTE+GradientBoosting : 0.525
3. ClassWeights+GradientBoosting:0.459
4. SMOTE+XGBoost : 0.576

ResultsofthesamefineͲtunedmodelsonKaggleLogͲloss(Public,PrivateScores):

1. SMOTE+RandomForest :0.49,0.50
2. SMOTE+GradientBoosting :0.49,0.48
3. **ClassWeights+GradientBoosting:0.43,0.42**
4. SMOTE+XGB :0.53,0.50

Conclusion:BasingourevaluationonourKaggleresult,wedeclare**ClassWeights+GradientBoosting**asourbestmodel.

To assess the models' stability and generalization capability, k-fold cross-validation was employed.This technique provided insights into the models' performance variance across different data subsets,reinforcing the robustness of the evaluation. Post- evaluation, an error analysis was conducted to identify and scrutinize instances where the models' predictions deviated from the actualoutcomes. This analysisprovideddeeperinsightsintothemodels'limitationsandpotentialareasforimprovement. Ourevaluationestablisheda clear understanding of the models' capabilities andlimitations, guiding future efforts to refine anddeploythemeffectively.

Resultsofcross-validationsonthesamefine-tunedmodels:

1. SMOTE+RandomForest : 0.332
2. SMOTE+GradientBoosting : 0.272

3. ClassWeights+GradientBoosting:0.447
4. SMOTE+XGB : 0.265

However, we concluded that these results are implausible and symptomatic of overfitting to thetraining data, which explainswhy these models perform so poorly on the test dataset. That is, except for the ClassWeights + Gradient Boosting model, whose 0.45loglossresulthereisplausibleandconsistentwithitsotherresults,includingonthetestdataset.

7. ModelDevelopmentandEvaluation

This section explains how we developed and evaluated the Gradient Boosting and Random Forest models, detailing each step frominitialtrainingtothefinalevaluation.

Forinitialmodeltraining,theGradientBoostingandRandomForestmodelswereinitiallytrained usingthe defaulthyperparameters. This preliminarytraining servedas a baseline to assess the models' naturalperformanceonthedatasetwithoutanytuning.ForGradientBoosting,themodel demonstrated promising results, capturing complex patterns in the data. However, there was room forimprovement,particularlyontheminorityclassCL.Wealsowishedtoreduceoverfittingand enhance its abilityto generalise. Similarly, our simplest RandomForest model provideda robust baseline performance. Itsinherent randomness and ensemble design offered a strong start, yet it was apparent that hyperparameter tuning could furtheroptimiseitsperformance.

HyperparameterTuning:

Hyperparameter tuningwas conductedusing a grid searchapproach, focusingon keyparameters that influence model complexity and learning dynamics. For Gradient Boosting, parameters like `n_estimators`, `learning_rate`, and `max_depth` were tuned. The optimal configuration significantly improved the model's accuracy and reduced the likelihood of overfitting. In thecase of RandomForest,tuningfocusedon`n_estimators`,`max_depth`,and`min_samples_split`.Adjustingthe `class_weight` parameter proved crucial in addressing the class imbalance, enhancing the model'sperformanceacrossallclasses.ForXGBoostClassifier,tuningfocusedoncontrollingoverfittingwith `n_estimators`, `max_depth` and `subsample`. The scoring metric for GridSearchCV was log loss andwemeasuredprecisiontoo.Duringdevelopment,accuracyandneg_log_losshadbeentestedas scoring method but both result in poorer log loss values for validation. K-fold cross-validation further validatedtherobustnessofthetunedmodels,providinginsightsintotheirstabilityacrossdifferent datasegments.Bothmodelsshowedconsistentperformanceacrossfolds,affirmingtheirreliability and generalisation capability.

FeatureImportanceAnalysis:

Feature importance analysis revealed critical predictors for each model, offering insights into thefactors most influential indetermining liver cirrhosis outcomes. In the Gradient Boosting model,features like Bilirubin (0.42), Days With Cirrhosis WithinTrial (0.15), and Prothrombin (0.10) stood out as top predictors, aligning with clinical expectations regarding their significance in liver disease progression. TheRandomForest model gave lessimportanceto eachfeature,withtop importance’s for Bilirubin(0.14), Age (0.11,) Copper (0.08), and Days With Cirrhosis Within Trial (0.08). So the RandomForestmodelalsounderscoredtherelativeimportanceof demographicfactors,illustrating the multifaceted nature of cirrhosis prognosis.

Error Analysis:

Erroranalysispinpointedspecificinstanceswherethemodelsfaltered,ofteninpredictingless prevalent classes. Ourinvestigation of these errors highlighted potential biases and areas whereadditional data or feature engineering might improvemodel accuracy, although we leave that as future work. Using the optimised Gradient Boost model on validation set, we analyzed the distribution of misclassifications (seeAppendixB). Our evaluationestablisheda clear understandingof the models' capabilitiesandlimitations,guidingfutureeffortstorefineanddeploythemeffectively.

8. KeyInsightsandInterpretations

Thissectionsynthesisesthefindingsfromthemodeldevelopmentandevaluationstages,translatingthe technicalresultsinto interpretableinsightsandactionableintelligenceforstakeholders.

- **Predictivepowerofclinicalfeatures:**Thefeatureimportanceanalysisshowstheclinicalrelevanceofcertainfeatures (e.g.Bilirubin,Albumin)inpredictingcirrhosisoutcomes.Theirprominenceinthemodelsalignswiththeprevailing medicalunderstanding,whichreinforcesthatourmodelsarevalid,andallowsourmodelstoofferdata-driven affirmation to clinicians who use our models.
- **ModelPerformance and Real-World Application:** BothGradient BoostingandRandom Forestmodels demonstrated robust performance in predicting cirrhosis outcomes. Their ability to handle incomplete datasets, as often encounteredin real-worldscenarios, makes themparticularlyvaluable toolsfor healthcarepractitioners. When applied tonew, incompletepatientdata,ourmodelscanstillprovidereliableoutcomepredictions,aiding clinicaldecision- making.
- **Scalabilitywithmoredata:**Ourmodellingapproachisdesignedtoscaleeffectivelywithadditionaldata.Should clientsprovideaccesstolarger,evenifincomplete,historicaldatasets,themodelscanberetrainedorfine-tunedto enhancetheirpredictiveaccuracyfurther.Thisadaptabilityensuresthatourmodelsremainrelevantandusefulasmore data become available, continually improving their predictive capabilities.

- **Implicationswithhealthcareproviders:**Byintegratingthesepredictivemodelsintotheirworkflow,healthcare providerscangainforesightintopotentialcirrhosisoutcomes,enablingmorepersonalisedandtimelyinterventions.This proactiveapproachcouldsignificantly improvepatientmanagementandoutcomesinliver cirrhosiscare.

9. RecommendationandFutureWork

Recommendation:

- We recommend using our Gradient Boost model to predict patient outcomes in a clinical setting from any new real-world data gathered by those clinicians.
- We also recommend our approach to modelling and encourage training similar models on larger datasets to improve performance in future.

FutureWork:

- **DataCollection:**Weencouragecontinuousdatacollectionforfurtherdevelopment,especiallyfortheunderrepresented transplant outcomeclassandthemost significantfeatures,in ordertoreducebiasandimprovemodelperformance. Incomplete data should not be a deterrent; our models are designed to accommodate and make the best use of available information.
- **Model Refinement:** As more data become available, iterative retraining and refinement will allow our models adapt to new patterns and information, optimising their predictive accuracy.
- **Integration into Clinical Practice:** Clinicians will find our models most useful upon the development user-friendlyinterfacesforhealthcareproviderstoleverage thepredictive modelsefficiently, ensuringthattheinsightsareaccessibleand actionable within the clinical workflow.
- **Future Research:** Further research should explore the integration of additional data types, such as genetic or lifestyle factors, into the models to enrich the predictive context and enhance outcome prediction.

10. Conclusion

Our models demonstrate significantpotential for predictingliver cirrhosis outcomes ina clinical setting. Clinicians may gather new real-world data then use our model to predict liver cirrhosis outcomes for their patients to guide their care. The models' ability to discern critical patterns and provide accuratepredictions,evenwithincompletedata,underscorestheirutilityinaclinicalsetting.Asweexpand our datasets and refine our models, we anticipate even greater utility for liver diseasemanagement,ultimately advancingpatient care and outcomes.

References

[1] Kaggle. (n.d.). Synthetic Liver Cirrhosis Dataset, Kaggle: Playground Series S3E26. Retrieved March 11, 2024, fromhttps://www.kaggle.com/competitions/playground-series-s3e26/

[2] Arvidsson, J. (n.d.). Mayo Clinic Liver Cirrhosis Dataset 1974-84, Kaggle: Mayo Clinic. Retrieved March 12, 2024, from https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-prediction

Appendices

Appendix A–Scoresforourcandidatemodels:

Kagglescoresforourcandidatemodels:

Model	PrivateScore	PublicScore	
XGB+SMOTE	0.53	0.53	
GB+SMOTE	0.49	0.48	
RF+SMOTE	0.49	0.50	
GB+WEIGHTED	**0.43**	**0.42**	**<–Best**

Classification ReportscoresforGB+WeightedontheValidationSet:

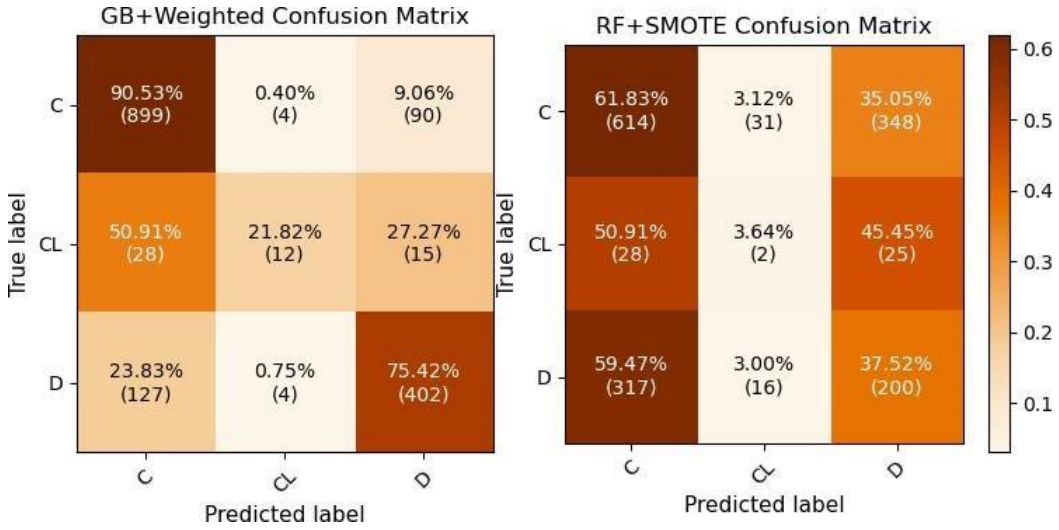
Class	Precision	Recall	F1-Score	Support
C	0.85	0.91	0.88	993
CL	**0.60**	**0.22**	**0.32**	55
D	0.79	0.75	0.77	533
Accuracy			0.83	1581
MacroAverage	0.75	0.63	0.66	1581
WeightedAverage	0.82	0.83	0.82	1581

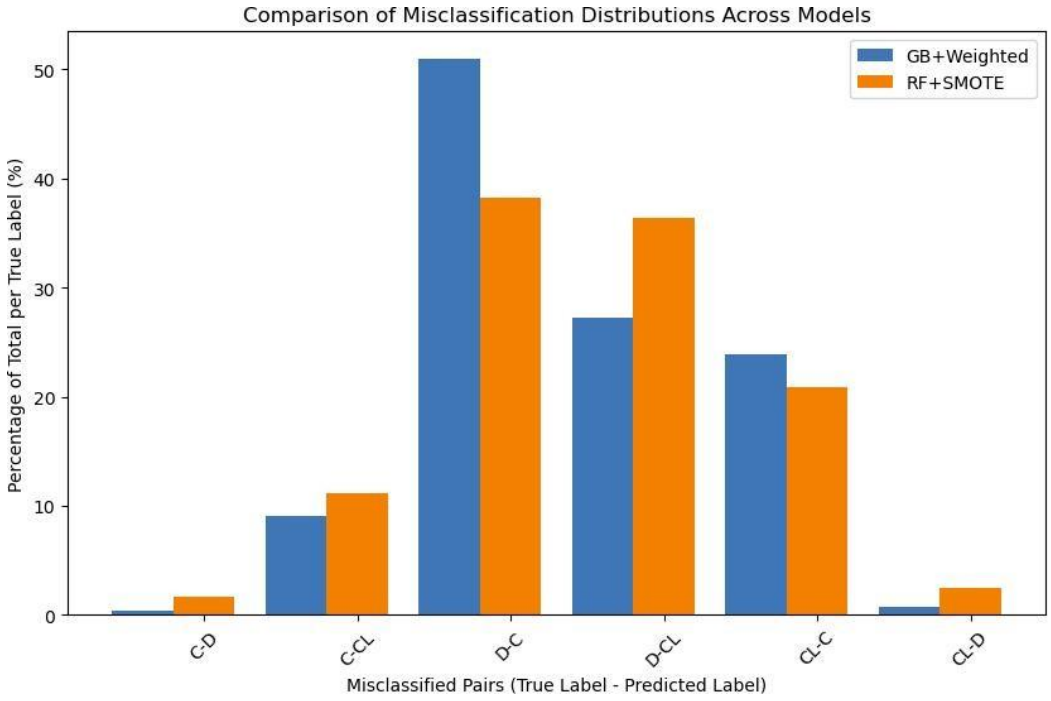
AppendixB–MisclassificationStudy:

ConfusionMatricesandHistogramofMisclassificationPairs

We prepared confusion matrices for both our best models to help us see whether and how they misclassify patients.

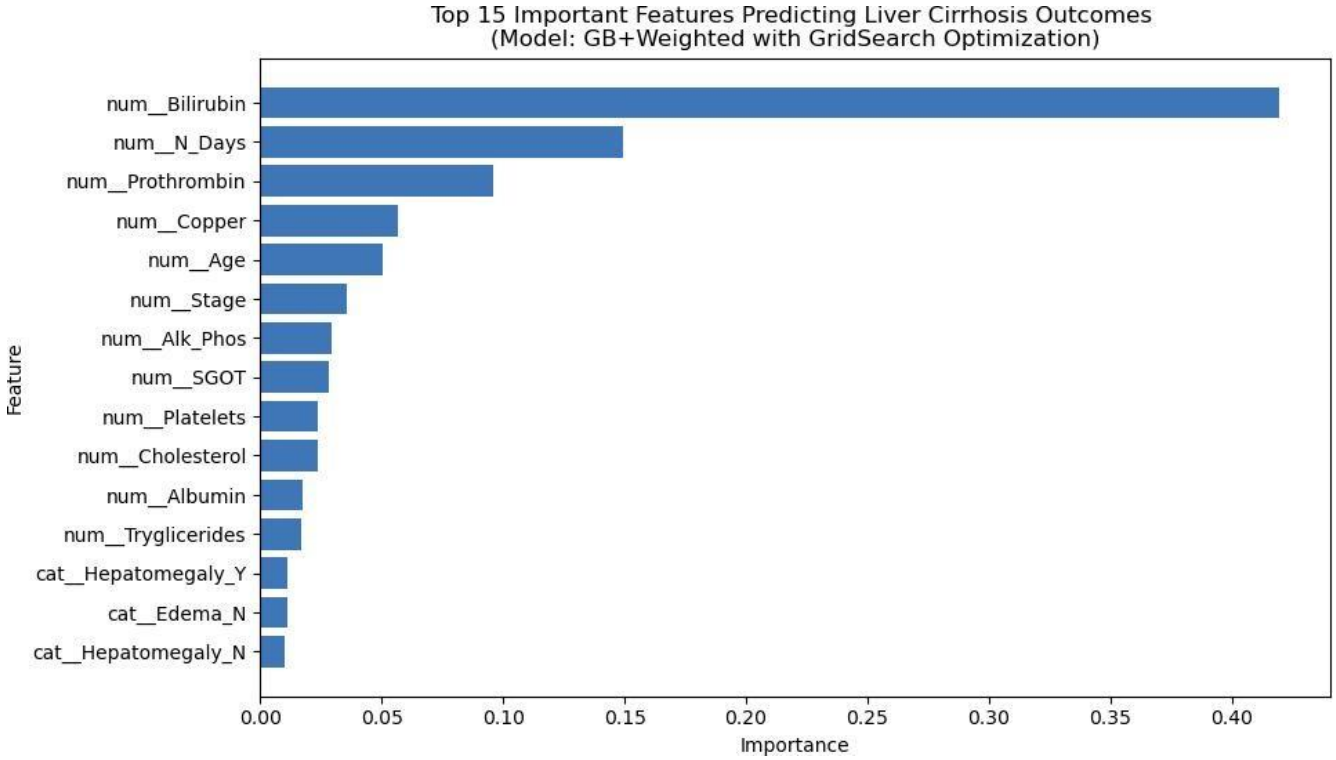
1. TheGB+Weightedmodellooksgood.Ofits20livertransplantpredictions,itsuggests12outof20livertransplants should go to those whose True label is CL.
2. TheRF+SMOTEmodellooksbadandlittlebetterthanguessing.Itsuggestswasting31outof49livertransplantson those who would live anyway with True label C.

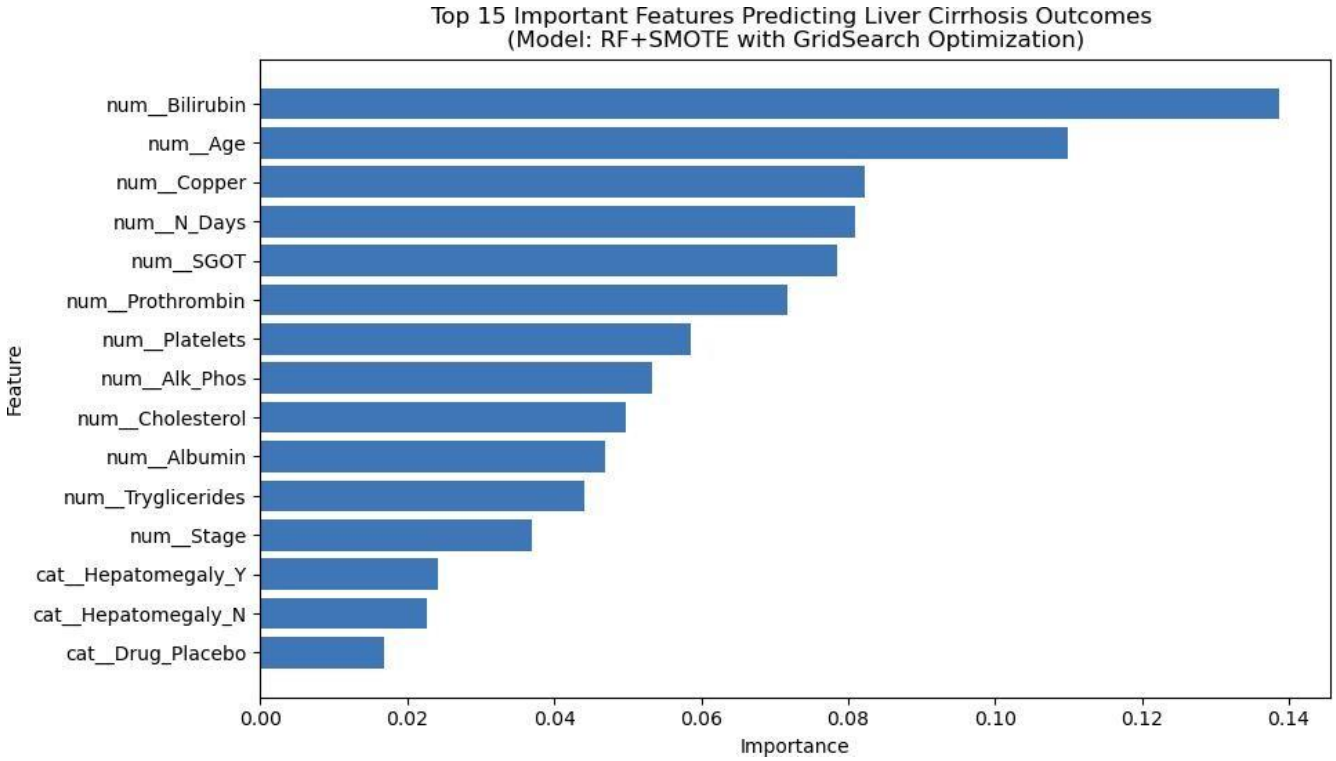




AppendixC–FeatureImportanceStudy:

We rankeddata features by their importance to eachof our best twomodels. Thisis a measure of how usefuldataoneachfeatureistohelpingthemodeldiscriminateanddecidebetweeneachclassC,CL, D. We recommend prioritizing gathering themost important data in a clinical setting in order to allowour model to best assist clinicians with predictions of disease outcome.





N Days: A count of days suffering from cirrhosis for the patient in the study. Cirrhosis is associated with a chance of deterioration towards death. So `N Days` may predict a deathoutcome.

Age: The age of the patient can be a critical factor in the progression and management of liver cirrhosis. Older patients might have a different prognosis due to the natural aging process and potential comorbidities.

Sex: There might be sex-specific differences in the progression of liver diseases, with some studies suggesting variations in the susceptibilityto liver injury andfibrosis between malesand females.

Bilirubin: This is a by-product of red blood cell breakdown, and elevated levels can indicate liver dysfunction, oftenassociatedwithworseoutcomesincirrhosis.

Albumin:A protein made by the liver, low albumin levels can indicate poor liver function and are associated with more advanced liver disease.

Cholesterol: While less directly related to cirrhosis, abnormal cholesterol levels can be associatedwithliverdysfunction and other metabolic conditions affecting the liver.

Drug (Treatment Type): The type of treatment can influence the progression of the disease. This could reflect different management strategies for cirrhosis and potentially differentoutcomes.

Ascites: The accumulation of fluid in the abdomen, indicating advanced liver disease, significantly impacts patient prognosis.

Hepatomegaly: Liver enlargement can be a sign of liver disease progression and is relevant in assessing the disease's severity.

Spiders (Spider Angiomata): These are small, spider-like blood vessels visible on the skin and can be an indicator of liver disease.

Edema: Swelling due to fluid accumulation, often in the lower legs, can indicate worsening liver function.

Copper: Elevatedcopper levels in the blood can be found in certain types of liver disease,such as Wilson's disease. Alk Phos (Alkaline Phosphatase): An enzyme related to the bile ducts; high levels can indicate bile duct obstruction or otherliverdiseases.

SGOT (AST): An enzyme found in the liver, heart, and other tissues. Elevated levels canindicate liver damage. **Tryglicerides:** While primarily a cardiovascular risk marker, abnormalities can also reflect metabolic issues related to liverdisease.

Platelets: Liver cirrhosis can lead to thrombocytopenia (low platelet count), which is a critical factor in assessing liver functionand diseaseseverity.

Prothrombin: Clotting factor made bythe liver. Liver damage can reduce levels, indicating liver dysfunction.

Stage: This reflects the severity of cirrhosis, with higher stages indicating more advanced disease and typically worse outcomes.

Status(Outcome): Thetargetvariable,indicatingthepatient'soutcome-eitheralivewith cirrhosis(C),alivewithtransplant(CL),ordeath(D),eachofwhichthemodelsaimtopredict based on the features.

AppendixE–SMOTEExplanation

SMOTE (Synthetic Minority Over-sampling Technique) enhances dataset balance by generatingsynthetic samples from the minority class in an imbalanced dataset to avoid model bias toward the majority class. SMOTE can help achieve more robust and accurate predictive models, although oneshouldtakecare to ensurethe syntheticdatapointsare medically plausible.

Howitworks:

1. **NeighborIdentification:** Foreachsampleintheminority class,identifyitsk-nearestneighborsinthefeature space. **Q:Howwouldwesorttheneighborsinfeaturespace?**Thereismorethanonefeature,sothere'smorethan onefeatureaxis,andtheknearestneighborswithrespecttoeachfeatureardifferent,sowhichlistshouldI draw the k nearest neighbors from?

   A: Weighted Distance Metric: We weight the distance metric according to the importance of each feature. The Euclidean distance between two points in an n-dimensionalspacecanbeweightedwithrespecttoeachfeaturewhereweightreflects the importance of each feature. Then we sort the nearest neighbors according tothis metricinordertorespecttherelativeimportanceofdifferentfeatures.Intheabsenceof ameasureoffeatureimportance,makeaneducatedguessbasedondomainknowledge or preliminary analysis.

2. **SyntheticSample Creation:**Createasyntheticdata entrythatliesbetweenachosendataentrybelongingtoa minorityclassandoneofitsk-nearestneighbors.Foreachdataentrybelongingtoaminorityclass,selectarandom neighborfromitsk-nearestneighbors.Imagineplottingboththesedatapointsanddrawinga'line'betweenthesetwo pointsin'featurespace'.Generateasyntheticsamplealongthelinesegmentconnectingtheminorityclassdataentry and its selected neighbor. The syntheticsample pointwe generate will havefeature values that are a linearcombination of thetwooriginaldatapoints,withthecombinationcoefficientsrandomlychosenbetween0and1.Sothefeaturesofthis synthetic data entry necessarily lie between those of our two original points.
3. **Repetition:**Continuethisprocessuntilthedatasetachievesthedesiredlevelofbalanceoruntiltherequirednumberof synthetic samples is generated.

Caution:

- SMOTEcanintroduceartificialbias,particularlyifsyntheticpointsdonotwellrepresenttheunderlyingmedical scenarios as hoped. One could scrutinize the medical plausibility of synthetic data.
- Validateonunmodifiedtestdataset,toensureimprovementsinperformancemetrics aregenuineinsteadofoverfittingto synthetic samples.