Evaluation Report on Microbiota Status Classification

Objective

Need to develop a model to classify patients' gut microbiota status into three categories ('Optimal', 'Suboptimal', and 'At Risk') using health indicators, medical history, dietary habits, and lifestyle factors.

Preprocessing Techniques Used

- Unnecessary feature removal
- Convert multi-item columns into numerical
- Target encoding
- Convert boolean columns into numerical
- Outlier detection and removal
- SHAP analysis
- Using SHAP-value based weight to group multi-item features
- Creating composite features using domain knowledge
- Feature scaling (Standard Scaling)
- Balancing target class (using SMOTE)

Model's used

- Logistic regression
- Random Forest
- XgBoost
- ANN
- LightGBM
- MLP with feature interaction
- Stacking of XGboost + MLP + Logistic regression
- Hierarchical XgBoost
- Tab Transformer
- Hierarchical Tab Transformer

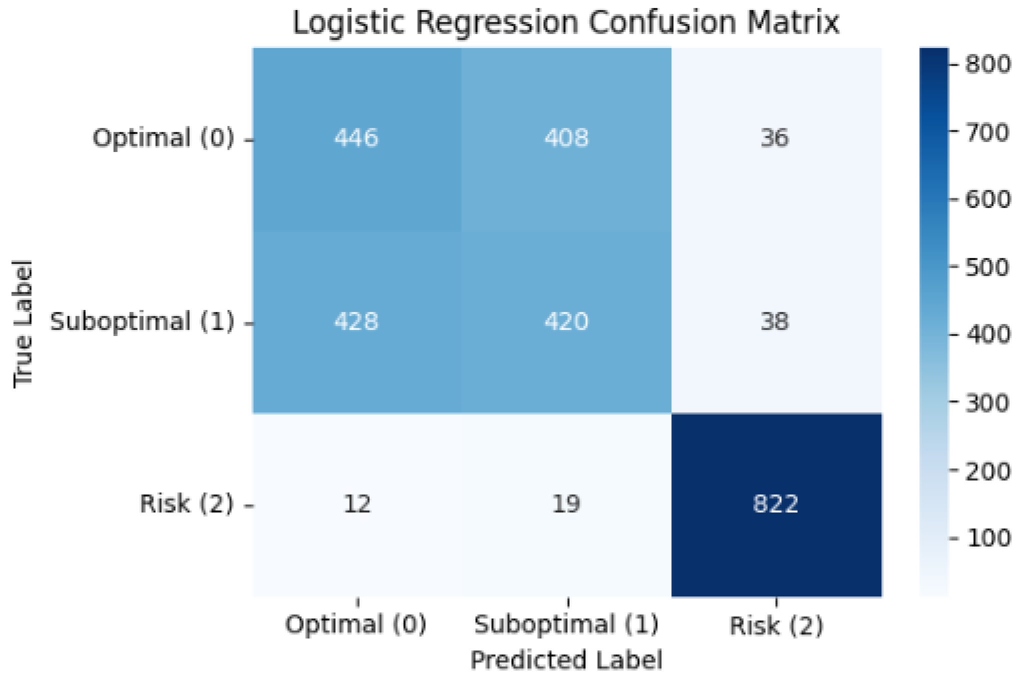
Hyperparameter tuning

- Bayesian Optimization
- Optuna

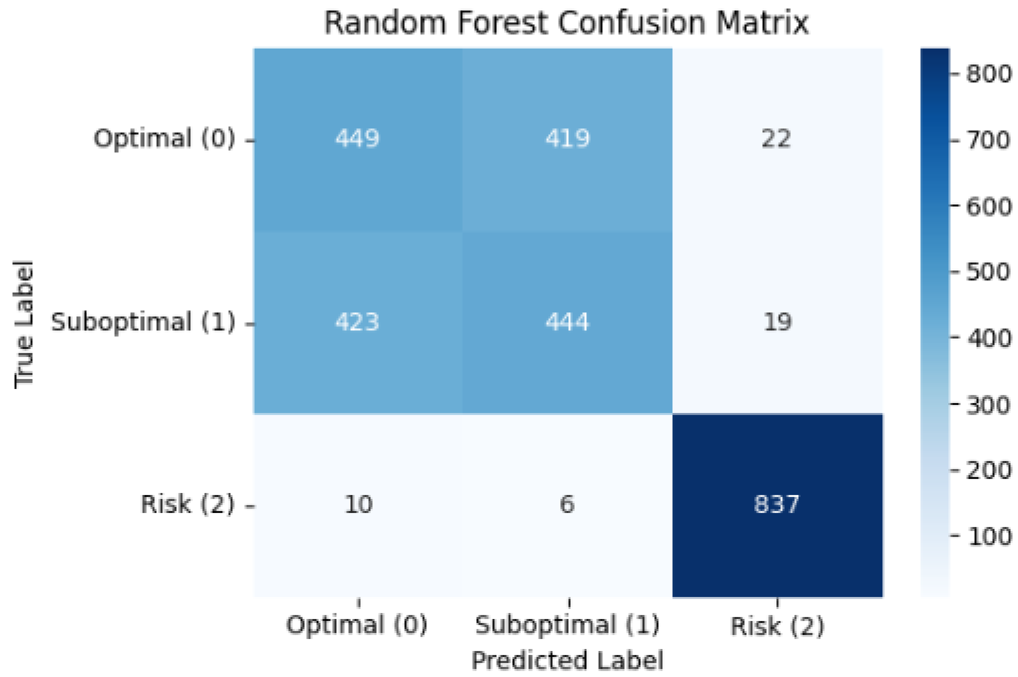
Performance Metrics Table

Model	Accuracy	Precision	Recall	F1 (macro)
Logistic Regression	0.64	0.64	0.65	0.64
Random Forest	0.66	0.66	0.66	0.66
XgBoost	0.6561	0.6533	0.6560	0.6546
ANN	0.6744	0.6759	0.6743	0.6741
LightGBM	0.55	0.57	0.61	0.58
MLP with feature interaction	0.6687	0.6669	0.6686	0.6571
Stacking (LR+MLP+XgBoost)	0.6695	0.6690	0.6693	0.6691
Hierarchical XgBoost	0.9182	0.9181	0.9183	0.9182
Hierarchical Tab Transformer	0.6529	0.6559	0.6596	0.6557
Hierarchical XgBoost with Bayesian Optimization	**0.9281**	**0.9281**	**0.9281**	**0.9280**
Hierarchical XgBoost with Optuna	0.9258	0.9258	0.9258	0.9258

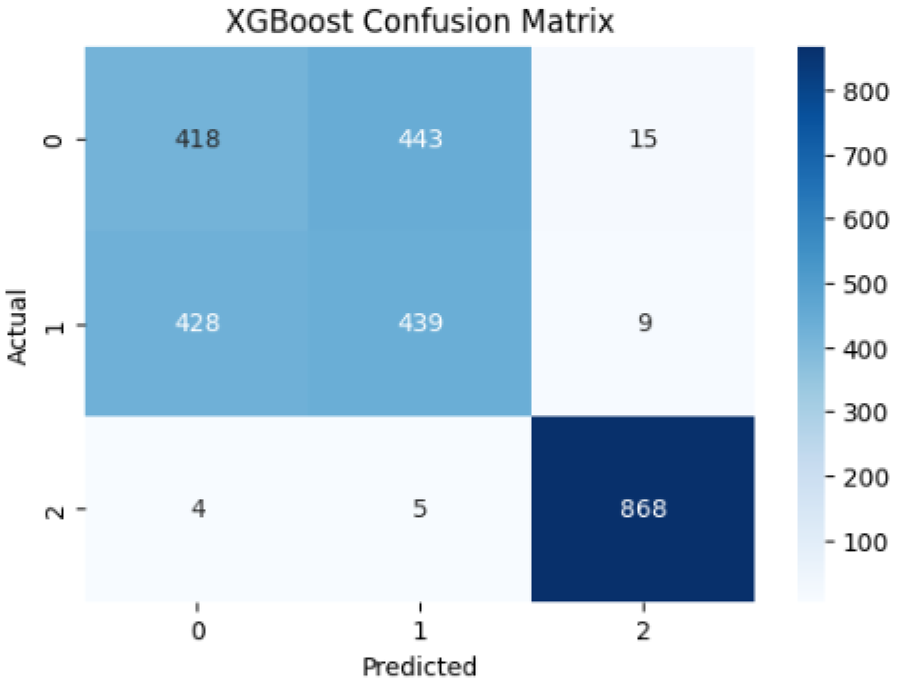
Confusion Matrix

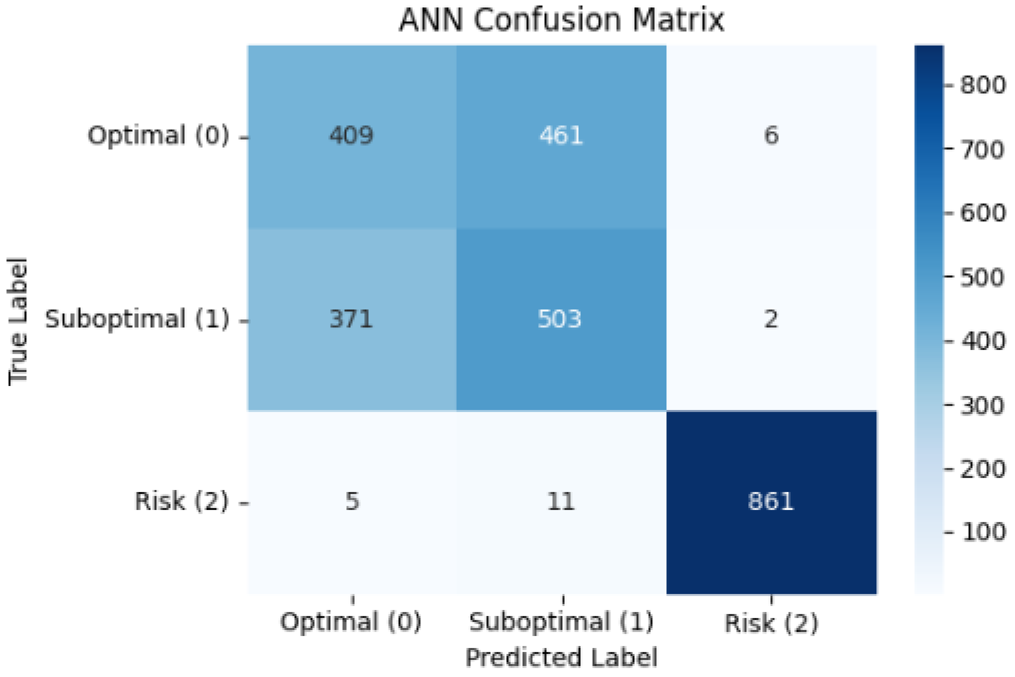

Logistic Regression Confusion Matrix

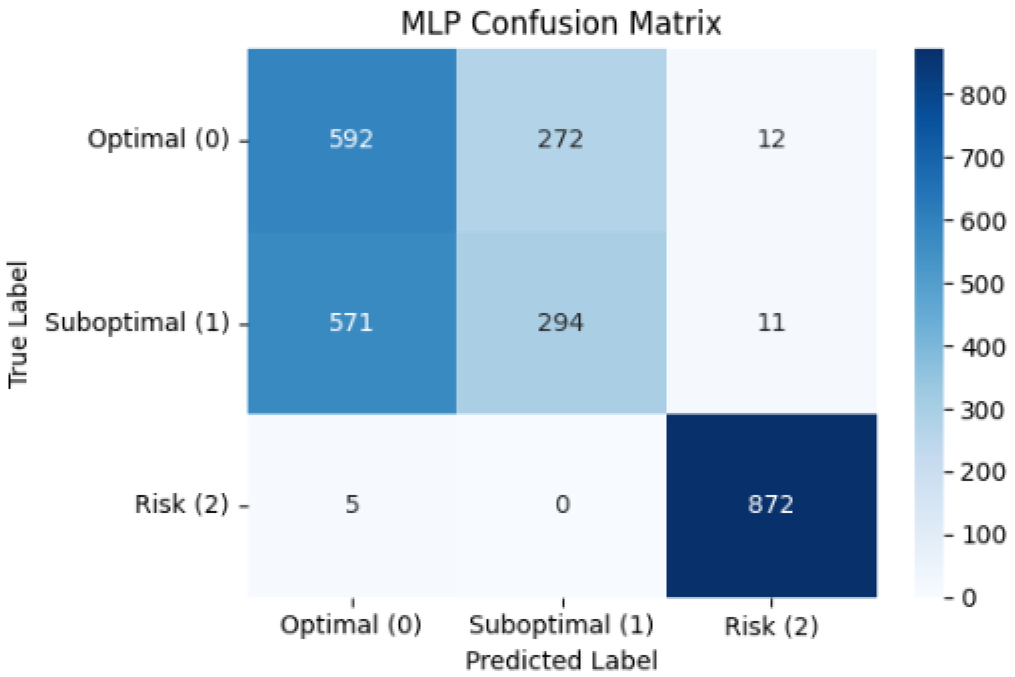
True Label \ Predicted Label	Optimal (0)	Suboptimal (1)	Risk (2)
Optimal (0)	446	408	36
Suboptimal (1)	428	420	38
Risk (2)	12	19	822

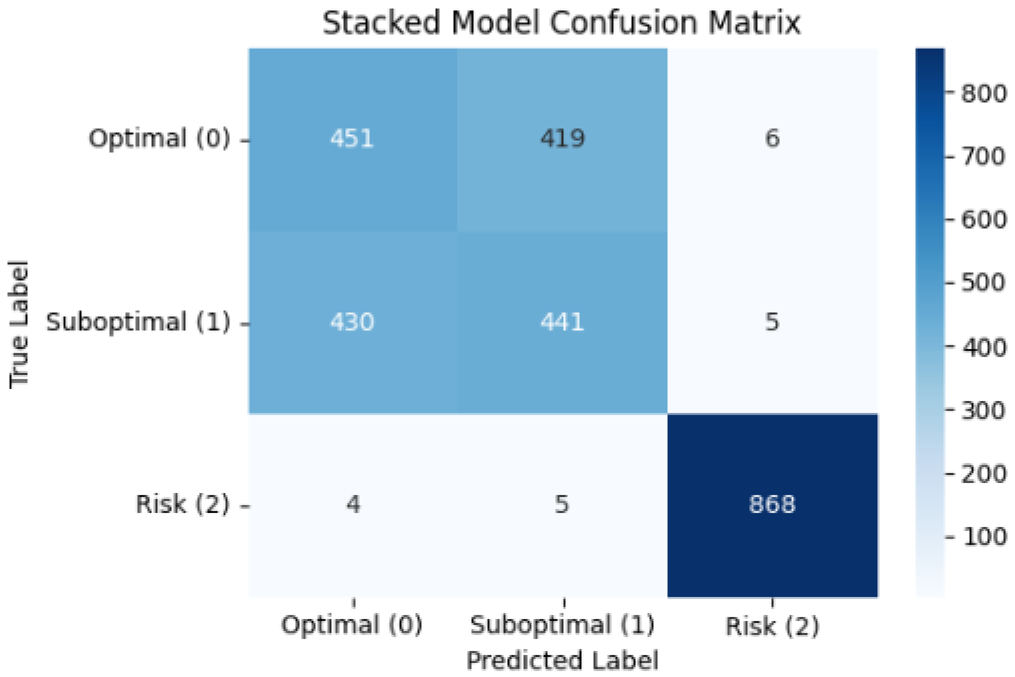

Random Forest Confusion Matrix

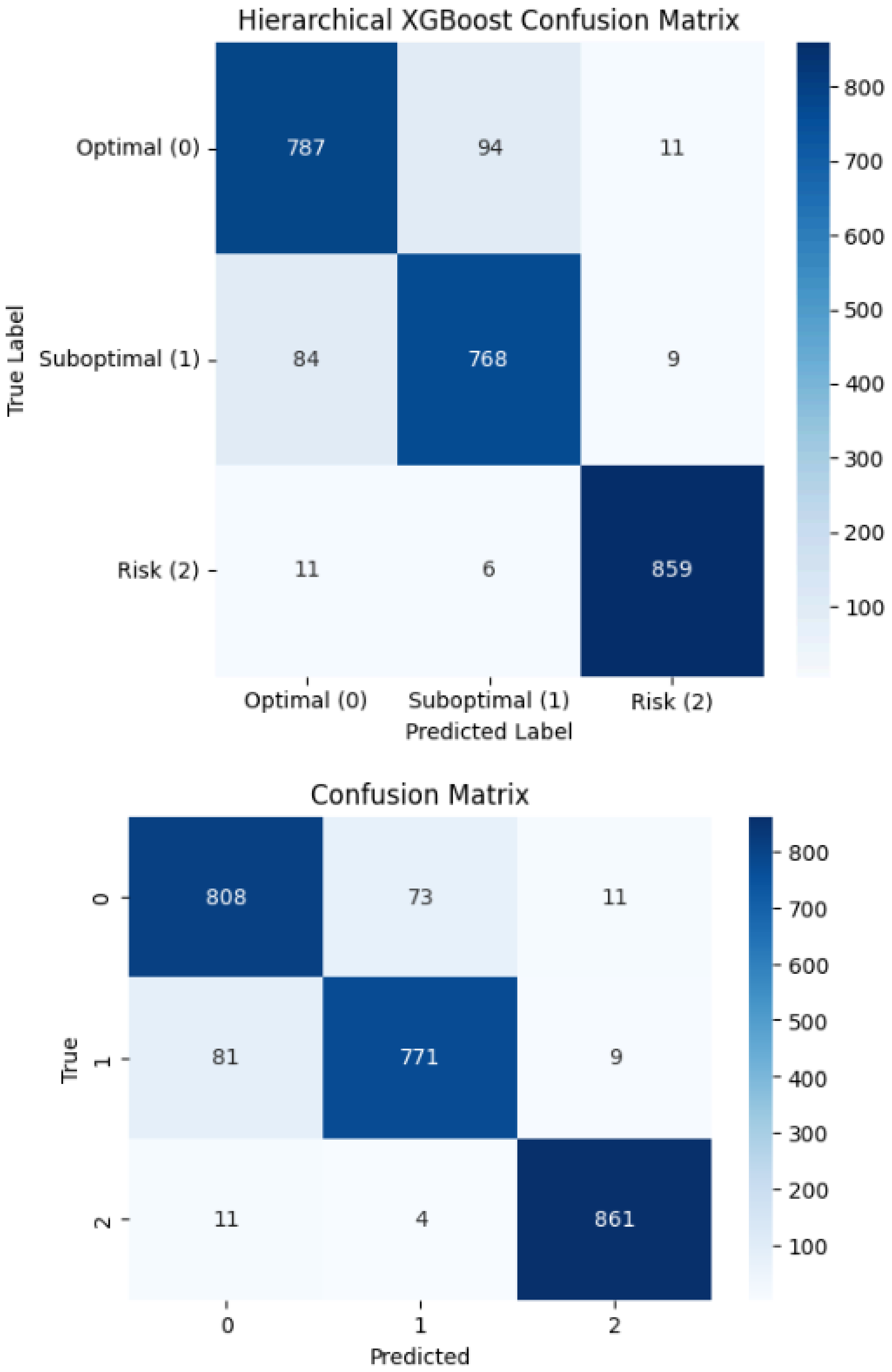
True Label \ Predicted Label	Optimal (0)	Suboptimal (1)	Risk (2)
Optimal (0)	449	419	22
Suboptimal (1)	423	444	19
Risk (2)	10	6	837


XGBoost Confusion Matrix
Actual
0
1
2
418
443
15
428
439
9
4
5
868
0
1
2
Predicted
800
700
600
500
400
300
200
100


ANN Confusion Matrix
True Label
Optimal (0)
Suboptimal (1)
Risk (2)
409
461
6
371
503
2
5
11
861
Optimal (0)
Suboptimal (1)
Risk (2)
Predicted Label
800
700
600
500
400
300
200
100


MLP Confusion Matrix
True Label
Optimal (0)
Suboptimal (1)
Risk (2)
592
272
12
571
294
11
5
0
872
Optimal (0)
Suboptimal (1)
Risk (2)
Predicted Label
800
700
600
500
400
300
200
100
0


Stacked Model Confusion Matrix
True Label
Optimal (0)
Suboptimal (1)
Risk (2)
451
419
6
430
441
5
4
5
868
Optimal (0)
Suboptimal (1)
Risk (2)
Predicted Label
800
700
600
500
400
300
200
100


Hierarchical XGBoost Confusion Matrix

Confusion Matrix

Fig: Confusion Matrix of Hierarchical XgBoost with Bayesian Optimization

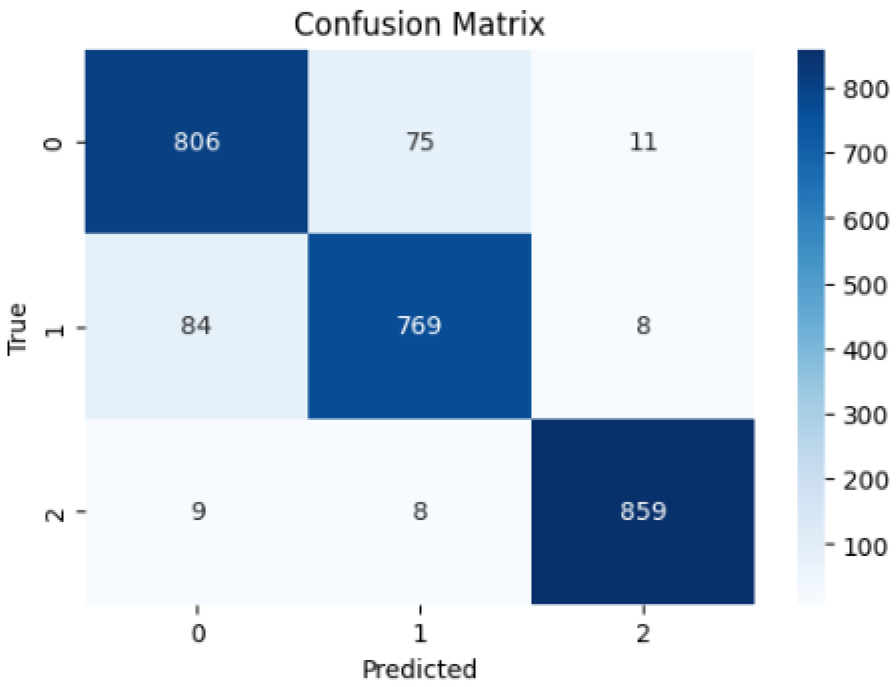
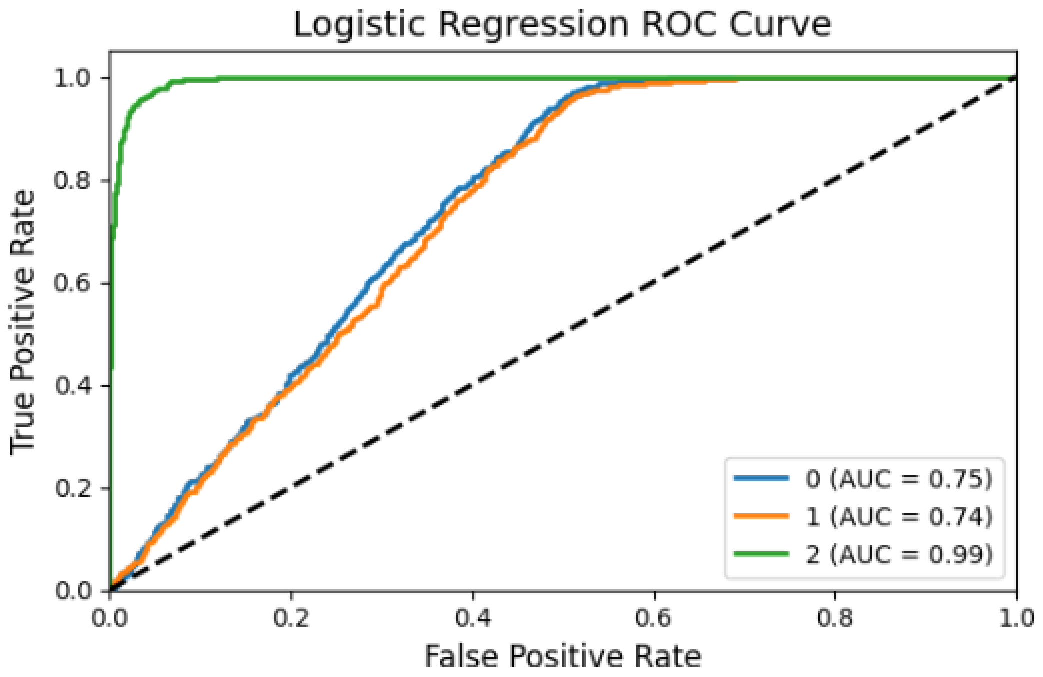
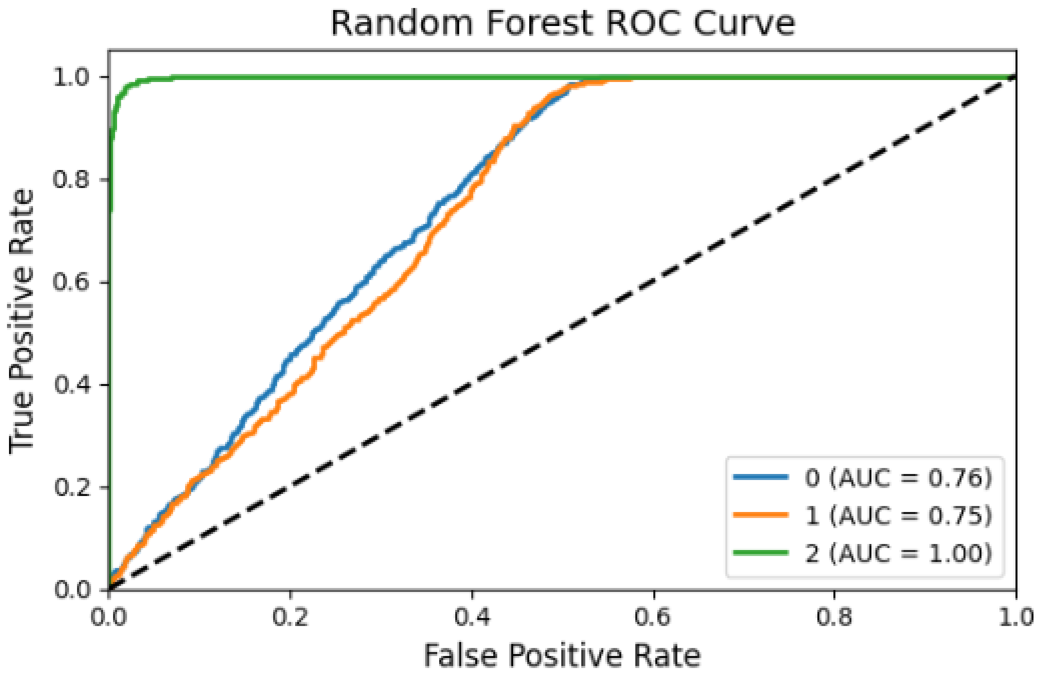
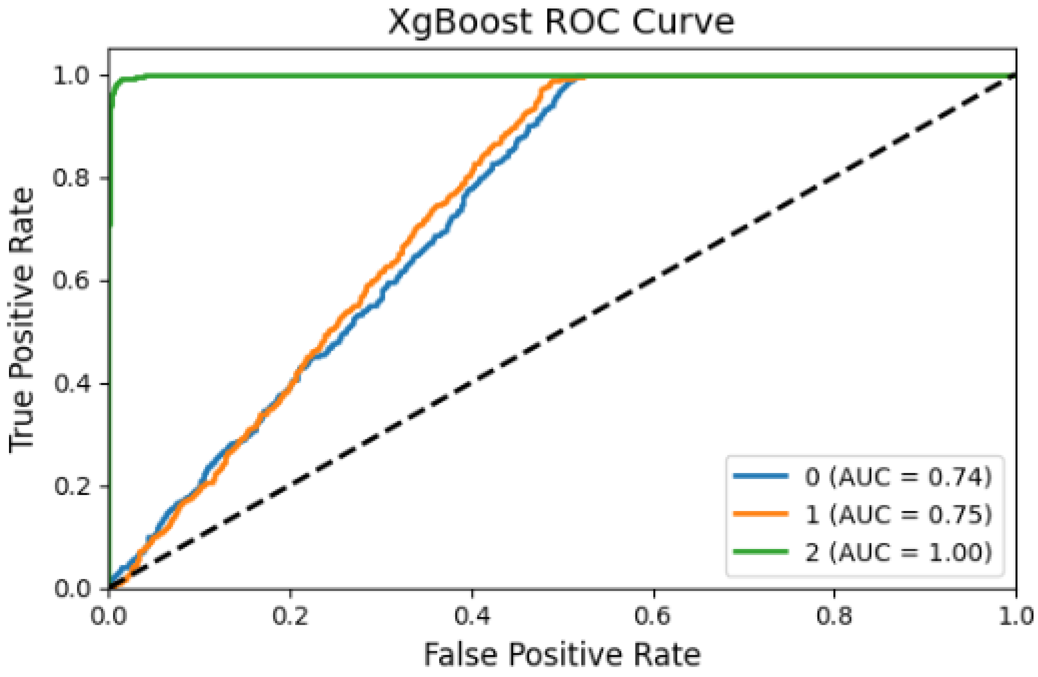


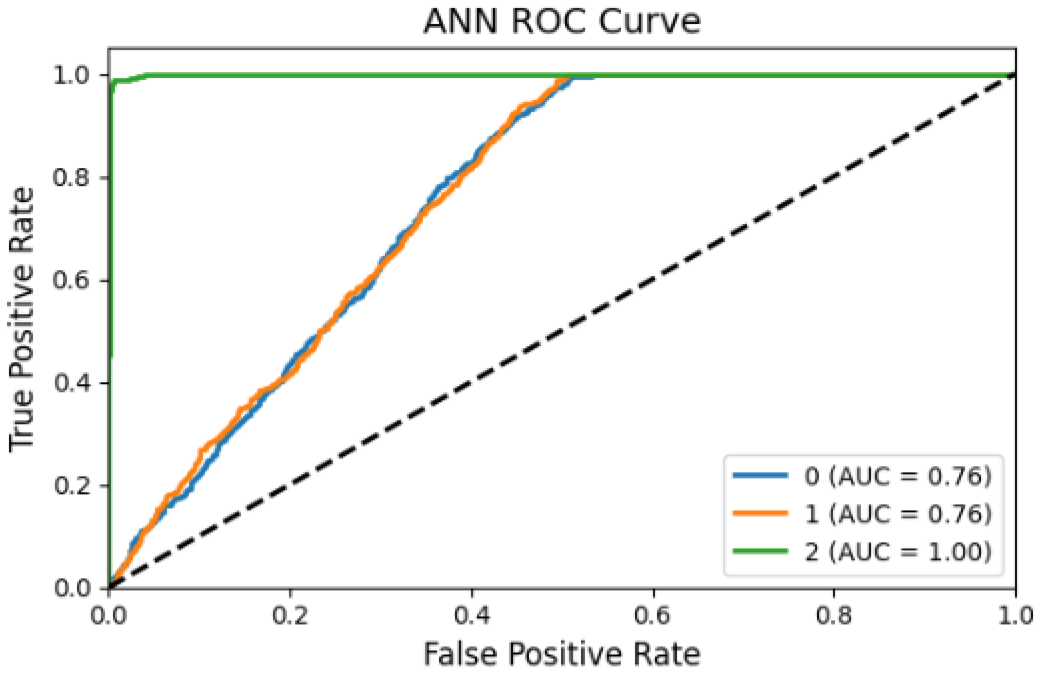
Fig: Confusion Matrix of Hierarchical XgBoost with Optuna

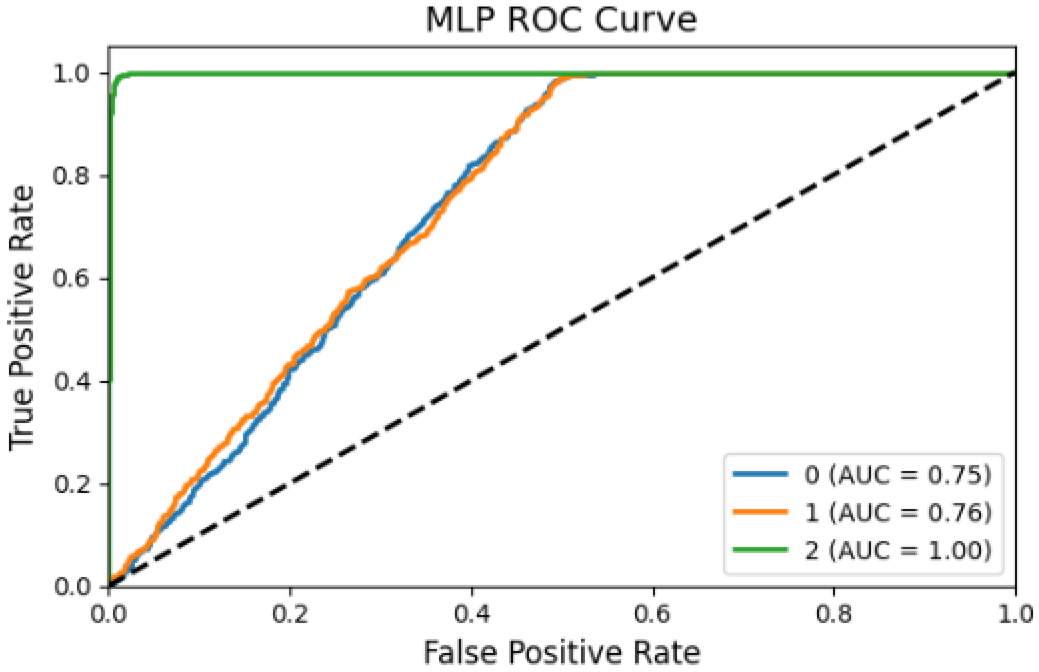
ROC Curve

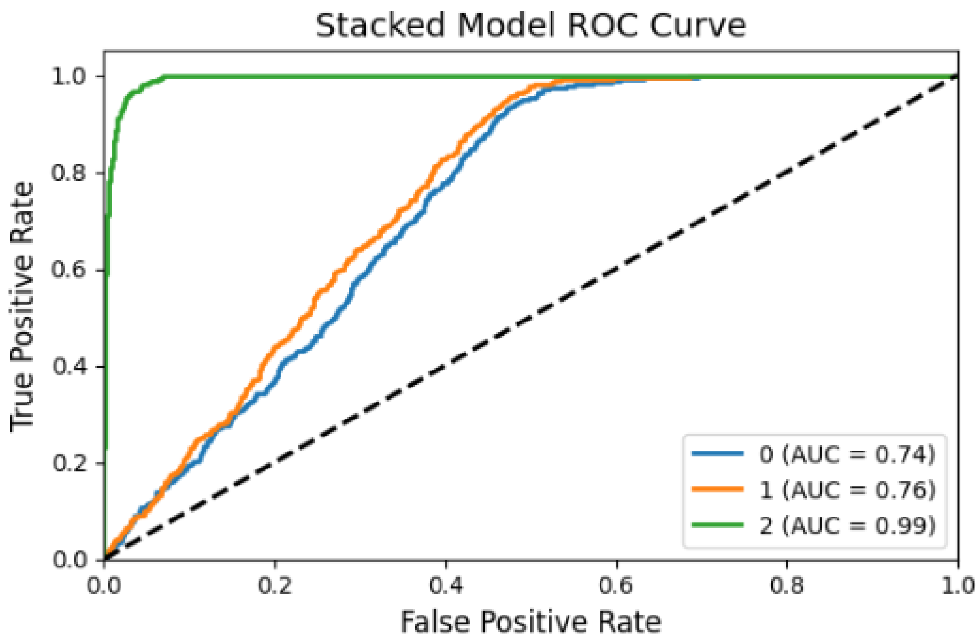


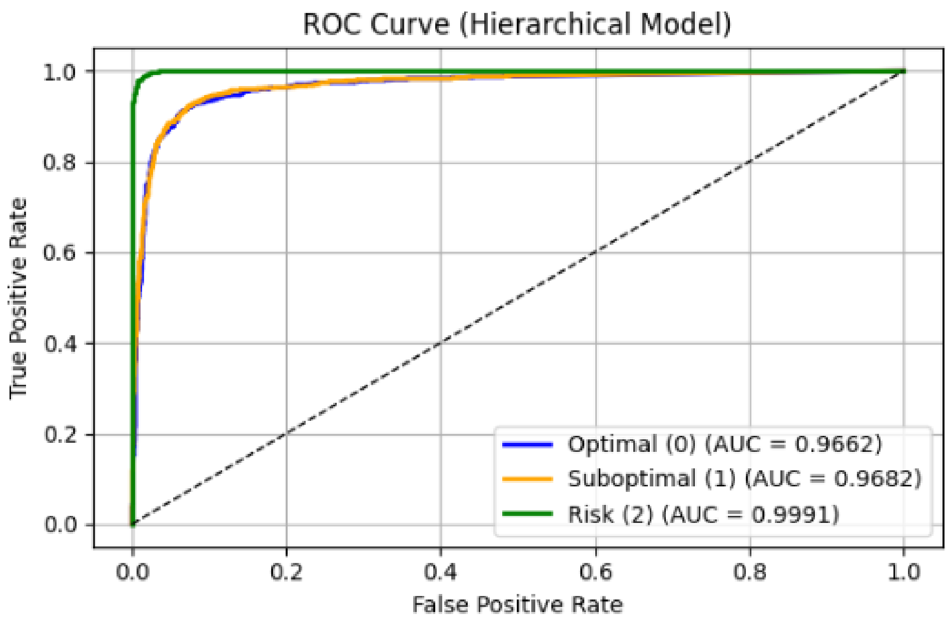

Random Forest ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0 (AUC = 0.76)
1 (AUC = 0.75)
2 (AUC = 1.00)


XgBoost ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0 (AUC = 0.74)
1 (AUC = 0.75)
2 (AUC = 1.00)


ANN ROC Curve


MLP ROC Curve





Summary of key findings

- Hierarchical classification performs better in this dataset.
- All the models struggled to distinguish between Optimal and Suboptimal
- Hyperparameter tuning can improve performance.
- Composite features, SHAP based features and target encoding played an important role to increase the performance.