H.E.P.A.R. (Hepatic Ensemble Predictive Analysis Resource): A Convergent Machine Learning Architecture for Drug-Induced Liver Injury Prediction

Laksh Agarwal, Neo Dong | Carmel High School

Q1: Problem & Objectives

- Companies **lose billions** when drugs don't pass Drug Induced Liver Injury(DILI) testing and computational methods utilize exascale systems which are inaccessible while **costing tens of millions**.
- Non-exascale systems cannot implement full chemical biological bi-modal approach.
- Recent studies by Dr. Waxman found success using stratified biological features in predicting sugar breakdown in the bloodstream.

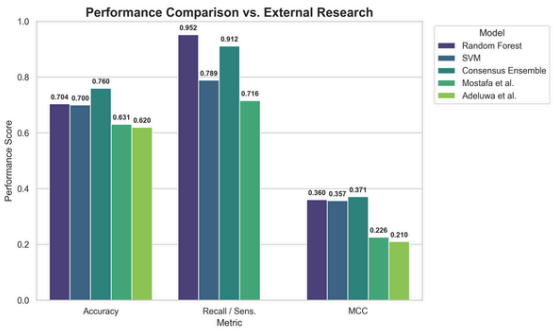
Can we implement a stratified bimodal approach to improve accuracy of accessible DILI models and reduce overall costs?

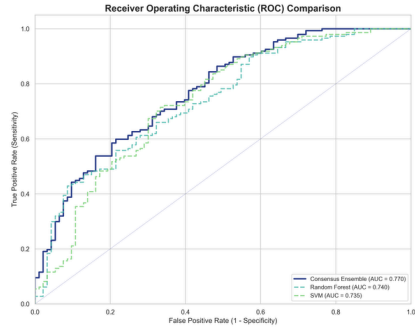
Q2: Project Design

- Derive chemical features using SMILES and biological features using stratification method adjusted for liver properties via genetic algorithms.
- Create a voting ensemble using heavy models (GaussianProcess, HistGradientBoosting, ExtraTrees, MLP, SVM) and light learners (LogisticRegression, SGDClassifier, KNN).
- Collect and validate results using 80-20 train test split and 5x2 cross validation with paired T-test.

Q3: Data Analysis & Results

- Ensemble voting improved accuracy, consistency, and AUC-AOC at slight cost of recall.
- HEPAR improved 13% in accuracy, 20% in recall, and 45% in consistency when compared to previous models.





Q4: Interpretations & Conclusions

- A stratified bi-modal approach **significantly increases accuracy** in DILI prediction
- Utilizing a voting ensemble with a slight heavy model skew **improves stability**.
- Genetic algorithms are sufficiently advanced for accurate metabolite formation prediction.
- **Next Steps**: Validate unexpected chemical features with lab testing. Conduct beta testing with real researchers. Check if improvements continue with super computers.