Heart Disease Risk Assessment: A Multi-Class Prediction Approach

Vi Thi Tuong Nguyen, Lam Nguyen, James Pham, Le Duy Vu

Problem Description

Globally, heart disease is the leading cause of death, responsible for about 32% of all deaths worldwide [1]. A key challenge is that current diagnostic methods rely on expensive tests that are often unaffordable or inaccessible for people with limited means, and difficult to implement in smaller hospitals and developing regions. As a result, many patients don't get early detection that could save their lives.

We are not simply trying to answer "does this person have heart disease?" We aim to predict the severity level, which helps doctors determine how urgently someone needs treatment. A patient with mild symptoms may only require lifestyle changes, whereas a more severe case might need immediate intervention. This streamlined and detailed risk assessment process could help people without access to healthcare, general practitioners without specialized equipment, emergency room doctors evaluating chest pain, and healthcare systems in resource-limited areas.

Aside from developing machine learning models, we plan to create a user-friendly web application to manifest our solution in action. Users will be able to input basic clinical information, such as age, gender, blood pressure, cholesterol levels, etc., through a simple interface, and our trained model will provide an assessment of heart disease risk and severity level. This web app will demonstrate the practical impact of our research and illustrate how machine learning could be integrated into real clinical workflows.

Dataset

We're using the **UCI Heart Disease Dataset** from Kaggle: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

The dataset contains 920 patient records from four medical centers: Cleveland Clinic, Hungarian Institute of Cardiology, University Hospital in Switzerland, and V A Long Beach Medical Center. It has 14 key clinical features including age, gender, chest pain type, blood pressure, cholesterol levels, ECG results, and exercise stress test outcomes. The target variable ranges from 0-4 representing increasing disease severity: 0 means no significant disease (less than 50% artery blockage) while 1-4 indicate progressively worse conditions.

For preprocessing, we'll handle missing values through careful imputation (analyzing whether they're random or systematic), encode categorical variables like chest pain types into numbers the algorithm can understand, and address the expected class imbalance where most patients likely have no disease or mild disease compared to severe cases.

Proposed Solution

We're focusing mainly on building machine learning models that predict the exact severity level (0 through 4) to give doctors the most detailed and actionable information. Rather than just telling a doctor whether or not a patient has heart disease, our model will determine whether it's no disease, mild, moderate, severe, or very severe, which directly impacts treatment decisions and resource allocation. Our approach will compare several algorithms:

1. **Random Forest** handles mixed data types well and provides feature importance rankings that doctors can interpret.
2. **XGBoost** typically performs excellently on tabular medical data and handles missing values naturally [2].
3. **Support Vector Machines** are good for smaller datasets like ours. We'll also experiment with ensemble methods that combine predictions from multiple models.

Our approach should work since the clinical features we have are the ones doctors use to assess heart disease risk, so it makes sense that an algorithm could learn these patterns as well. Previous studies on this dataset have achieved 80-85% accuracy [3][4], which suggests there are genuine patterns in the data that algorithms can learn. We'll use cross-validation to ensure each fold maintains the same severity level proportions, and tune hyperparameters through grid search.

Expected Challenges

The biggest challenge will be **class imbalance**. We expect most patients to have no disease or mild disease, with fewer severe cases. This could make our model great at predicting the common cases but terrible at catching the severe ones, exactly the opposite of what we want clinically. We're planning to try techniques like SMOTE (which creates synthetic examples of rare classes) and cost-sensitive learning (which penalizes mistakes on severe cases more heavily).

Dataset size is another concern. With only 920 patients, we need to avoid overfitting complex models. We'll use regularization techniques and careful cross-validation to make sure our training and test sets maintain the same proportions of severity levels.

Missing values in medical data are often not random. Certain tests might be missing because patients were too sick or doctors didn't think they were necessary. We'll analyze the patterns first (are they random or systematic?) then choose appropriate imputation methods. Specifically, we plan to evaluate mean/median imputation for numerical features, mode imputation for categorical variables, and KNN imputation for more sophisticated missing value estimation when patterns suggest the missingness is not completely random.

Evaluation Plan

First, for the **performance metric**, we'll use weighted F1-score as our primary metric since it balances precision and recall while accounting for class imbalance. Regular accuracy can be misleading with unbalanced classes.

Second, the **success measurement** will be at least 75% weighted F1-score overall. But more importantly, we want high sensitivity (>90%) for detecting any level of disease. Missing a sick patient is much worse than worrying a healthy one.

Finally, the **detailed analysis** will include creating confusion matrices to see per-class performance, calculating precision and recall for each severity level, and comparing our results against published benchmarks on this dataset. We'll implement stratified k-fold cross-validation (k=5) to enable robust model evaluation while maintaining class distribution across folds. We'll also analyze which clinical features are most important for predictions to ensure our model makes medical sense.

Team Roles

Our team will divide lead responsibilities as follows, although every member will contribute to all aspects of the project:

- **Lam Nguyen** - *Data preprocessing and analysis lead* - data cleaning, missing value imputation, exploratory data analysis, and feature engineering
- **James Pham** - *Model development and training lead* - training and implementing machine learning algorithms (Random Forest, XGBoost, SVM), hyperparameter tuning, and ensemble methods
- **Le Duy Vu** - *Web application development lead* - designing and implementing the user-friendly web interface, integrating trained models for real-time predictions, and UX testing
- **Vi Thi Tuong Nguyen** - *Evaluation and documentation lead* - model evaluation, performance analysis, creating visualizations and reports, and writing documentation

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," *WHO Fact Sheets*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. Y. Yıldız and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data," *arXiv preprint arXiv:2410.03705*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.03705

[3] K. M. Alfadli and A. O. Almagrabi, "Feature-Limited Prediction on the UCI Heart Disease Dataset," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 5871-5883, 2023. [Online]. Available: https://doi.org/10.32604/cmc.2023.033603

[4] S. Anderies et al., "Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms," *Engineering, Mathematics and Computer Science (EMACS)*, vol. 4, pp. 87-93, 2022. [Online]. Available: https://doi.org/10.21512/emacsjournal.v4i3.8683