Group X Progress Report:
Heart Disease Prediction using Machine Learning

ZhiChong Lin, ZiDi Yao, Ke Ma
yaoz25@mcmaster.ca, mak11@mcmaster.ca, linz8@mcmaster.ca

1 Introduction

This section introduces the problem and motivation of your project. You may adapt the motivation from your original proposal. Typical content includes: (1) What problem you are solving, (2) Why it matters, (3) Why machine learning is suitable, (4) Your project objective. This should be about 0.25–0.5 pages.

2 Related Work

This section summarizes the most relevant previous work. If no identical problem exists, describe the most similar tasks such as: – Medical risk prediction – Heart disease datasets – Classic ML models like logistic regression / SVM in healthcare Cite at least five references (use custom.bib). Length: 0.25–0.5 pages.

3 Dataset and Preprocessing

Describe the dataset, number of samples, features, data source, and what preprocessing was required.

3.1 Dataset Description

Present the raw features in a table:

Feature	**Description**
Age	Age of patient (years)
Sex	Biological sex (M/F)
ChestPainType	Chest pain type (ATA, ASY, NAP, TA)
RestingBP	Resting blood pressure (mm Hg)
Cholesterol	Serum cholesterol (mg/dL)
FastingBS	Fasting blood sugar (0/1)
RestingECG	ECG results (Normal, ST, LVH)
MaxHR	Maximum heart rate achieved
ExerciseAngina	Exercise-induced angina (Y/N)
Oldpeak	ST depression value
ST_Slope	Slope of ST segment (Up/Flat/Down)
HeartDisease	Target label (1 = disease, 0 = healthy)

Table 1: Raw dataset features.

3.2 Target Extraction

Describe how `HeartDisease` was extracted as the binary label vector.

3.3 Feature Preprocessing

Explain your preprocessing:

- Removing whitespace in column names
- Encoding binary variables (Sex, ExerciseAngina, FastingBS)
- One-hot encoding for multi-class categorical features (ChestPainType, RestingECG, ST_Slope)
- Saved processed data to: `processed/X_encoded.csv`

3.4 Final Processed Dataset

State final shape (e.g., 918 rows × 18 columns) and where it is stored.

4 Model Inputs (Features)

Describe the input representation to the model, e.g.:

- The 18-dimensional processed feature vector
- Whether any normalization was applied
- Whether feature engineering or selection was performed

This section corresponds to Item 3 in the project instructions.

5 Model Implementation

Describe the machine learning model(s) used.
Examples:

- RBF-kernel Support Vector Machine (SVM)
- Justification for using SVM
- Hyperparameters used (C, gamma)
- Training pipeline and libraries (scikit-learn)

This section corresponds to Item 4 of the project instructions.

6 Evaluation Strategy and Results

6.1 Evaluation Method

Explain why you used stratified K-fold cross validation (e.g., small dataset size, need for robust evaluation).

6.2 Metrics

Explain why accuracy, precision, recall, F1, and AUC-ROC are important in medical diagnosis.

6.3 Results

Insert your figures:

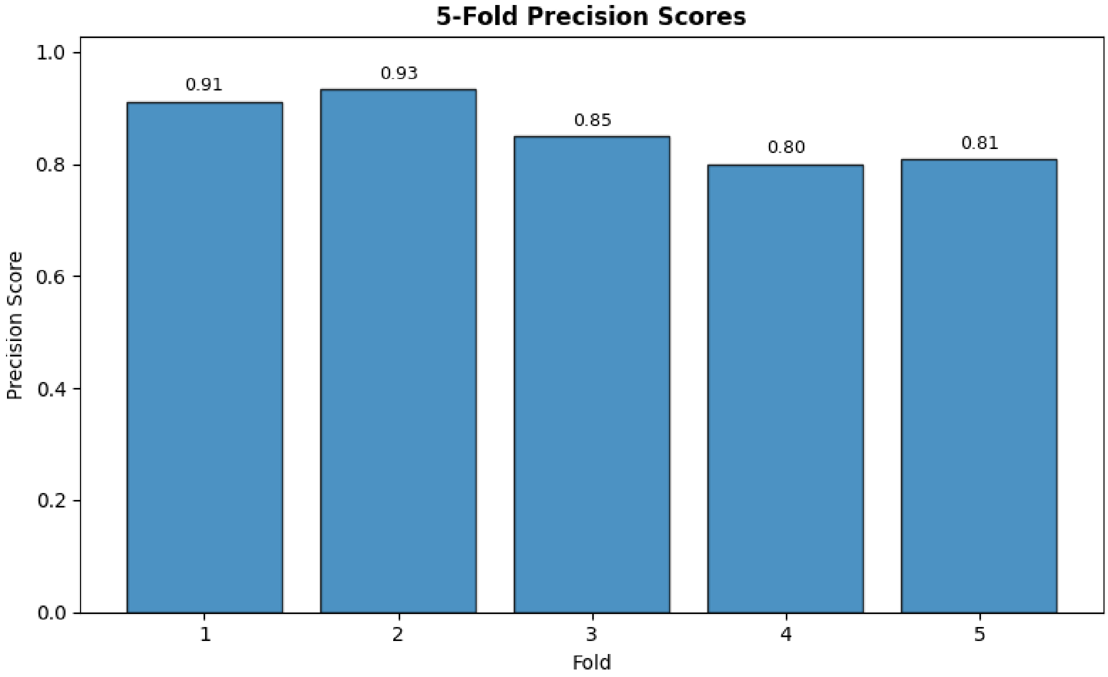
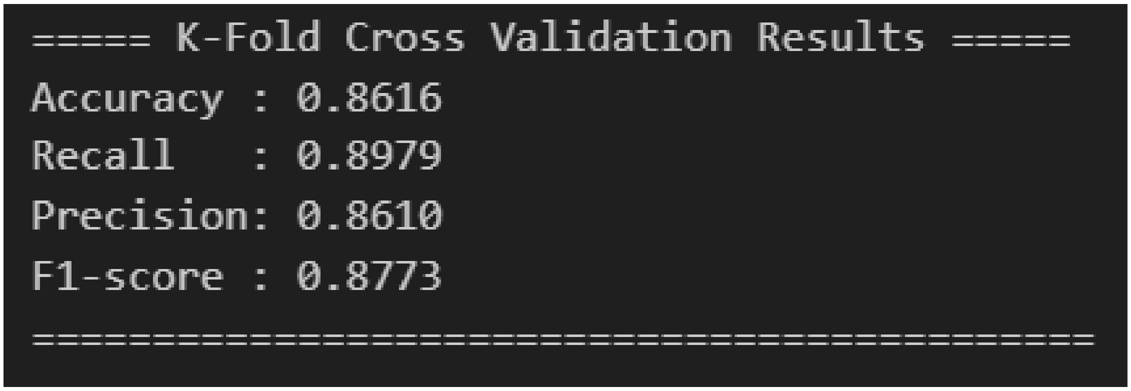


Figure 1: 5-fold precision scores.



```
===== K-Fold Cross Validation Results =====
Accuracy : 0.8616
Recall   : 0.8979
Precision: 0.8610
F1-score : 0.8773
============================================
```

Figure 2: Summary metrics across folds.

7 Feedback and Future Plans

Summarize TA feedback and your improvements:

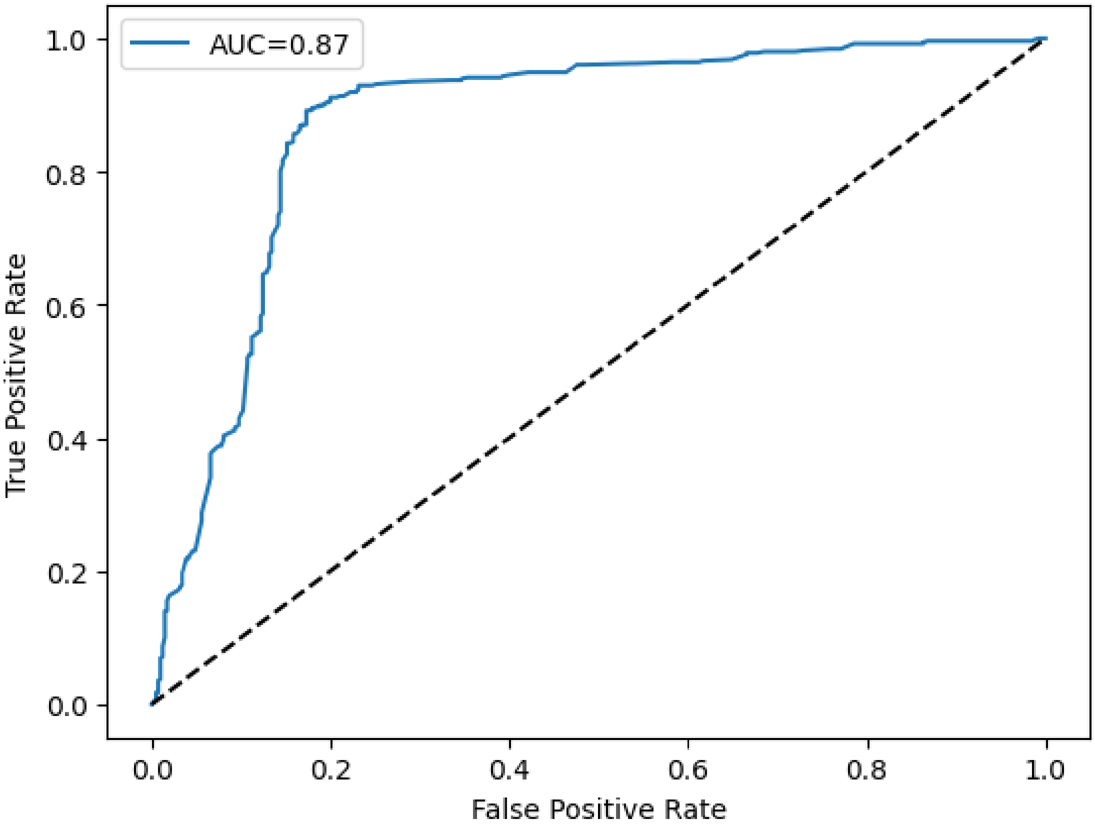


Figure 3: AUC–ROC curves.

- Replace label encoding with one-hot encoding
- Consider switching from SVM to neural networks for performance gains
- Create a new GitHub branch for experiments

Team Contributions

Describe what each team member worked on.

References

[Ando and Zhang(2005)] Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

[Andrew and Gao(2007)] Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

[Gusfield(1997)] Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

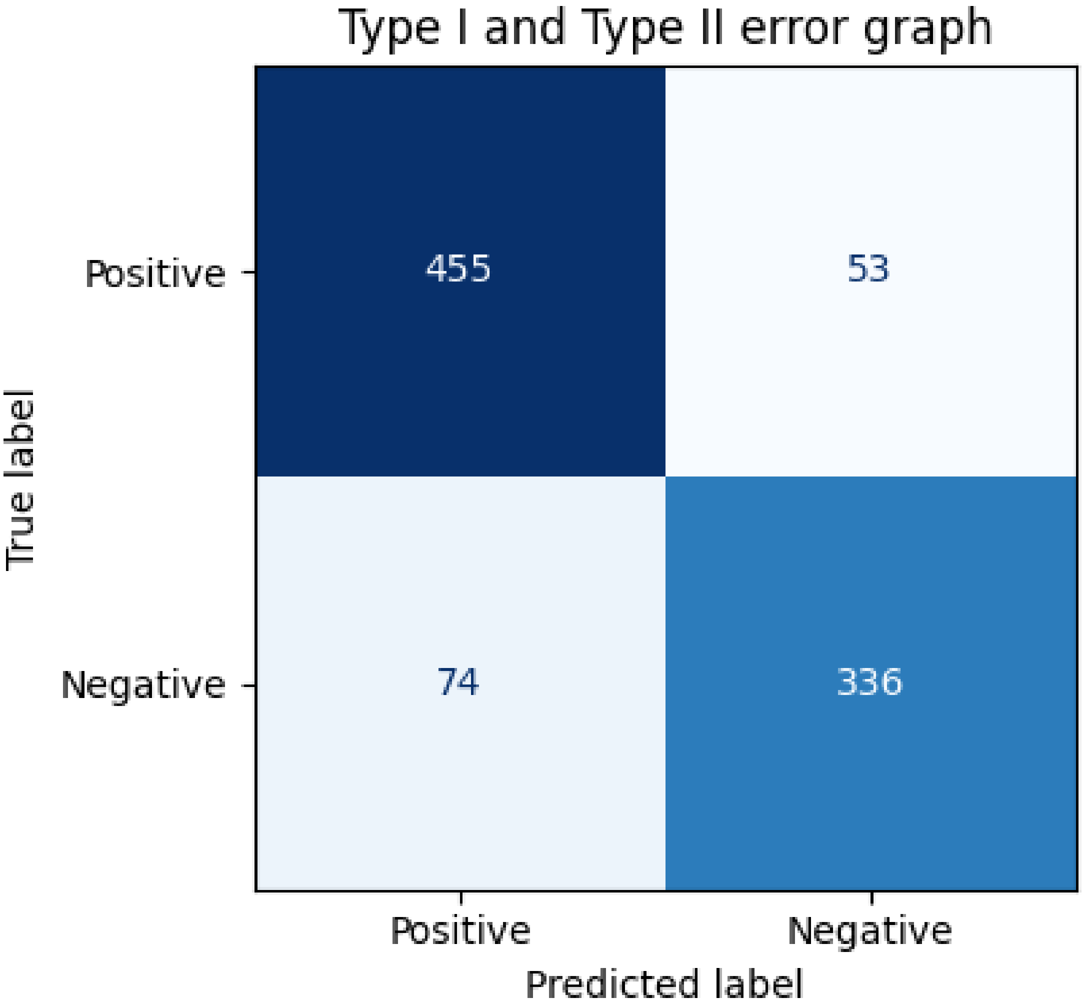


Figure 4: Confusion matrices.

[Rasooli and Tetreault(2015)] Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. http://arxiv.org/abs/1503.06733 Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.