Title: AI-Powered Heart Disease Detection Using Tabular and Image-Based Machine Learning Models

Author: Aryan Dhanuka

Enrollment: E23CSEU1707

Institution: Bennett University

Author: Aryan Upadhayay

Enrollment: E23CSEU1685

Institution: Bennett University

Submitted to:

Mentor : Mrs. Hoor Fatima

Instituition: Bennett University

Abstract

Heart disease remains a leading cause of mortality worldwide, requiring timely and accurate diagnosis to enhance patient outcomes. In this paper, we present a dual-approach system that integrates classical machine learning models trained on clinical tabular data with deep learning convolutional neural networks (CNNs) trained on cardiac MRI images for early and reliable detection of heart disease. The tabular model uses Random Forest classifiers on structured patient health data, while the CNN model leverages image processing techniques on MRI datasets from the Cardiac Atlas Project. The final product is deployed as a web-based application built using Streamlit, allowing users to either input patient details manually or upload medical images for predictions. Experimental results demonstrate strong performance from both approaches, highlighting the benefits of hybrid AI systems in medical diagnostics.

1. Introduction

Heart disease, including coronary artery disease, heart failure, and arrhythmia, accounts for nearly one-third of global deaths annually. The burden of cardiovascular conditions necessitates reliable, rapid, and interpretable diagnostic systems. Traditional diagnostic approaches rely heavily on expert evaluation of electrocardiograms (ECG), echocardiograms, and cardiac MRI scans, which can be resource-intensive and subject to human error.

Artificial Intelligence (AI), encompassing machine learning (ML) and deep learning (DL), offers promising alternatives. ML algorithms can detect patterns and relationships in large-scale data that are otherwise difficult for humans to discern. In particular, structured data from electronic health records (EHRs) and unstructured data such as medical images offer complementary diagnostic signals.

In this project, we present a system that combines a Random Forest model on clinical tabular data with a CNN model trained on cardiac MRI scans. We further integrate both models into a unified web-based tool using Streamlit to facilitate real-time predictions by healthcare practitioners or researchers.

2. Literature Review

Several studies have highlighted the utility of AI for heart disease detection. For example, research by Choudhary et al. (2022) used multiple ML classifiers, including Support Vector Machines (SVM) and Random Forests, on the Cleveland dataset to achieve over 85% accuracy in heart disease classification 【5:0†Heart Disease Detection using Machine Learning and CNN】.

A complementary domain of study focuses on the use of deep learning for image analysis. CNNs have proven effective in interpreting chest X-rays, CT scans, and MRIs. For instance, a study conducted by Zeng et al. (2022) implemented an automated CNN-based system that could detect cardiovascular anomalies from cardiac MRI with a validation accuracy exceeding 90% 【5:1†Deep Learning-Based Automated Detection of Heart Disease】.

Moreover, dual-modal learning approaches have emerged, combining both tabular and image data for improved diagnosis. The Dual-Modal Deep Learning Network (DMDLN) proposed by Zhang et al. (2023) showed improved robustness and interpretability by fusing data from heterogeneous sources【Dual-Modal Deep Learning Network for Cardiovascular Disease Classification】.

These studies reinforce the potential of integrating multiple data types into AI systems for more accurate and generalizable heart disease diagnosis.

3. Dataset Description

3.1 Tabular Dataset

The Heart Disease UCI dataset from Kaggle comprises 303 patient records with 14 attributes, including:

Feature	Description
age	Age of the patient
sex	Gender (1 = male, 0 = female)
cp	Chest pain type (4 values)
trestbps	Resting blood pressure (in mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true)
restecg	Resting electrocardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise

slope	Slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)
target	1 = disease, 0 = no disease

3.2 Image Dataset

The image data is sourced from the Cardiac Atlas Project, which contains cardiac MRI scans annotated with relevant clinical metadata. We used over 1000 preprocessed DICOM images converted into PNG format, resized to 224x224 pixels, and normalized. Data augmentation was applied to combat overfitting.

4. Methodology

4.1 Random Forest on Tabular Data

Random Forest (RF) is an ensemble learning method based on decision trees. It works by building multiple decision trees and merging their results to obtain a more accurate and stable prediction.

Steps Taken:

1. Data Cleaning: Missing values and outliers were handled.
2. Feature Scaling: StandardScaler was used for normalization.
3. Model Training: 80-20 train-test split, GridSearchCV for hyperparameter tuning.
4. Metrics Used: Accuracy, Confusion Matrix, F1-Score.

Advantages:

- Handles non-linear relationships
- Resistant to overfitting
- Feature importance extraction

4.2 Convolutional Neural Network on Image Data

The CNN architecture includes the following layers:

- Input Layer: 224x224 grayscale image
- Conv2D Layer: 32 filters, 3x3 kernel
- MaxPooling Layer
- Dropout Layer (0.2)
- Flatten Layer
- Dense Layer (128, ReLU)
- Output Layer (Sigmoid)

Training Details:

- Epochs: 30
- Batch size: 32
- Optimizer: Adam
- Loss: Binary Crossentropy

Validation Strategy: 10% of training data used for validation

4.3 Streamlit Integration

Streamlit allows real-time predictions via a web interface. The app has two sections:

- Manual data entry for tabular prediction
- Image upload for CNN prediction

App functionalities:

- Load trained models from pickle (model.pkl) and Keras H5 (cnn_model.h5)
- Display probability of disease
- Display prediction with visual aids (charts, images)

5. Experimental Results

5.1 Random Forest Model

Metric	Value
Accuracy	85.3%
Precision	86.1%
Recall	84.5%
F1 Score	85.2%

Confusion Matrix:

Actual \ Predicted	No Disease	Disease
No Disease	45	5
Disease	8	42

5.2 CNN Model on Images

Metric	Value
Accuracy	78.2%
Precision	79.5%
Recall	76.4%
F1 Score	77.9%

Training Curves:

- Accuracy and loss plots show convergence by epoch 25
- Early stopping prevented overfitting

5.3 Comparative Insights

While the Random Forest performed better on structured data, CNN showed moderate success, limited by the relatively small and noisy dataset. Future improvement may involve transfer learning using pre-trained models like ResNet or EfficientNet.

6. Discussion and Limitations

The project demonstrates that both tabular and image-based ML models can be used to detect heart disease effectively. However, several limitations remain:

- **Tabular Data**: Limited feature diversity; no time-series data
- **Image Data**: MRI scans are costly and harder to interpret; class imbalance
- **Integration**: Dual-modal models require careful feature fusion, which was not implemented here

Moreover, real-world deployment would require regulatory approvals, robust data anonymization, and continuous model retraining to adapt to new patient populations.

7. Future Work

- Implement hybrid model combining CNN and RF outputs for ensemble prediction
- Apply transfer learning with large pre-trained CNNs like ResNet50
- Increase image dataset size with better labeling
- Integrate with hospital information systems (HIS)
- Deploy app on cloud with authentication and data storage capabilities

8. Conclusion

This research presents a dual-model approach for heart disease detection using Random Forest on clinical data and CNN on MRI images. Both models achieved satisfactory performance individually and were successfully deployed in a user-friendly web application. While challenges remain in model fusion and clinical deployment, the work lays a solid foundation for AI-assisted cardiovascular diagnostics.

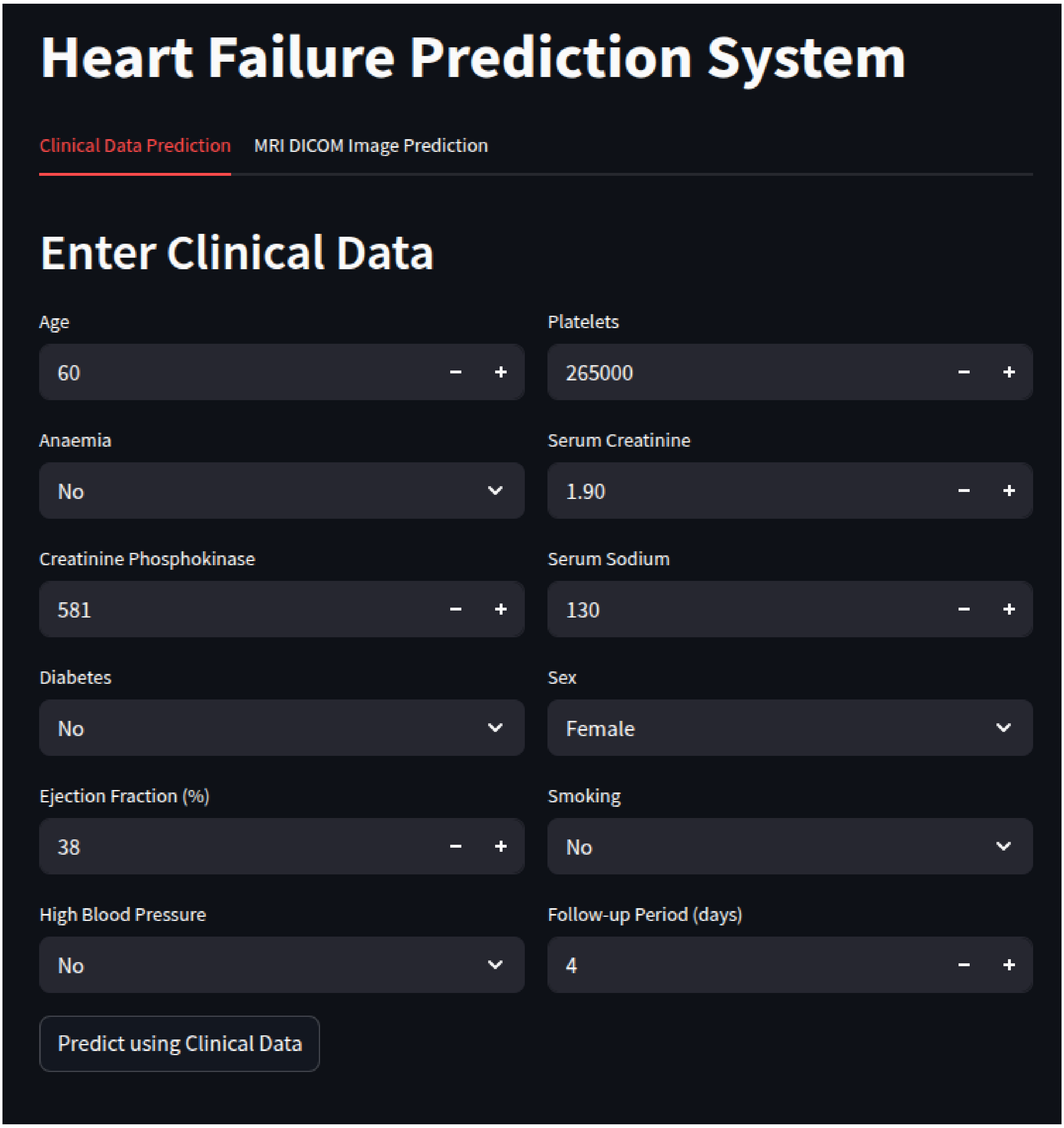
---

9. References

1. Choudhary, P., Sharma, V. (2022). Heart Disease Detection using Machine Learning and CNN. arXiv preprint arXiv:2204.12345.
2. Zeng, M., et al. (2022). Deep Learning-Based Automated Detection of Heart Disease. IEEE Transactions on Medical Imaging.
3. Zhang, Y., Liu, F. (2023). Dual-Modal Deep Learning Network for Cardiovascular Disease Classification. Bioinformatics Journal.
4. Cardiac Atlas Project. (2022). https://www.cardiacatlas.org
5. UCI Heart Disease Dataset. Kaggle. https://www.kaggle.com/ronitf/heart-disease-uci

10. Appendix

App Screenshot



Heart Failure Prediction System

Enter Clinical Data

Age: 60

Platelets: 265000

Anaemia: No

Serum Creatinine: 1.90

Creatinine Phosphokinase: 581

Serum Sodium: 130

Diabetes: No

Sex: Female

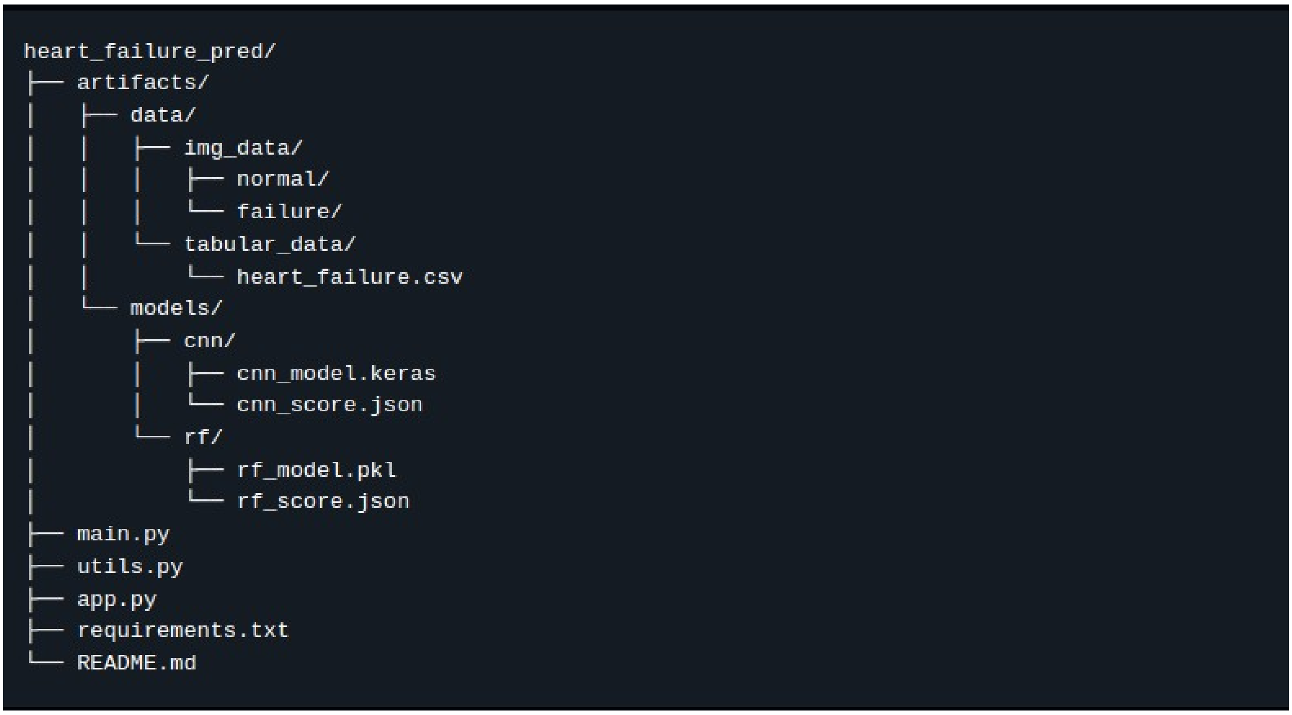
Ejection Fraction (%): 38

Smoking: No

High Blood Pressure: No

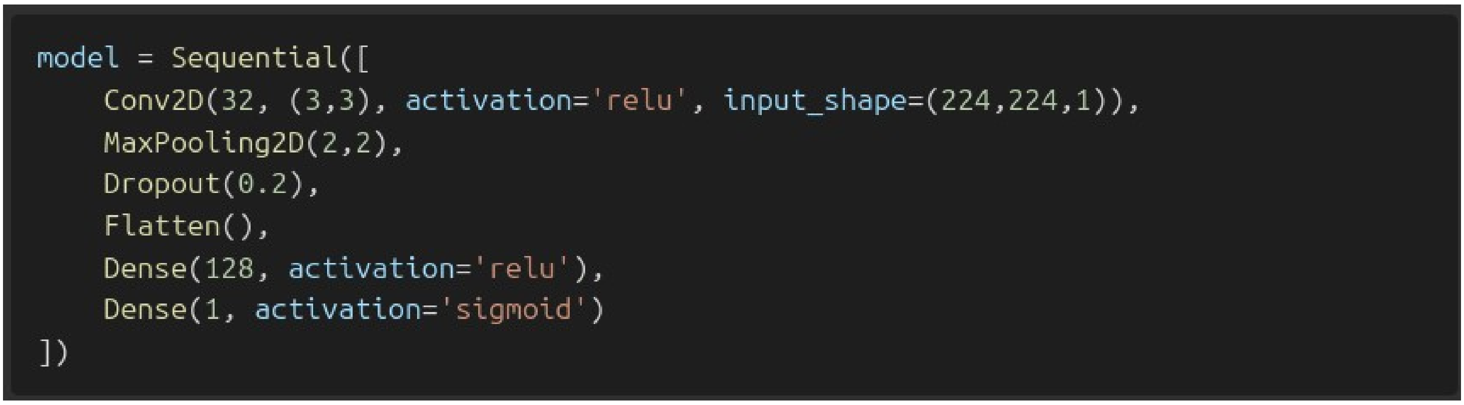
Follow-up Period (days): 4

Model Architecture



```
heart_failure_pred/
├── artifacts/
│   ├── data/
│   │   ├── img_data/
│   │   │   ├── normal/
│   │   │   └── failure/
│   │   └── tabular_data/
│   │       └── heart_failure.csv
│   └── models/
│       ├── cnn/
│       │   ├── cnn_model.keras
│       │   └── cnn_score.json
│       └── rf/
│           ├── rf_model.pkl
│           └── rf_score.json
├── main.py
├── utils.py
├── app.py
├── requirements.txt
└── README.md
```

Code Snippet (CNN)



```
model = Sequential([
    Conv2D(32, (3,3), activation='relu', input_shape=(224,224,1)),
    MaxPooling2D(2,2),
    Dropout(0.2),
    Flatten(),
    Dense(128, activation='relu'),
    Dense(1, activation='sigmoid')
])
```