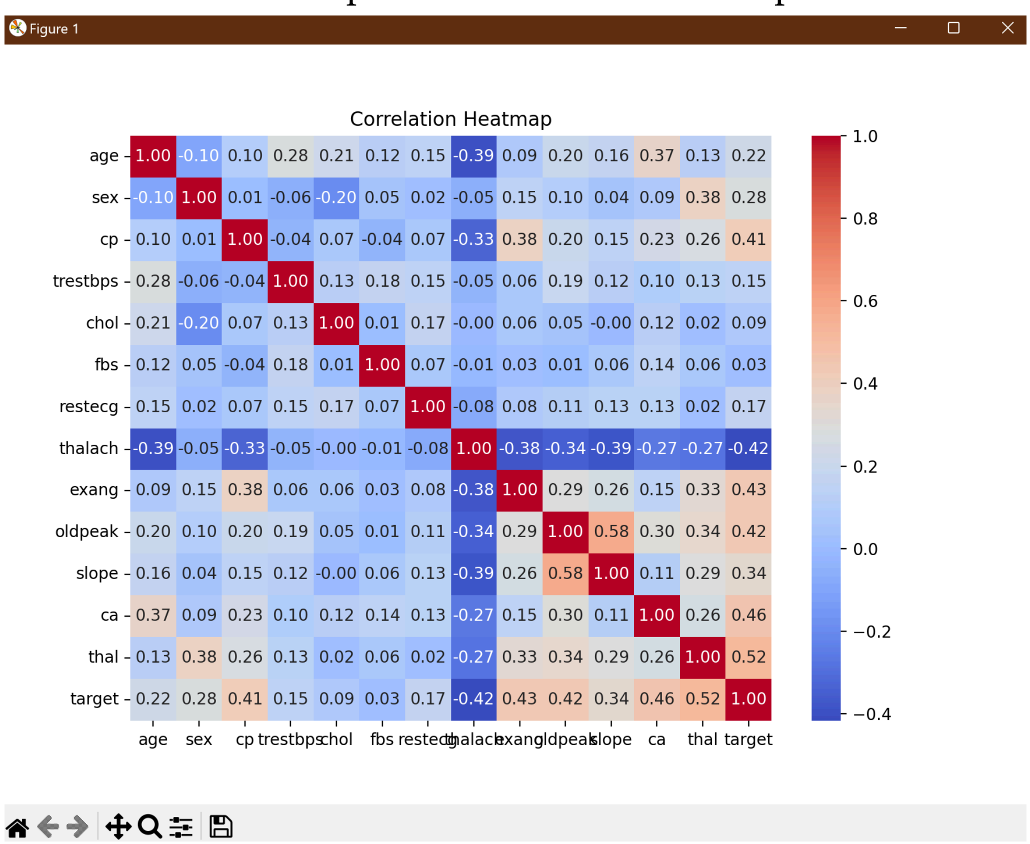
Week 1: Disease Prediction Using Patient Data

The Cleveland Heart Disease dataset from the UCI repository was used for this task. After loading the data, missing values represented by “?” were replaced with **NaN** so they could be handled properly. The target column was converted into a binary variable, where **0** represents **no disease** and an**y value greater than 0 represents the presence of disease**. For preprocessing, numeric columns with missing values were filled with their median, while categorical columns were filled with their mode. All numeric features (except the target) were normalized between **0 and 1 using Min–Max scaling,** and categorical features were converted into numerical form using **Label Encoding**. Exploratory Data Analysis was performed using describe() to view summary statistics, and a correlation heatmap was plotted to observe feature relationships with the target. The processed dataset was then split into training and testing sets (80/20), and two models were trained: Logistic Regression and Random Forest. Logistic Regression achieved an accuracy of **85.25%**, while Random Forest achieved an accuracy of **90.16%**. Based on this comparison, Random Forest performed better and was chosen as the preferred model for disease prediction.



Week 2: Cancer Detection Using Histopathological Images

For this task, a subset of the **Breast Cancer Histopathological Image Dataset** from Kaggle was used, consisting of 500 images divided into benign and malignant classes. The dataset was organized into class-specific subfolders, with **400 images for training** and **100 images for validation**. All images were resized to **128×128 pixels** to standardize input dimensions, and pixel values were normalized to the range [0, 1] for the CNN model.

A **Convolutional Neural Network (CNN)** was trained from scratch with three convolutional layers, each followed by max-pooling, and dense layers for classification. The model was compiled with the Adam optimizer and binary cross-entropy loss and trained for **10 epochs**. The scratch CNN achieved a validation accuracy of **98.0%** with a validation loss of **0.055**, demonstrating strong classification performance on the small dataset.

To improve generalization, **Transfer Learning** was applied using a pre-trained **VGG16** model with ImageNet weights. The convolutional base was frozen to act as a feature extractor, and fully connected layers were added on top for binary classification. Instead of normalization, the dataset was preprocessed using VGG16's preprocess_input function. The model was trained for **5 epochs**, achieving a validation accuracy of **99.0%** with a validation loss of **0.013**, outperforming the scratch CNN.

In conclusion, both models achieved high performance, but the **VGG16 transfer learning approach** provided better results, confirming the effectiveness of pre-trained feature extractors for medical imaging tasks with limited data.

Week 3: Skin Cancer and Pneumonia Detection

For this task, a subset of the ISIC Skin Cancer Dataset was used, consisting of 500–1,000 images divided into cancerous and non-cancerous classes. The dataset was reorganized into two main folders: *cancer* and *non-cancer*, with separate training and testing splits. All images were resized to 128×128 pixels and normalized to the range [0, 1].

A **Transfer Learning** approach was applied using a pre-trained **ResNet50** model with ImageNet weights. The convolutional base was frozen to act as a feature extractor, and custom dense layers were added for binary classification. The model was compiled using the **Adam optimizer** and **binary cross-entropy loss**, and trained for several epochs.