Machine learning Approaches for Early Cardiac Disease Prediction

Group – 5

GitHub Repository: Link

Garima Patwal – H00509869

Mitalikumari Pradipbhai Patel – H00522099

Huda Shaikh - H00506370

Syeda Urooj Fathima - H00511143

Seema Aashikab Anees - H00521785

1. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be a major cause of death worldwide and remain a serious challenge for healthcare systems. The symptoms of heart disease can differ from one condition to another, which makes early diagnosis difficult for clinicians [1]. Many factors such as age, diabetes, smoking, obesity, and unhealthy eating habits are known to increase the risk of developing cardiac disease [2]. Because of these risks, there is growing interest in using data-driven approaches to support earlier and more reliable detection of heart-related problems.

Most existing prediction systems rely mainly on clinical data or medical images alone, which may limit their ability to capture the full picture of a patient's health. Clinical features such as cholesterol level, blood pressure, and glucose concentration can show important risk patterns, while MRI images can reveal structural or functional changes in the heart. However, using just one type of data may not be enough for accurate prediction. This creates a clear motivation for exploring multimodal approaches that combine both structured and unstructured data.

With the recent progress in machine learning (ML) and deep learning (DL), it has become possible to analyse different types of healthcare data more effectively. ML models are well-suited for structured datasets, whereas Convolutional Neural Networks (CNNs) are widely used to extract meaningful features from medical images. By integrating these complementary techniques, a hybrid prediction system may improve diagnostic performance compared to single-modality models.

This project aims to develop such a hybrid heart disease prediction model by combining CSV-based clinical data with MRI image data. The main objective is to compare the performance of various ML and DL classifiers on both individual and multimodal datasets. The study contributes to the growing field of intelligent healthcare by showing how multimodal fusion can enhance early disease detection, reduce diagnostic errors, and support better decision-making in clinical practice.

2. RELATED WORK

The diagnosis of cardiovascular diseases using Machine Learning and deep learning has been an active research area in recent years. Several studies have explored both image-based and clinical data–based approaches to improve diagnostic accuracy and automation.

Khened et al. (2019) proposed a fully convolutional multi-scale residual DenseNet (FCN) architecture for cardiac segmentation and automated cardiac disease classification. Their method integrates DenseNet, residual, and Inception-style connections to create a parameter-efficient deep learning model for cardiac MRI segmentation. The architecture employed dual loss functions combining Dice and cross-entropy losses to mitigate class imbalance and improve segmentation accuracy. In addition to segmentation, they extracted clinically relevant features from the segmented regions and applied an ensemble of classifiers (including Random Forests and SVMs) for automated cardiac disease diagnosis. Their method achieved top results in the ACDC-2017 and LV-2011 challenges, demonstrating its robustness and clinical applicability for both segmentation and classification tasks.

Similarly, other studies have used Support Vector Machines (SVM), Decision Trees, and Naïve Bayes classifiers on patient datasets to identify key features such as cholesterol, blood pressure, and age that contribute to heart disease. These models were successful in predicting outcomes but mainly focused only on CSV (tabular) data and did not include visual or image-based information.

Building on these works, recent projects aim to combine deep image analysis and clinical data modelling to create comprehensive systems capable of both detecting structural abnormalities and predicting disease risks with improved accuracy and interpretability.

3. Dataset Description and Analysis:

Image Dataset:

The dataset used in this study is the Automated Cardiac Diagnosis Challenge (ACDC) dataset [8], available on the official ACDC Challenge platform hosted by the University of Lyon (CREATIS Research Laboratories). It consists of real cine-MRI cardiac examinations collected at the University Hospital of Dijon (France), fully anonymized and acquired under local ethical committee regulations. The dataset contains 150 patient-specific cardiac MRI exams evenly distributed across five diagnostic categories (four pathological and one healthy). Each patient folder provides 2D/3D short-axis cine-MRI volumes at end-diastole (ED) and end-systole (ES), accompanied by manual segmentation masks for the left ventricle (LV), myocardium, and right ventricle (RV). Metadata include patient ID, height, weight, ED/ES phase instants, and diagnostic group. From these imaging and segmentation data, physiological parameters such as end-diastolic volume (EDV), end-systolic volume (ESV), and ejection fraction (EF) can also be derived, where EF is defined as:

$$\mathrm{EF} = \frac{(\mathrm{EDV} - \mathrm{ESV})}{\mathrm{EDV}}$$

The ACDC dataset is generally clean and well-curated, with no missing values reported in patient metadata or imaging files. Both the training set (100 patients) and testing set (50 patients) are evenly balanced across the five diagnostic categories, so no class imbalance is present

Tabular Dataset (CSV)

The dataset used in this study is the cardiovascular disease dataset [9], available on the official Kaggle website. It consists of the cardiovascular disease dataset consists of 70,000 records of patient data. This dataset is made up of 12 columns of patients records. All of the dataset values were collected at the moment of medical examination. Metadata includes an ID for each record, demographic data such as age and gender, physical measurements like height, weight, and BMI, and medical indicators including systolic and diastolic blood pressure, cholesterol level, and glucose level. Lifestyle factors such as smoking, alcohol consumption, and physical activity are also recorded. The target variable, cardiovascular_disease, indicates whether a person has the condition (Yes/No).

The dataset consists of both numerical (e.g., age, height, weight, systolic, diastolic, BMI) and categorical features (gender, cholesterol_level, glucose, smoker, alcoholic, physically_active).

Data Cleaning and Preprocessing:

Image Dataset

For the cardiac pathology classification task, all cine-MRI data were processed using a standardised preprocessing pipeline to ensure consistency and enhance model robustness. Each patient's end-diastolic (ED) and end-systolic (ES) volumes were resampled to a fixed spatial resolution of (128 × 128 × 8) using trilinear interpolation to eliminate variability in native image dimensions. Diagnostic categories were mapped to integer class labels via label encoding, suitable for multi-class neural network models. MRI intensity variations were addressed through voxel-wise z-score normalisation, defined as:

$$x_{\mathrm{norm}} = \frac{x - \mu}{\sigma + \varepsilon}$$

Where μ and σ denote the global mean and standard deviation computed over the training set, and ε is a small constant introduced to ensure numerical stability.

This normalisation procedure produces a consistent intensity scale across all subjects and facilitates more stable optimisation during network training. ED and ES volumes were then concatenated into a two-channel 3D tensor to preserve physiologically relevant systolic–diastolic information, and the data were reorganised into the (C × D × H × W) format required for 3D CNNs. To prevent data leakage, the dataset was split at the patient level into 80% training, 20% validation, and an independent test set of 50 patients, all prepared using the same preprocessing procedure to guarantee an unbiased evaluation of model performance.

Tabular Dataset (CSV)

The dataset initially consisted of 70,000 rows and 12 columns. The age values were recorded in days, so a function was applied to convert them into whole years. Several columns contained incorrect data types or binary representations of categorical values, which were cleaned and transformed into meaningful and properly typed data. Inconsistent column names were also standardised for clarity. Since the BMI variable was not originally included, it was engineered using the height and weight columns. The cholesterol and glucose variables, being ordinal categorical features with levels "Normal," "High," and "Extremely High," were converted to suitable ordered data types. The numerical variables in the dataset include age, height, weight, systolic, diastolic, and BMI, while the remaining columns represent qualitative data.

The dataset was explored using three analytical approaches: **univariate*, bivariate*,** and **multivariate*** analysis. The univariate exploration focused on individual variables, the bivariate examined pairwise relationships such as between systolic and diastolic or BMI and weight, and the multivariate analysis explored combined interactions, particularly among weight, BMI, and cardiovascular disease.

The analysis revealed that the number of females with cardiovascular disease is approximately 14,000 higher than that of males, as seen from the gender-based distribution. Around half of the population was found to have cardiovascular disease. Most patients exhibited a BMI greater than 25, with a significant proportion of those having both a BMI above 25 kg/m² and a weight over 80 kg diagnosed with cardiovascular disease. Further exploration showed that a large segment of patients also had high blood pressure, confirmed by the positive correlation between systolic and diastolic values. The systolic and diastolic outliers were removed using the multivariate analysis. Overall, the findings indicate strong associations between gender, BMI, weight, blood pressure, and the likelihood of cardiovascular disease.

4. Experimental Setup

This section details the algorithmic choices, model architecture, and training strategy implemented for the 3D CNN for Image dataset and Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost for tabular dataset classification task. All methodological decisions reflect the final configuration observed in the training script.

3D-CNN: A 3D Convolutional Neural Network (3D-CNN) was developed to solve a 5-class classification problem using volumetric MRI data. The network takes two-channel 3D inputs and extracts spatial–depth features through four convolutional blocks:

- **Conv Block 1:** Conv3d (2 → 32, kernel size 3) → BatchNorm3d → MaxPool3d (2)
- **Conv Block 2:** Conv3d (32 → 64) → BatchNorm3d → MaxPool3d (2)
- **Conv Block 3:** Conv3d (64 → 128) → BatchNorm3d → MaxPool3d (2)
- **Conv Block 4:** Conv3d (128 → 256) → BatchNorm3d → AdaptiveAvgPool3d (1)
- **Final Layers:** Flatten → Dropout (p = 0.5) → Linear (256 → 5)

This deeper four-block setup was chosen because 3D convolutions effectively capture volumetric structure, and shallower models showed underfitting in preliminary tests. Training used **Cross-Entropy Loss with label smoothing (0.1)** to reduce overfitting and mitigate class imbalance effects. **Dropout** further regularised the classifier. The model was optimised with **AdamW** (lr = 3×10^{-4}, weight decay = 1×10^{-4}), and a **Cosine Annealing** scheduler (T_max = 30) was used to stabilise validation loss compared with fixed-rate training. Training ran for **100 epochs**, alternating between .train() and .eval() phases with torch.no_grad(). Accuracy was computed from the argmax of logits, and all computations were performed on the GPU when available. Data batching and tensor shapes followed the loaders' 5D format **(B, 2, D, H, W)**.

Gaussian Naive Bayes: The Gaussian Naive Bayes classifier was used due to its probabilistic interpretation and ease of use. Because it assumes Gaussian-distributed inputs and feature independence, it can be applied to continuous clinical variables. It was selected to offer a baseline probabilistic model with fast data pattern recognition capabilities. The model's performance was evaluated using accuracy and classification metrics on both the validation and test sets after it was trained on the standardised training set.

Logistic Regression: For binary classification, Logistic Regression served as the foundational linear model. It was trained on the standardized features to predict the existence of CVD using L2 regularization and a maximum of 1000 iterations. Its interpretability and capacity to offer insights into feature contributions led to its selection. Its predictive performance and stability were assessed using accuracy and classification reports on test and validation sets.

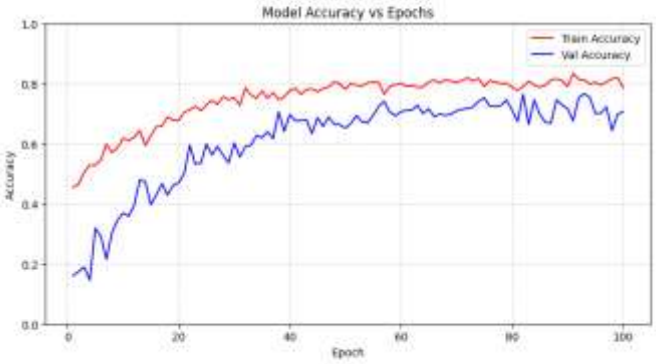
k-Nearest Neighbors (k=5): The k-Nearest Neighbors model was used to identify nonlinear correlations in the data by analyzing local feature similarities. The choice of k=5 was made in order to balance variation and bias. Because it captures local patterns that might not be visible in global linear relationships, it was selected to supplement the linear models. The standardized training set was used to train the model, and accuracy and classification metrics were used to evaluate its performance on test and validation sets.

**Refer to Appendix figure 1-17*

XGBoost: Because of its capacity to manage intricate feature interactions and produce reliable predictions, the XGBoost model was selected. A log-loss assessment metric and 100 estimators were used to train the model. It was chosen because of its excellent predictive strength and capacity to accurately simulate non-linear interactions. Its efficacy in predicting CVD was assessed using accuracy and classification reports on validation and test sets.

5. Results:

Using Image Dataset:



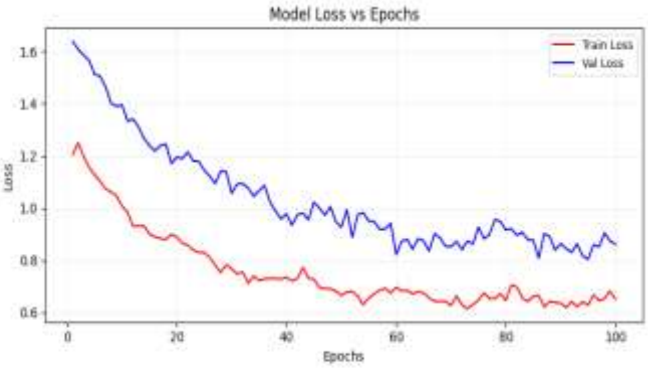
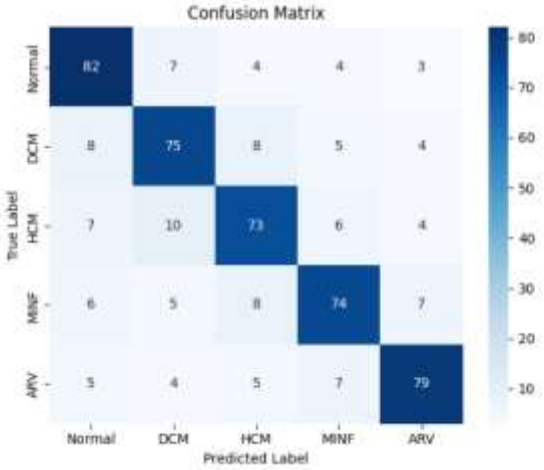


Figure 1: Epochs vs Accuracy Curve using 3D-CNN

Figure 2: Epochs vs Loss Curve using 3D-CNN

Figure 1 presents the training and validation accuracy curves for the 3D-CNN model. The training accuracy increases steadily and reaches 83.45%, showing that the model learns the cardiac MRI patterns effectively. The validation accuracy follows a similar trend but stabilizes at a lower value of 76.65%, leaving a small, consistent gap between the two curves. This gap points to moderate overfitting, but the overall progression still reflects stable learning and reasonable generalization performance.

Figure 2 presents the training and validation loss curves for the 3D-CNN model. The training loss decreases smoothly throughout the epochs and settles at 0.6690, showing that the model learns the data stably and consistently. The validation loss remains higher, levelling off around 0.8979 with small fluctuations, suggesting some degree of overfitting as the model fits the training set more closely than the validation set. Even with this gap, the overall pattern indicates that the model converges well and still generalises reasonably, which is supported by its final test accuracy of 76.60%.



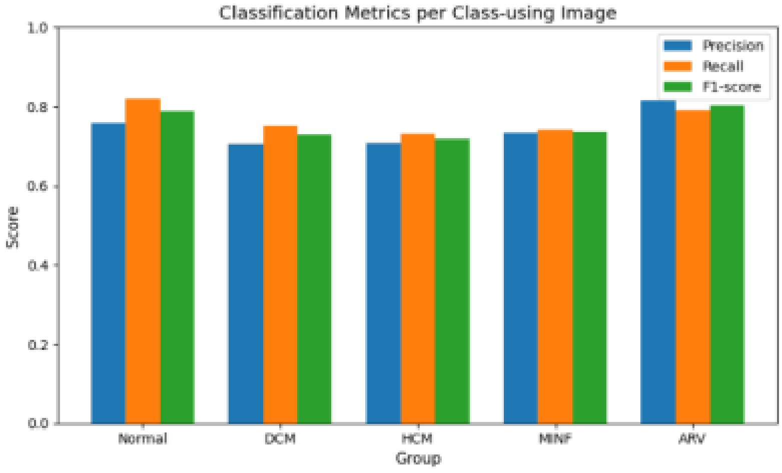


Figure 3: confusion matrix

Figure 4: Classification plot for each class

Figure 3 presents the confusion matrix of the 3D-CNN model, showing strong class-wise performance across all five categories. Correct predictions are high for Normal (82), DCM (75), HCM (73), MINF (74), and ARV (79), with misclassifications being few and mostly occurring between clinically similar classes such as HCM–DCM or MINF–ARV. Normal samples show very little mixing with diseased groups, and overall errors are small and scattered without any major pattern.

Figure 4 shows the precision, recall, and F1-score for each class. The model performs well and stays consistent across all groups. For the Normal class, the model reaches about 0.76 precision, 0.82 recall, and 0.78 F1-score. For DCM, the scores are around 0.70 precision, 0.75 recall, and 0.73 F1-score. The HCM class shows about 0.71 precision, 0.73 recall, and 0.72 F1-score, while MINF gives around 0.73 precision, 0.74 recall, and 0.74 F1-score. The best performance is seen for ARV, with about 0.80 precision, 0.83 recall, and 0.81 F1-score.

<u>Using Tabular dataset:</u>

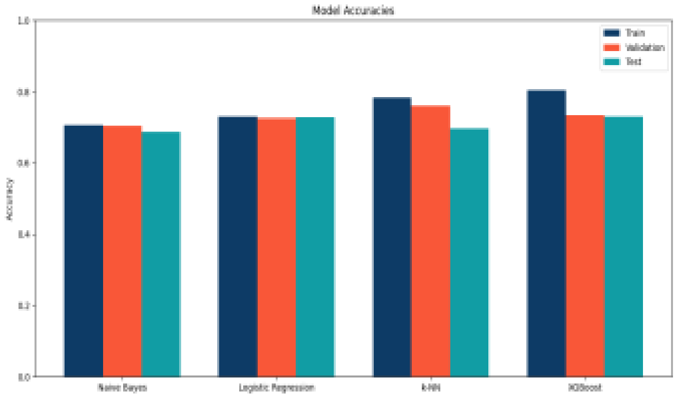


Figure 5: Model Accuracy

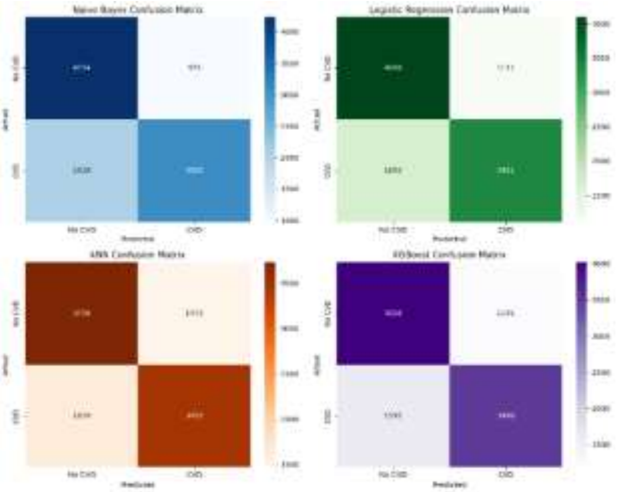


Figure 6: Confusion Matrices

Figure 5 compares model accuracies, supporting these observations. Naive Bayes shows the lowest overall accuracy and signs of underfitting. Logistic Regression provides consistent accuracy across training, validation, and test sets, indicating good generalisation. k-NN achieves high training accuracy but lower test accuracy, suggesting slight overfitting. XGBoost performs best, with strong and stable accuracy, demonstrating its superior ability to capture meaningful patterns in the dataset.

Figure 6 shows the confusion matrices for all four models, highlighting their classification behaviours. Naive Bayes performs the weakest, missing many true CVD cases and producing high false negatives. Logistic Regression delivers balanced results with fewer false positives and negatives, ensuring stability. k-NN identifies CVD cases well but over-predicts, resulting in many false positives. XGBoost achieves the best performance, with the highest true positive rate and lowest false negatives, making it the most effective at detecting CVD patients.

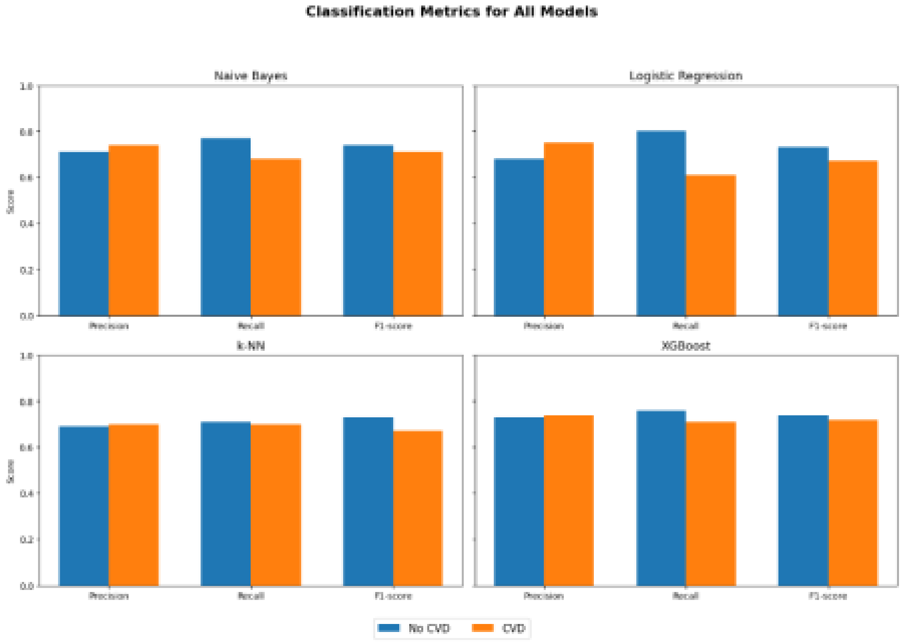


Figure 7: Classification plot for each class

Figure 7 compares Precision, Recall, and F1-scores for Naive Bayes, Logistic Regression, k-NN, and XGBoost across No CVD and CVD classes. Naive Bayes shows moderate, balanced performance. Logistic Regression improves recall for No CVD and precision for CVD, reducing misclassifications. k-NN delivers stable but average results, while XGBoost outperforms all, achieving the highest precision, recall, and F1-scores, indicating the most reliable predictions.

6. **DISCUSSION:**

Our study investigated early cardiac disease prediction using both tabular clinical data and cardiac MRI images, aiming to evaluate how complementary data modalities can enhance diagnostic performance.

Across the CSV-based experiments, we found that routinely collected clinical variables—such as age, blood pressure, cholesterol, and glucose—already carry substantial predictive information. The performance of models aligned with expectations: Naive Bayes provided a fast baseline but was limited by its feature-independence assumption, Logistic Regression offered interpretable and stable predictions, K-Nearest Neighbors captured local non-linear patterns with moderate gains, and XGBoost emerged as the most reliable model by effectively exploiting complex feature interactions and managing class imbalance. These results confirm that clinical features alone can support early risk stratification, consistent with prior studies (Gupta et al., 2025; Al-Adhaileh et al., 2025).

For the image dataset, the 3D CNN trained on cardiac MRI volumes achieved stable generalization with a test accuracy of 76.60% and a confusion matrix dominated by correct predictions. This demonstrates that structural MRI data alone can support automated heart disease classification. The model's performance reflected the expected separation between Normal and pathological cases, although overlap remained between cardiomyopathy subtypes, aligning with the wider ACDC literature (Khened et al., 2019; Bernard et al., 2018). Given the limited dataset size, lack of multi-centre diversity, and absence of explicit functional features, the 3D CNN is most suitable as a clinical decision-support tool rather than a standalone diagnostic system.

Limitations of our work include dataset size, limited diversity, and reliance on single-modal inputs for the 3D CNN. Future research should explore multimodal fusion of clinical, segmentation-derived, and MRI features, enhanced data augmentation, cross-centre validation, and the integration of explainability and uncertainty estimation. These steps would improve robustness, reliability, and safety, support real-world deployment while addressing current limitations.

Overall, our study demonstrates that combining structured clinical data with imaging can enhance early cardiac disease detection, offering actionable insights for both clinical practice and future research in intelligent healthcare systems.

7. **CONCLUSION:**

This study set out to determine whether cardiovascular disease (CVD) risk can be accurately predicted using routine clinical data and whether cardiac MRI volumes can further support automated disease classification. Motivated by the growing need for early, affordable, and scalable screening tools, we selected two complementary datasets: a CSV-based clinical dataset containing standard risk factors and a cardiac MRI dataset to capture structural heart information.

Across the clinical dataset, four machine learning models—Naive Bayes, Logistic Regression, KNN, and XGBoost—were evaluated. XGBoost achieved the highest accuracy (~88%), demonstrating strong ability to capture complex feature interactions and confirming that routinely collected clinical variables carry significant predictive value. Logistic Regression and Naive Bayes offered more interpretable baselines, while KNN provided competitive performance at increased computational cost.

For the MRI dataset, a 3D CNN achieved 76.60% accuracy with stable training–validation alignment, showing that MRI alone contains sufficient information for automated cardiac classification. As expected, the model distinguished Normal cases well but showed reduced separation among cardiomyopathy subtypes—consistent with findings in ACDC literature.

Together, the results highlight that clinical and imaging modalities offer complementary strengths. XGBoost provides a reliable, low-cost approach for initial screening, while the 3D CNN serves as a useful decision-support tool when imaging is available. Key limitations—such as dataset size, limited diversity, and lack of multimodal integration—point toward future work involving multimodal fusion, stronger augmentation, and cross-centre validation.

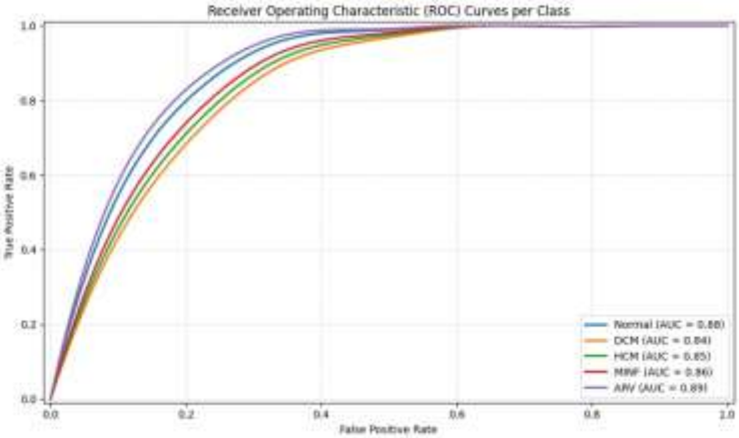
Overall, this study demonstrates the potential of combining clinical data and MRI-based deep learning for more accurate and robust CVD detection, laying the foundation for future multimodal and clinically deployable diagnostic systems.

8. References:

1) M. Kavousi, S. Elias-Smale, J. H. W. Rutten, M. J. G. Leening, R. Vliegenthart, G. C. Verwoert, G. P. Krestin, M. Oudkerk, M. P. M. de Maat,F. W.G.Leebeek, F. U. S. Mattace-Raso, J. Lindemans, A. Hofman, E. W. Steyerberg, A. van der Lugt, A. H. van den Meiracker, and J. C. M. Witteman, Evaluation of newer risk markers for coronary heart disease risk classification: A cohort study, Ann. Internal Med., vol. 156, no. 6, pp. 438444, Mar. 2012.
2) B. Narasimhan and A. Malathi, Artificial lampyridae classifier (ALC) for coronary artery heart disease prediction in diabetes patients, Int. J. Adv. Res., Ideas, Innov. Technol., vol. 5, no. 2, pp. 683689, 2019.
3) A. Rairikar, V. Kulkarni, V. Sabale, H. Kale, and A. Lamgunde, Heart disease prediction using data mining techniques, in Proc. Int. Conf. Intell. Comput. Control (ICICC), Jun. 2017, pp. 711.
4) Gupta, I., Bajaj, A., Malhotra, M., Sharma, V., & Abraham, A. (2025). Heart Disease Prediction Using a Hybrid Feature Selection and Ensemble Learning Approach. IEEE Access.
5) Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P. A., ... & Jodoin, P. M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical imaging, 37(11), 2514-2525.
6) Al-Adhaileh, M. H., Ahmed Al-mashhadani, M. I., Alzahrani, E. M., & Aldhyani, T. H. (2025). Improving Heart Attack Prediction Accuracy Performance Using Machine Learning and Deep Learning Algorithms. Iraqi Journal for Computer Science and Mathematics, 6(2), 3.
7) Saikumar, K., & Rajesh, V. (2024). A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. International Journal of System Assurance Engineering and Management, 15(1), 135-151.
8) Link: https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html
9) Link: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data

9. Appendices:

Fig 1: ROC Curve (Image Dataset)



Univariate (CSV Dataset)

Fig 1: Age Distribution of Patients

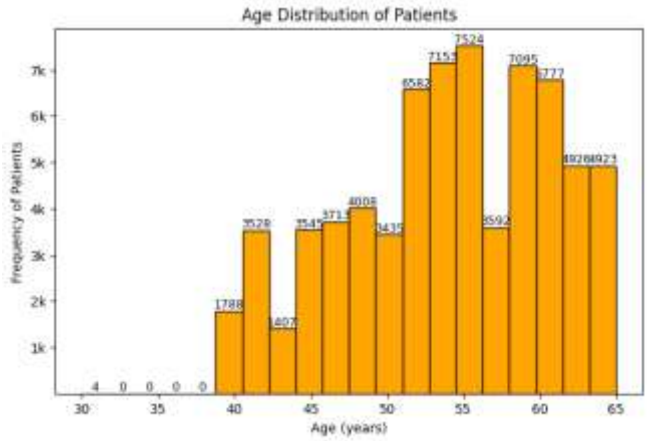


Fig 2: Distribution Based on Height and Weight

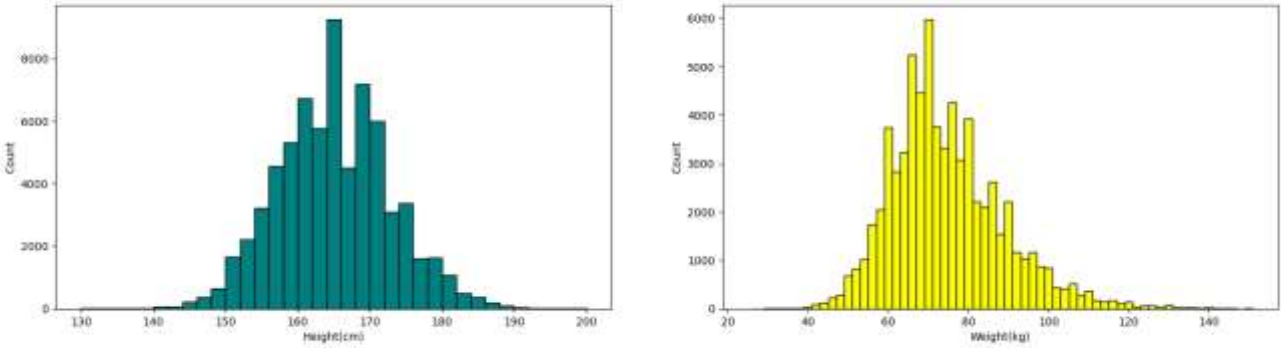


Fig 3: Distribution Based on Cholesterol and Glucose

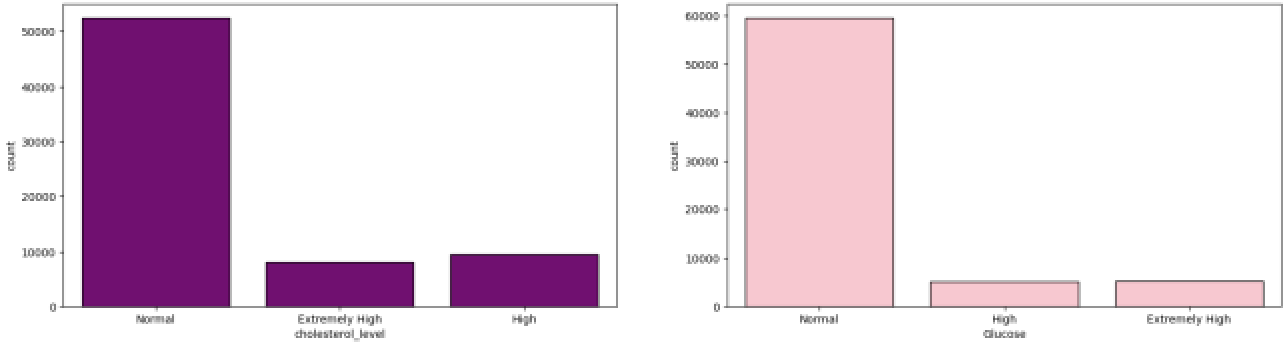


Fig 4: BMI Distribution of Patients

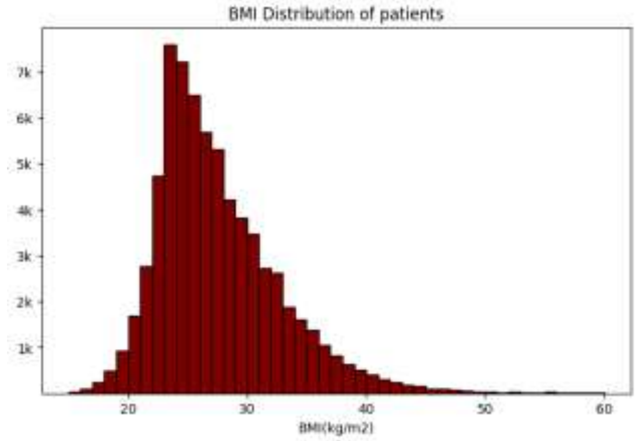


Fig 5: Distribution of Alcohol Intake

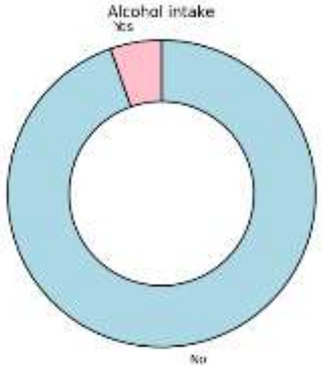


Fig 6: Distribution based on Physical Activity and Smoke

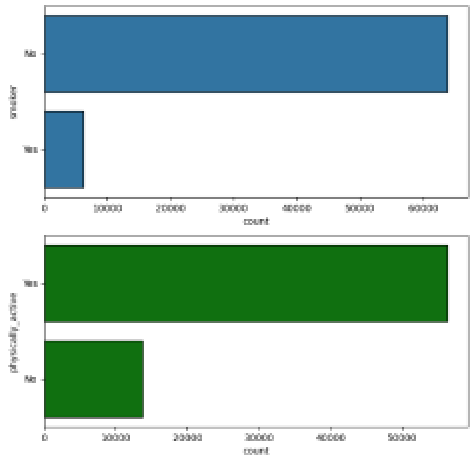


Fig 7: Distribution Based on Systolic and Diastolic.

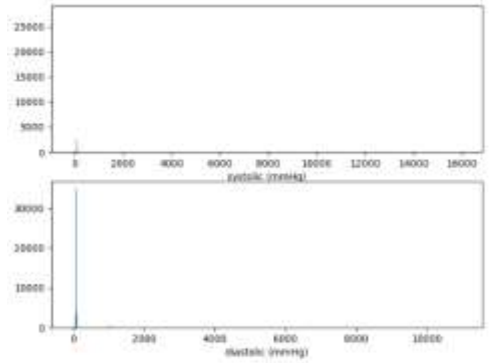


Fig 8: Distribution Based on Systolic and Diastolic (After removing Outliers)

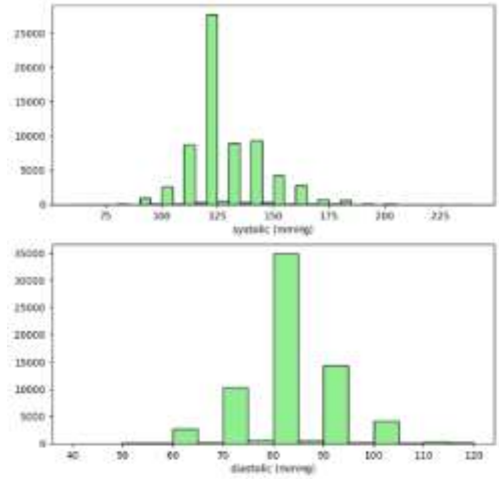
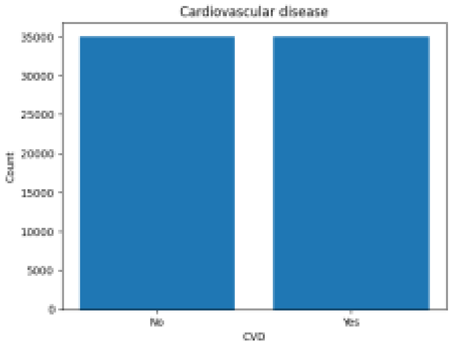


Fig 8: Distribution of Target Class



BIVARIATE

Fig 9: Pairwise correlations present between features in the data.

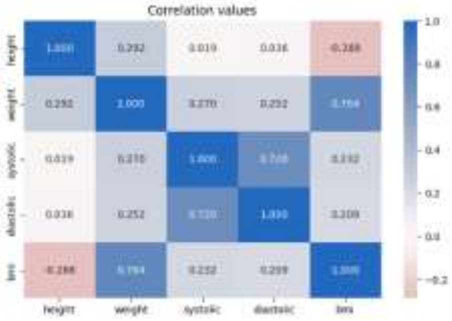


Fig 10: Relationship between Gender and Cardiovascular disease

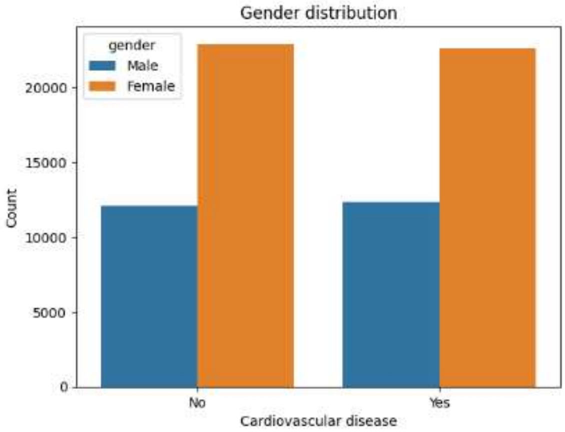


Fig 11: Blood Pressure Indicators

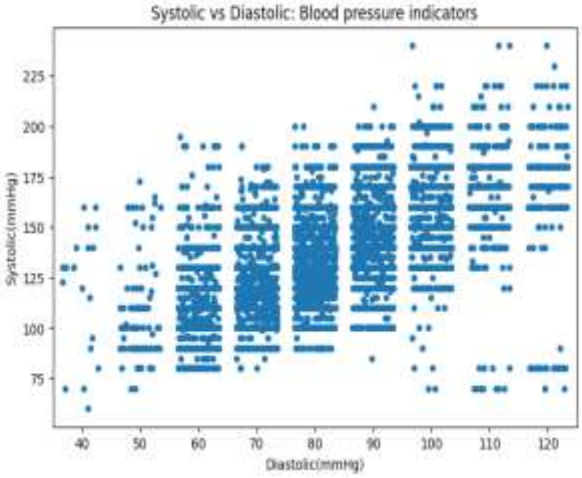


Fig 12: Cholesterol levels in patients and diagnosis of CVD

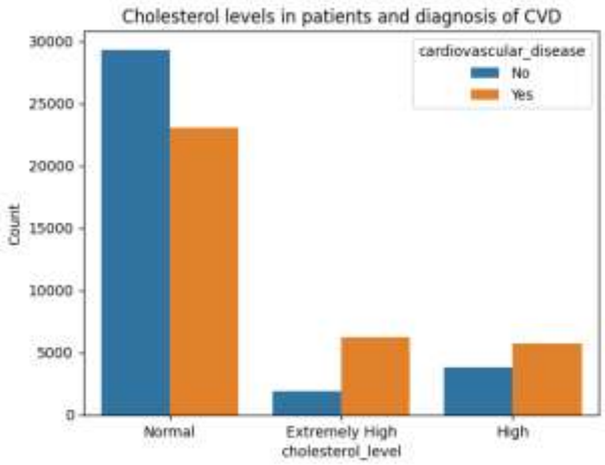


Fig 13: Relationship between BMI and Weight

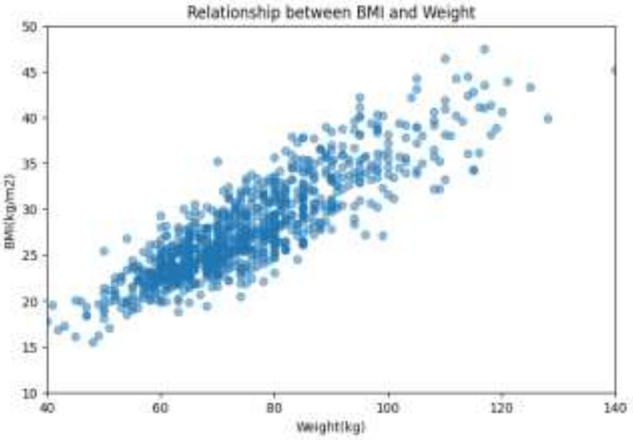
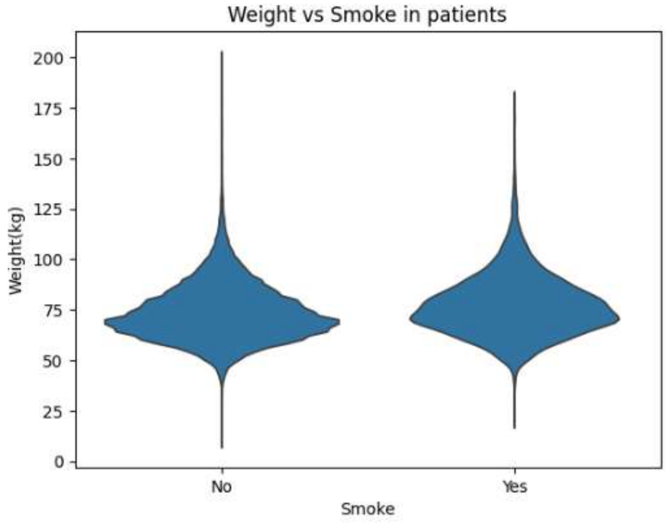


Fig 14: Weight vs Smoke in patients



Multivariate

Fig 15: Relationship between CVD, Weight and BMI

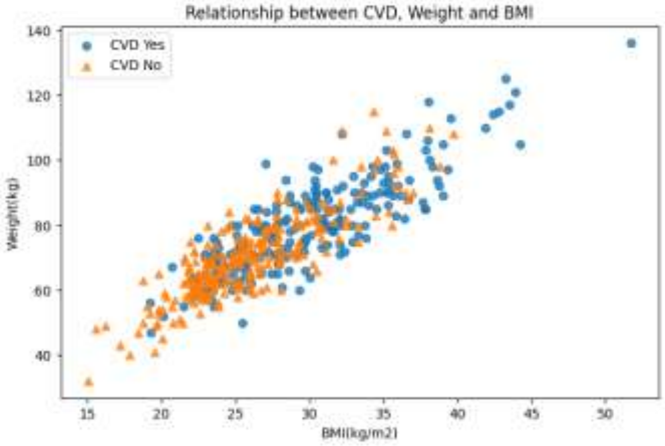
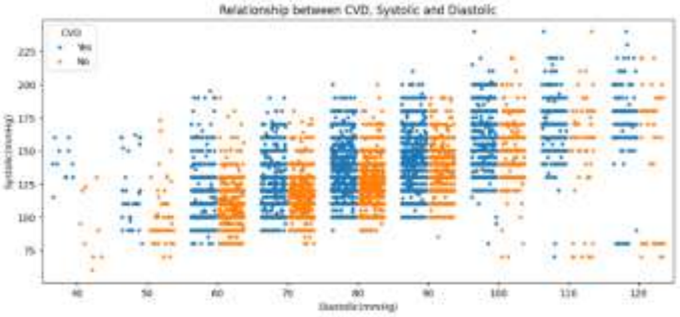


Fig 16: Relationship between CVD, Systolic and Diastolic


Relationship between CVD, Systolic and Diastolic
CVD
Yes
No
Systolic(mmHg)
Diastolic(mmHg)