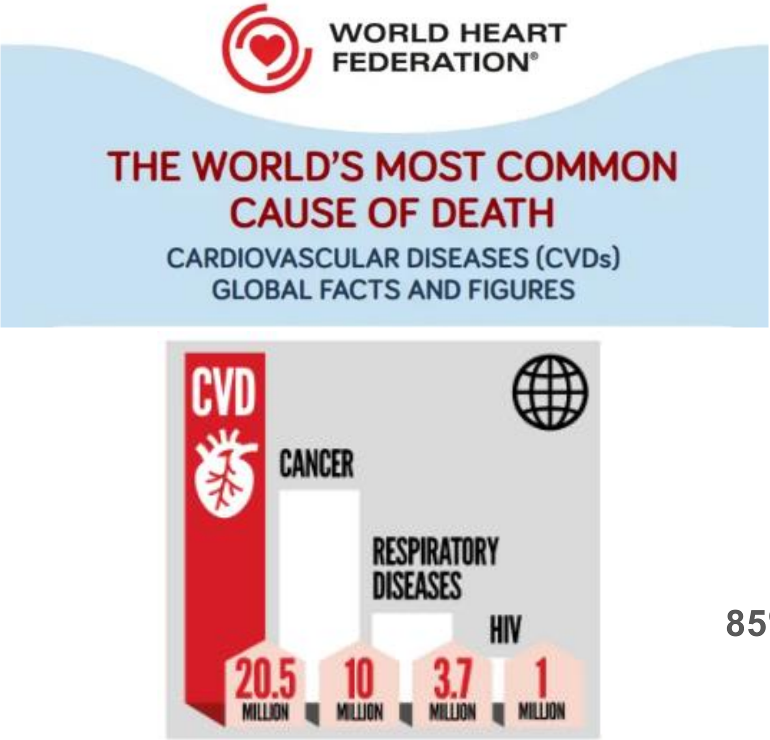
National College of Ireland
Research Day 2026
June 18, 2026

Early Diagnosis Prediction for People with Cardiovascular Conditions using Machine Learning Methods

Charlyn Rosales, Ph.D.
Gabriel Galang, MSIT
Engr. Lester Phil Cruz
Mr. Jonathan Tomacruz
Michelle Tansinsin, MSIT

NCI Coláiste Náisiúnta na hÉireann | National College of Ireland

INTRODUCTION



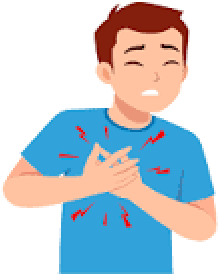
https://world-heart-federation.org/resource/cardiovascular-disease-infographic/

World Heart Federation & World Health Organization reported that Cardiovascular Diseases (CVDs) are the leading cause of death globally with estimated

20.5M deaths every year,
33% of all global deaths.

85% of these deaths were due to heart attack and stroke

MOST COMMON SYMPTOMS



CHEST PAIN

Pain or Discomfort in the center of the chest.

DIFFICULTY IN BREATHING

Shortness of breath or a sensation of not getting enough air as normal.

NAUSEA OR VOMITING

Unpleasant feeling in your stomach with an urge to throw-up.

FAINTING

Feeling that you might faint, become dizzy, or lose balance.

MOST COMMON SYMPTOMS

COLD SWEAT

Cool and moist feeling or sweating even when environment is not hot.

TURNING PALE

Unusually light or pale color of the skin compared to normal.

STROKE

Sudden weakness / numbness of the face, arm, or leg (one side of the body)

ONSET OF:

Confusion, Speech, Seeing, or Walking difficulty and severe headache with unknown reason.

PROBLEM

A Cardiovascular Diseases can lead to:

- Heart attack
- Stroke
- Death

World Health Organization

- It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin.

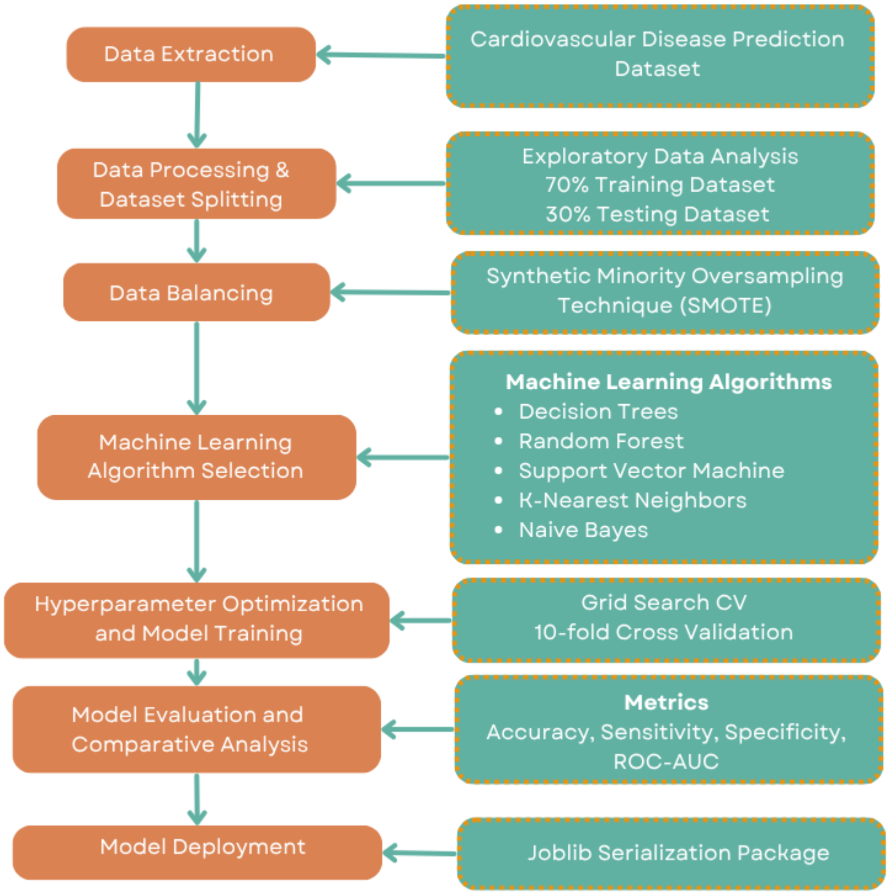
SOLUTION

Early Diagnosis, Prevention, and Intervention.

To aide in early diagnosis, there is a need to develop a Cardiovascular Disease Prediction Model using:

- Historical Records
- Machine Learning Algorithms

Block Diagram of CVD Prediction Model Development



PROGRAMMING SOFTWARES AND DEPENDENCIES

python + colab

pandas

NumPy

seaborn

matplotlib

scikit learn

DATASET

kaggle

Cardiovascular Disease Prediction Dataset

14 Columns

(13 Symptoms and 1 Target Class)

303 Rows

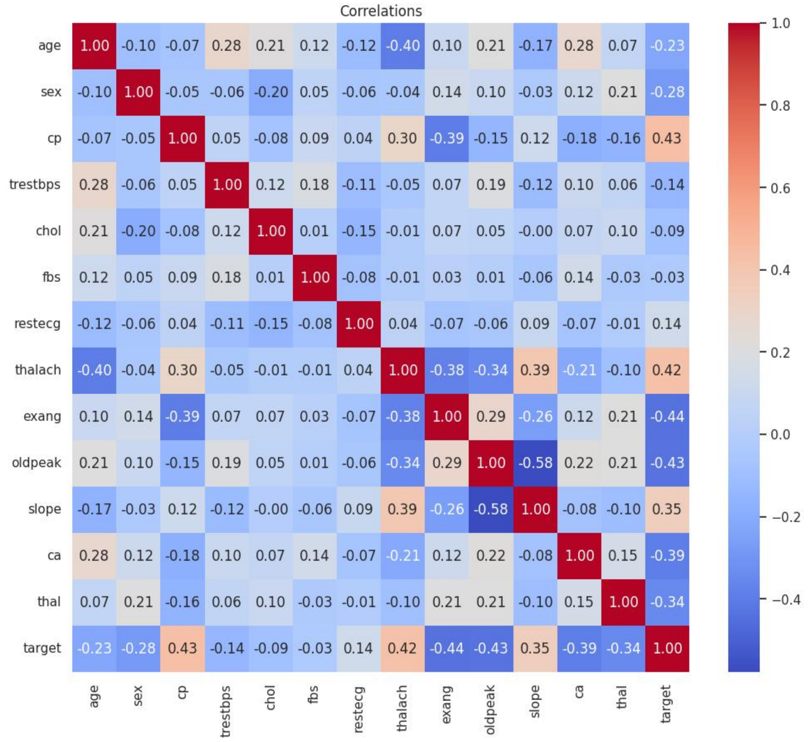
Source: https://www.kaggle.com/datasets/riyapatel1697/cdv-healthcare-data

Predictors

Attribute Name	Description
age	Person’s Age presented in years
sex	Gender of the person number 1 or male, and number 0 for female
cp	The person is experiencing chest pain and its type
trestbps	The person’s resting blood pressure measured in mm Hg upon admission to the hospital
chol	The person’s serum cholesterol measured in mg/dl
fbs	The person’s fasting blood sugar is greater than 120 mg/dl number 1 or true, and number 0 for false
restecg	The person’s results of their resting electrocardiographic
thalach	The person’s maximum heart rate recorded
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	The person’s ST depression influenced by exercise in relation to rest
slope	The person’s recorded slope of the peak exercise ST segment
ca	Person’s number of major blood vessels from 0-3 colored by fluoroscopy
thal	Characterized as number 3 for normal, number 6 for a fixed defect, then number 7 for a reversible defect
target	1 or 0

Cardio Vascular Diseases Prediction Dataset

Pearson Correlation Coefficient (PCC) Score of Predictors to the Target Class



Strongest Correlations (>0.40)

- Chest Pain (cp)
- Maximum Heart Rate (thalach)
- Exercise Induced Angina (exang)
- ST depression induced by exercise (oldpeak)

Weakest Correlations (<0.15)

- Fasting blood sugar
- Cholesterol
- Resting blood pressure

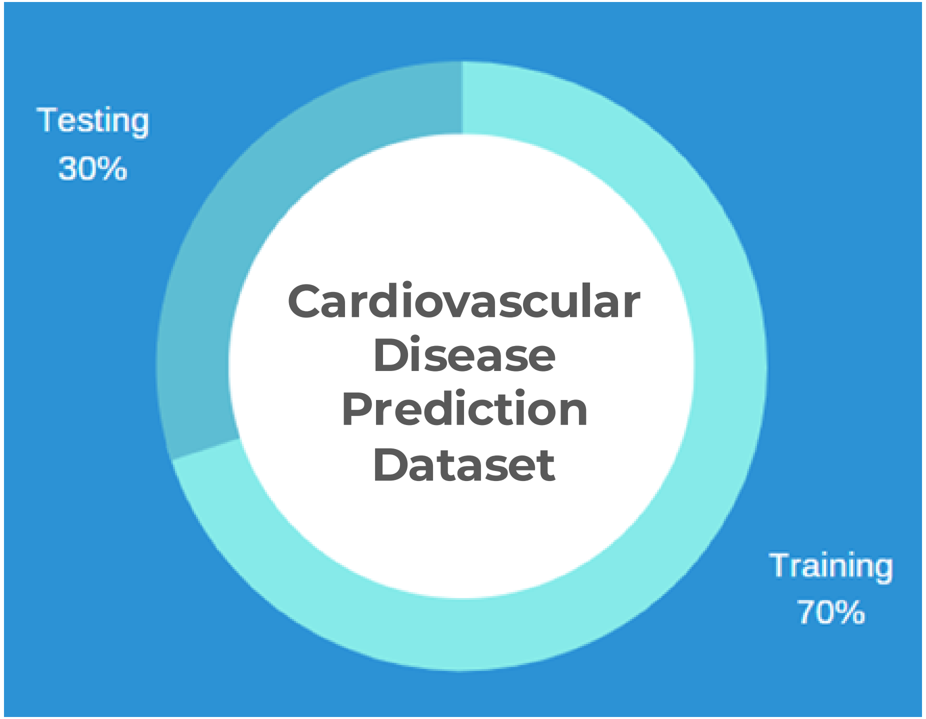
Variance Inflation Factor (VIF) Feature Selection

Predictors	VIF
age	1.44
sex	1.23
cp	1.40
trestbps	1.18
chol	1.15
fbs	1.08
restecg	1.06
thalach	1.65
exang	1.44
oldpeak	1.74
slope	1.66
ca	1.29
thal	1.19

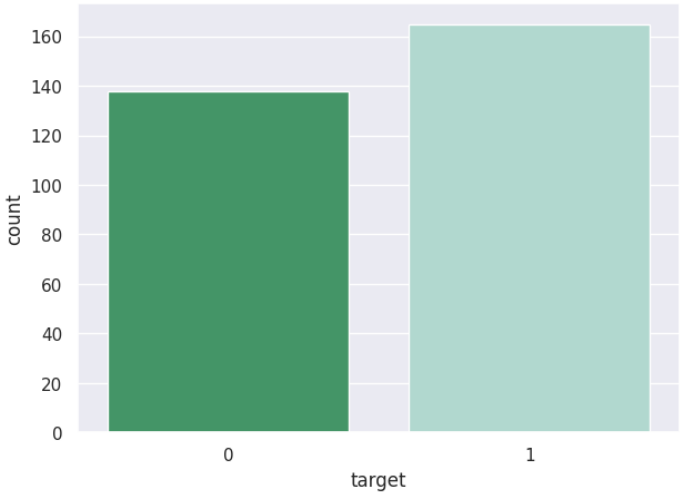
Standard Rules

1 no collinearity
< 5 moderate collinearity
>10 high collinearity and
needs to be removed

DATASET SPLITTING

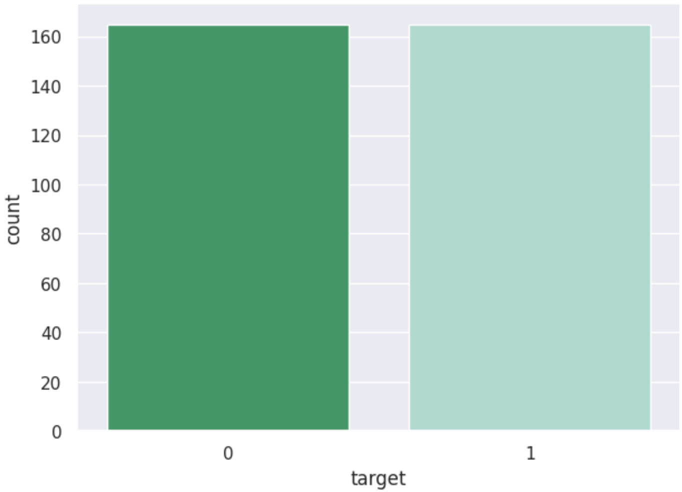


TRAINING DATASET BALANCING



Heart Disease is **Not Present**

Heart Disease is **Present**



Heart Disease is **Not Present**

Heart Disease is **Present**

Raw Training Dataset → **SMOTE** → **Balanced Training Dataset**

Benchmark Machine Learning Algorithm

Logistic Regression

Process	Accuracy	Sensitivity	Specificity	ROC – AUC
training	86.96	82.61	91.30	86.96
testing	71.43	68.29	74.00	71.15

Machine Learning Algorithm Selection

Decision Tree

```
params = { 'max_depth': [2, 3, 5, 10, 20],
'min_samples_leaf': [5, 10, 20, 50, 100],
'criterion': ["gini", "entropy"] }
```

Hyperparameter Optimization Results

DT Criterion	Minimum Samples	Maximum Depth	Score	Ranking
gini	**2**	**50**	**79.13**	**1**
gini	3	50	79.13	1
gini	5	50	79.13	1
gini	10	50	79.13	1
gini	20	50	79.13	1

Machine Learning Algorithm Selection

Random Forest

```
param_grid = {
    'criterion': ["gini", "entropy"],
    'max_depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [1,2,5,10,20],
    'n_estimators': [100, 200, 300],
    'bootstrap': [True, False]
}
```

Hyperparameter Optimization Results

RF Criterion	Maximum Depth	Minimum Samples	# of Estimators	Bootstrap	Score	Ranking
gini	**5**	**2**	**200**	**True**	**85.65**	**1**
entropy	3	1	200	True	85.65	1
entropy	5	5	100	True	85.65	1
gini	10	2	100	True	85.22	4
gini	5	1	300	False	85.22	4

Machine Learning Algorithm Selection

Support Vector Machine

```
param_grid = {
    'svm__kernel': ['linear', 'rbf'],
    'svm__C': [0.1, 1, 10],
    'svm__gamma': ['scale', 'auto']
}
```

Hyperparameter Optimization Results

C	gamma	Kernel	Score	Ranking
0.1	**scale**	**rbf**	**83.48**	**1**
0.1	auto	rbf	83.48	1
10	auto	linear	83.04	3
10	scale	linear	83.04	3
1	scale	rbf	82.61	5

Machine Learning
Algorithm Selection
K Nearest Neighbors

```
params = {
    'knn__n_neighbors': [3, 5, 7, 9, 11, 13],
    'knn__weights': ['uniform', 'distance'],
    'knn__metric': ['euclidean', 'manhattan']
}
```

Hyperparameter Optimization Results

Metrics	Neighbors	Weights	Score	Ranking
manhattan	**11**	**distance**	**85.65**	**1**
manhattan	7	distance	85.65	1
euclidean	7	uniform	85.22	3
euclidean	7	distance	84.78	4
manhattan	7	uniform	84.78	4

Machine Learning Algorithm Selection

Naïve Bayes

```
params = {'var_smoothing': np.logspace(0,-9, num=100)}
```

Hyperparameter Optimization Results

Var Smoothing	Score	Ranking
4.33	**84.78**	**1**
1.23	84.35	2
2.85	84.35	2
1.87	84.35	2
2.31	84.35	2

(TRAINING)

Prediction Model Statistical Performance Results

Algorithm	Accuracy	Sensitivity	Specificity	ROC-AUC
Benchmark Algorithm (Logistic Regression)	86.96	82.61	91.30	86.96
Decision Tree	79.13	80.00	78.26	79.13
Random Forest	**93.91**	**93.04**	**94.78**	**93.91**
Support Vector Machine	**88.26**	**86.09**	**90.43**	**88.26**
K Nearest Neighbors	**100**	**100**	**100**	**100**
Naïve Bayes	83.91	76.52	91.30	83.91

(TRAINING)

Prediction Model Statistical Performance Results

Algorithm	Accuracy	Sensitivity	Specificity	ROC-AUC
Benchmark Algorithm (Logistic Regression)	86.96	82.61	91.30	86.96
Decision Tree	79.13	80.00	78.26	79.13
Random Forest	**93.91**	**93.04**	**94.78**	**93.91**
Support Vector Machine	**88.26**	**86.09**	**90.43**	**88.26**
K Nearest Neighbors	**100**	**100**	**100**	**100**
Naïve Bayes	83.91	76.52	91.30	83.91

(TESTING)

Prediction Model Statistical Performance Results

Algorithm	Accuracy	Sensitivity	Specificity	ROC-AUC
Benchmark Algorithm (Logistic Regression)	71.43	68.29	74.00	71.15
Decision Tree	70.33	68.29	72.00	70.15
Random Forest	**79.12**	**73.17**	**84.00**	**78.59**
Support Vector Machine	**74.73**	**70.73**	**78.00**	**74.37**
K Nearest Neighbors	64.84	58.54	70.00	64.27
Naïve Bayes	**74.73**	**68.29**	**80.00**	**74.15**

(TESTING)

Prediction Model Statistical Performance Results

Algorithm	Accuracy	Sensitivity	Specificity	ROC-AUC
Benchmark Algorithm (Logistic Regression)	71.43	68.29	74.00	71.15
Decision Tree	70.33	68.29	72.00	70.15
Random Forest	**79.12**	**73.17**	**84.00**	**78.59**
Support Vector Machine	**74.73**	**70.73**	**78.00**	**74.37**
K Nearest Neighbors	64.84	58.54	70.00	64.27
Naïve Bayes	**74.73**	**68.29**	**80.00**	**74.15**

Prediction Model Performance in Training Dataset

Confusion Matrix

Training set: n = 230 (115 positive, 115 negative)

Actual	Predicted Positive	Predicted Negative
Positive	107	8
Negative	6	109

Performance Evaluation Result

Accuracy	**93.91%**
Sensitivity	**93.04%**
Specificity	**94.78%**
ROC-AUC	**93.91%**

Prediction Model Performance in Testing Dataset

Confusion Matrix

Testing set: n = 91 (41 positive, 50 negative)

Actual/Predicted	Positive	Negative
Positive	30	11
Negative	8	42

Performance Evaluation Result

Accuracy	**79.12%**
Sensitivity	**73.17%**
Specificity	**84.00%**
ROC-AUC	**78.59%**

JOBLIB SERIALIZATION PACKAGE

Transform the prediction model into a file that can be transmitted to other platforms.

CVDpredictionmodel.pkl

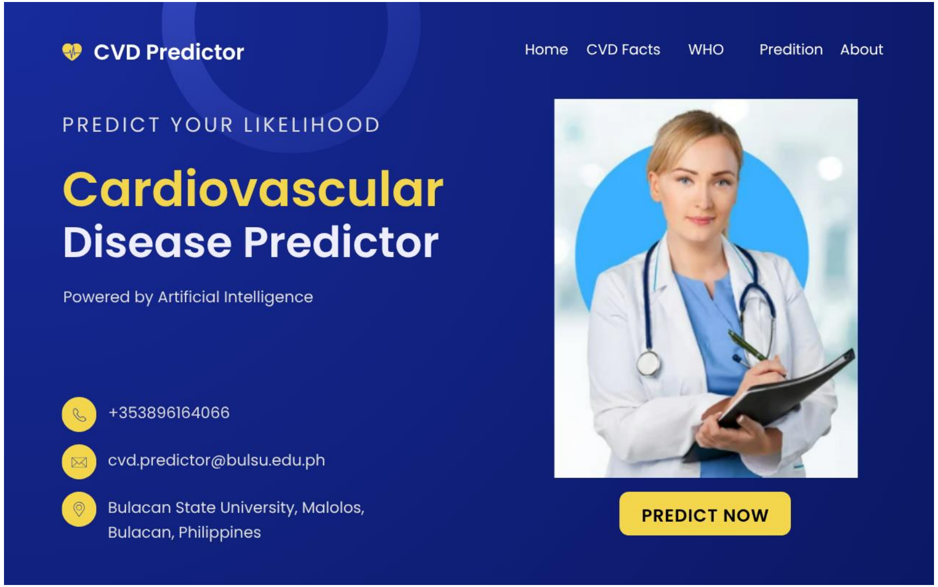
CONCLUSIONS

- **Random Forest Algorithm** can be used to build the CVD prediction model.
- The configuration that will yield the best performance of the algorithm was successfully determined through hyperparameter optimization using **GridSearchCV** and **10-fold cross-validation**.
- The prediction model yielded **93.91%** accuracy, **93.04%** sensitivity, **94.78%** specificity, and **93.91%** Area Under the ROC curve score.
- The output of this study can be used to develop **machine learning based applications** that can serve as a health monitoring and management tool.

RECOMMENDATIONS

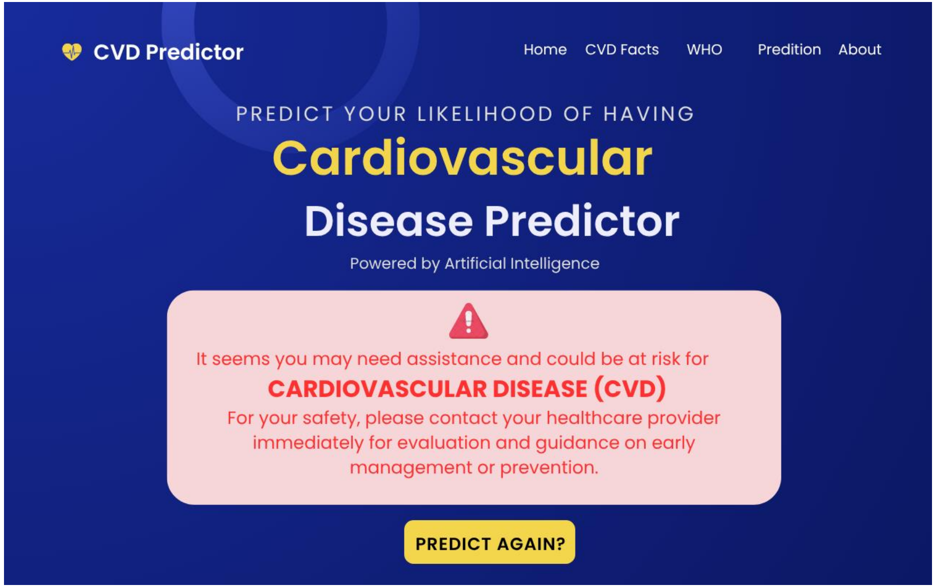
- Improve generalizability of the findings by increasing patient records through **additional clinical data** (hospital and clinic records)
- Exploring **broader risk factors** such as demographics, lifestyle, and family history.
- Prediction of patient's **severity** and **mortality**.
- Include a **10-year** risk assessment.
- Further testing on **larger datasets** before clinical application.
- Development of a real-time clinical **decision support system** via web and mobile platforms to assess symptoms with immediate notification of possibility of developing CVD.

Sample UI

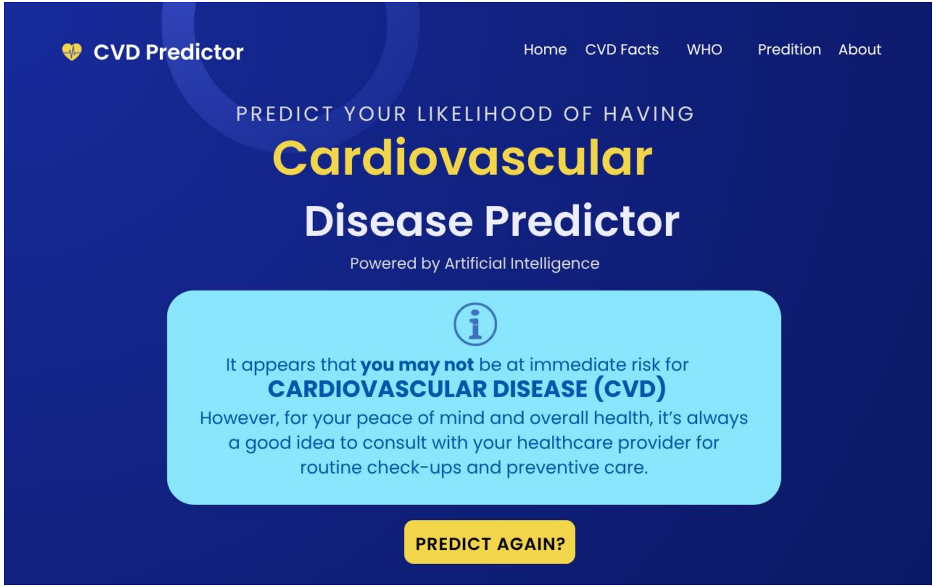

CVD Predictor
Home
CVD Facts
WHO
Predition
About
PREDICT YOUR LIKELIHOOD
Cardiovascular
Disease Predictor
Powered by Artificial Intelligence
+353896164066
cvd.predictor@bulsu.edu.ph
Bulacan State University, Malolos,
Bulacan, Philippines
PREDICT NOW


CVD Predictor
Home
CVD Facts
WHO
Predition
About
PREDICT YOUR LIKELIHOOD OF HAVING
Cardiovascular
Disease Predictor
Powered by Artificial Intelligence
AGE
Age
Gender
Chest Pain
Systolic BP
Blood Sugar
PREDICT NOW

Sample UI



CVD Predictor
Home CVD Facts WHO Predition About
PREDICT YOUR LIKELIHOOD OF HAVING
Cardiovascular
Disease Predictor
Powered by Artificial Intelligence
It seems you may need assistance and could be at risk for
CARDIOVASCULAR DISEASE (CVD)
For your safety, please contact your healthcare provider immediately for evaluation and guidance on early management or prevention.
PREDICT AGAIN?

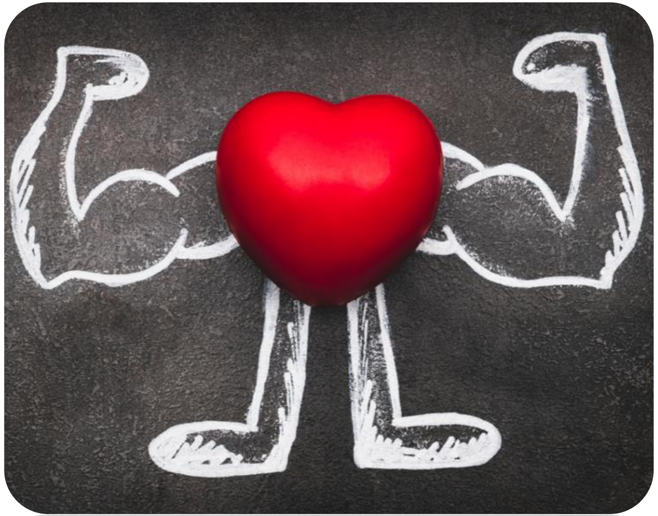


CVD Predictor
Home CVD Facts WHO Predition About
PREDICT YOUR LIKELIHOOD OF HAVING
Cardiovascular
Disease Predictor
Powered by Artificial Intelligence
It appears that you may not be at immediate risk for
CARDIOVASCULAR DISEASE (CVD)
However, for your peace of mind and overall health, it's always a good idea to consult with your healthcare provider for routine check-ups and preventive care.
PREDICT AGAIN?

REFERENCES

[1] "Cardiovascular diseases (CVDs)." Accessed: May 06, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] R. Patel, "Cardio Vascular Diseases Prediction Dataset," Cardio Vascular Diseases Prediction Dataset. Accessed: May 06, 2026. [Online]. Available: https://www.kaggle.com/datasets/riyapatel1697/cdv-healthcare-data
[3] "GridSearchCV," scikit-learn. Accessed: May 06, 2026. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[4] C. N. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Development of a Machine Learning Based Web Application for Early Diagnosis of COVID-19 Based on Symptoms," Diagnostics, vol. 12, no. 4, p. 821, Apr. 2022, doi: 10.3390/diagnostics12040821.
[5] Mohit, "Classification Problem: Relation between Sensitivity, Specificity and Accuracy," Analytics Vidhya. Accessed: May 06, 2026. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/classification-problem-relation-between-sensitivity-specificity-and-accuracy/
[6] "Why joblib: project goals — joblib 1.6.dev0 documentation." Accessed: May 06, 2026. [Online]. Available: https://joblib.readthedocs.io/en/latest/why.html

THANK YOU FOR YOUR ATTENTION !



https://www.fyzical.com/littleton-co/newsmedia/img/14148/Improve_your_Heart_Health_with_Exercise_and_Physical_Therapy.png