

International Journal Research Publication Analysis

Page: 01-13

HEART DISEASE PREDICTION AND PATIENT ANALYTICS: AN INTEGRATED MACHINE LEARNING PIPELINE AND INTERACTIVE DASHBOARD

***Mondrety Trailokya Vaishnavi, Dr. M. Ramchander**

Internship Guide Chaitanya Bharathi Institute of Technology (Autonomous) Gandipet,
Hyderabad, Telangana, India.

Article Received: 06 September 2026

*Corresponding Author: Mondrety Trailokya Vaishnavi

Article Revised: 26 September 2026

Internship Guide Chaitanya Bharathi Institute of Technology (Autonomous)
Gandipet, Hyderabad, Telangana, India.

Published on: 16 October 2026

DOI: <https://doi-doi.org/101555/ijrpa.1363>

ABSTRACT

Patient records held as raw rows in spreadsheets are slow to interrogate, hide the interaction between clinical attributes, and offer no quantified estimate of whether heart disease is present. Manual review of this kind motivates automated, comparable, and visually explorable heart disease analytics. This paper presents an end-to-end, layered analytics framework for heart disease prediction and patient-data exploration. The pipeline ingests a 303-record, 14-attribute patient dataset, validates it with Python and Pandas, performs exploratory analysis with Matplotlib and Seaborn, applies one-hot encoding to eight categorical attributes and standardization to five numeric attributes (yielding 30 model inputs), and trains and compares K-Nearest Neighbors, Decision Tree, and Random Forest classifiers whose hyperparameters are selected by cross-validation. An interactive Power BI dashboard with four KPI cards, four slicers, and four charts exposes the same data to non-programmers. Across the 303 validated records (165 with heart disease, 54.46%; average age 54.37 years), the analysis surfaces three principal observations: (i) chest pain type is strongly associated with disease status, with 82.00% and 79.31% of atypical-angina and non-anginal-pain patients diagnosed against 27.27% of typical-angina patients, as coded in this dataset; (ii) patients with heart disease reach a higher maximum heart rate (158.47 vs. 139.10 bpm) and show lower exercise-induced ST depression (0.58 vs. 1.59) than those without; and (iii) the disease share by gender and age group is heavily shaped by sample composition (96 female vs. 207 male patients; only 11 patients under 30 or over 70) and should not be generalized. The cross-validated accuracies are 84.48% for KNN ($K = 12$), 84.16% for Random Forest (90

estimators), and 78.51% for the Decision Tree (maximum depth 3). Twelve functional test cases pass, and dashboard figures reconcile with the source data. The framework offers a reproducible, low-cost, and transparent template for academic healthcare analytics; it is not a clinical diagnostic tool.

INDEX TERMS: Heart Disease Prediction, Machine Learning, K-Nearest Neighbors, Random Forest, Decision Tree, Cross-Validation, Exploratory Data Analysis, Power BI Dashboard, Healthcare Analytics.

I. INTRODUCTION

A. Background of Cardiac Risk Assessment and Health Analytics

Cardiovascular disease remains one of the leading causes of death worldwide [2], and timely identification of at-risk patients depends on interpreting many demographic and clinical measurements together. Hospital and clinic systems record each patient's age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate, exercise response, and electrocardiographic findings, producing rich, multi-attribute clinical footprints. Systematic transformation of these footprints into visual insights and predicted disease status helps analysts and students understand which factors are associated with disease, how those factors relate to one another, and whether disease presence can be predicted from them collectively [3], [4].

B. Limitations of Manual and Spreadsheet-Based Analysis

In a basic analytical setting, patient records are inspected manually, filtered in spreadsheets, or summarized through isolated queries. Each approach uses the same data but shows recurring structural limitations:

- **Time-Consuming Multi-Attribute Review:** Analysing many patient attributes by hand does not scale as the number of records grows.
- **Hidden Relationships:** Patterns across attributes such as chest pain type, heart rate, and ST depression are difficult to see in raw rows.
- **No Quantified Prediction:** Manual review gives no data-driven estimate of disease presence from the recorded attributes collectively.
- **Limited Interactivity:** Static tables and queries cannot be re-sliced by gender, age group, or chest pain type on demand.

C. Transition to Layered Analytics Frameworks

Modern practice separates data loading, validation, feature preparation, modelling, and presentation into distinct, testable stages. Python, augmented with Pandas [11] and NumPy [12], provides a substrate for loading and encoding data; Matplotlib [13] and Seaborn [14] support exploratory visualization; and scikit-learn [10] supplies preprocessing utilities, classifiers, and cross-validation. Business-intelligence tools such as Power BI [15] extend the resulting analysis to non-programmers through interactive dashboards.

D. Research Challenges in Heart Disease Modeling

- **Small Sample Size:** With 303 records, a single train-test split gives an unstable estimate, which necessitates cross-validation [8].
- **Categorical Codes Stored as Numbers:** Attributes such as cp, restecg, slope, ca, and thal use integer codes that a model may misread as ordered quantities.
- **Heterogeneous Feature Scales:** Age, blood pressure, cholesterol, heart rate, and ST depression span very different ranges, which distorts distance-based classifiers.
- **Evaluation Leakage:** Fitting preprocessing or tuning hyperparameters on the same data used for reporting can inflate accuracy [9].
- **Deployment Friction:** Clinical analysts and students rarely have the programming fluency to query a notebook, so insight stays locked inside code.

E. Overview of Proposed Solution and Key Contributions

This research presents an end-to-end analytics framework that ingests a heart disease dataset, validates and encodes it, trains three classifiers under cross-validation, and exposes the data through an interactive dashboard. Contributions are:

- **Reproducible Layered Pipeline:** A decoupled workflow (data, preprocessing, model, and presentation layers) documented step by step in a single notebook.
- **Feature Preparation for Mixed Data:** One-hot encoding of eight categorical attributes and standardization of five numeric attributes, producing 30 model inputs.
- **Comparative Model Selection:** KNN, Decision Tree, and Random Forest compared on identical processed data, with K, tree depth, and forest size chosen by cross-validation.
- **Empirical Findings:** Quantified associations between disease status and chest pain type, heart rate, ST depression, gender, and age group across 303 records.
- **Interactive Dashboard:** A Power BI dashboard with KPI cards, slicers, and charts whose

figures reconcile with the source dataset.

II. LITERATURE SURVEY

A. Evolution of Heart Disease Prediction

Computational diagnosis of coronary artery disease was formalized by Detrano et al. [3], who evaluated a probability algorithm across international patient cohorts, contributing the clinical data lineage from which widely used heart disease benchmarks descend [17]. Later work applied machine-learning ensembles to such data; Mohan et al. [4] combined random-forest and linear techniques to improve prediction on a heart disease benchmark. These studies concentrate on the classifier itself rather than on an integrated, explorable workflow around it.

B. Classification Algorithms

K-Nearest Neighbors, introduced by Cover and Hart [5], classifies an observation by the majority class of its nearest neighbours and is therefore sensitive to feature scaling. Decision trees, formalized by Quinlan [6], recursively split data on feature conditions and are easy to interpret but prone to overfitting when deep. Random forests, introduced by Breiman [7], average many randomized trees to reduce variance. These three families were selected here because they represent distance-based, rule-based, and ensemble learning respectively.

C. Cross-Validation and Evaluation Leakage

Kohavi [8] showed that k-fold cross-validation gives a more reliable accuracy estimate than a single split, especially on small datasets. Kaufman et al. [9] caution that information from validation data leaking into preprocessing or model selection produces optimistic results. Saito and Rehmsmeier [16] further note that, under class imbalance, precision-recall analysis is more informative than accuracy alone. These cautions directly shape the limitations and future work stated in Sections V and VI.

D. Research Gap Analysis

Three gaps recur: (1) **Model-Only Focus:** classifiers are reported in isolation from exploratory analysis and presentation;

(2) **Reproducibility Gap:** preprocessing and hyperparameter choices are rarely documented step by step for learners; and (3) **Usability Gap:** analytical output remains inside notebooks instead of interactive dashboards usable by non-programmers. This work addresses all three.

Table I: Summary of Representative Related Work.

Ref.	Study Focus	Data Modality	Reported Gap / Limitation
[3]	Probability algorithm for CAD diagnosis	Clinical, multi-site	Diagnostic model without an interactive exploration layer.
[4]	Hybrid ensemble-linear prediction	Heart disease benchmark	Model-centred; no dashboard or step-by-step pipeline.
[5]–[7]	KNN, decision-tree, random-forest theory	General-purpose ML	Algorithm theory without an end-to-end healthcare application.
[9]	Leakage in data mining	General-purpose ML	Methodological guidance without a healthcare pipeline.
[1]	Heart disease internship system	303 patient records	Engineering report; lacked formal peer-reviewed evaluation.

III. SYSTEM METHODOLOGY & ARCHITECTURE

A. Proposed System Overview and Feasibility

The proposed system replaces manual, spreadsheet-based inspection with a structured analytical pipeline. Feasibility was verified across three dimensions. **Economic:** the implementation relies entirely on open-source tooling (Python, scikit-learn, Jupyter) and a free desktop edition of Power BI, incurring no licensing cost in the academic setting. **Technical:** with only 303 records, preprocessing, cross-validation, and dashboard interaction complete quickly on a standard workstation (Intel Core i5, 8 GB RAM). **Operational:** the dashboard abstracts code from analysts, who explore plain-language slicers and charts.

B. Layered Data Processing Pipeline

The framework is organized into four cooperating layers: (1) *Data Layer*: loads heart.csv, validates shape, data types, and missing values, and performs exploratory analysis; (2) *Preprocessing Layer*: one-hot encodes categorical attributes and standardizes numeric ones; (3) *Analytics and Model Layer*: sweeps hyperparameters and evaluates KNN, Decision Tree, and Random Forest by cross-validation; and (4) *Presentation Layer*: reports model accuracy and renders the interactive Power BI dashboard.

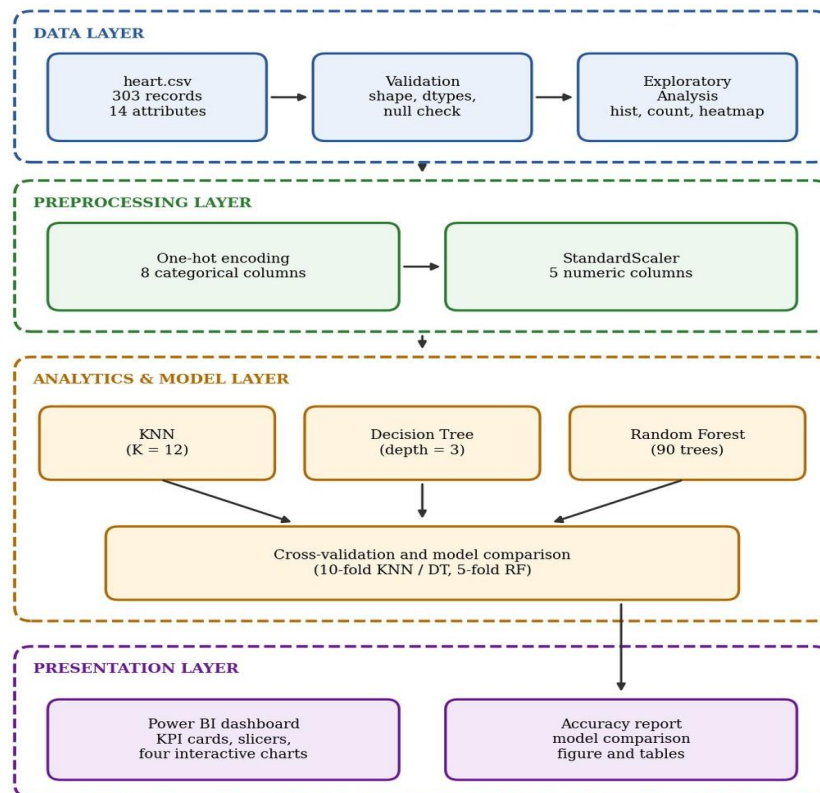


Fig. 1. Layered system architecture of the heart disease analytics framework.

Table II: Core Dataset Schema and Analytical Roles.

Attribute	Description	Type / Values
age	Age of the patient	Numeric (29–77)
sex	Gender	Binary (1 = male, 0 = female)
cp	Chest pain type	Categorical (0–3)
trestbps / chol	Resting blood pressure / serum cholesterol	Numeric
fbs	Fasting blood sugar	Binary (0/1)
restecg	Resting ECG result	Categorical (0–2)
thalach	Maximum heart rate achieved	Numeric
exang	Exercise-induced angina	Binary (0/1)
oldpeak	ST depression induced by exercise	Numeric (decimal)
slope / ca / thal	Peak-exercise ST slope / major vessels / thalassemia measure	Categorical (0–2 / 0–4 / 0–3)
target	Presence of heart disease	Binary (1 = disease, 0 = none)

C. Data Validation, Preprocessing and Feature Transformation

The dataset contains 303 patient records and 14 columns: 13 attributes stored as integers, oldpeak as a float, and the binary target. The check `df.isnull().any()` returns False for every column and `info()` shows 303 non-null entries per column, so no imputation or row deletion is

required. Eight attributes (sex, cp, fbs, restecg, exang, slope, ca, thal) are converted into separate binary columns by `get_dummies()`, so that integer codes are not misread as ordered quantities. The five numeric attributes (age, trestbps, chol, thalach, oldpeak) are standardized with `StandardScaler`, which matters especially for the distance-based KNN classifier. The result is 30 input columns and the target. Algorithm 1 summarizes the procedure.

Algorithm 1: Data Preparation, Model Selection & Evaluation Pipeline

Input: Raw patient CSV Draw(303 rows, 14 columns)

Output: Feature matrix X (30 columns), target y, CV accuracy per model

```

1: Load Draw into DataFrame df via read_csv()
2: Assert df.isnull().any() is False for every column
3: dataset ← get_dummies(df, [sex, cp, fbs, restecg, exang, slope, ca, thal])
4: Standardize [age, trestbps, chol, thalach, oldpeak] with StandardScaler
5: X ← dataset.drop(target); y ← dataset[target]
6: for each model M in {KNN, DecisionTree, RandomForest} do
7:   for each hyperparameter value h in range(M) do
8:     s[h] ← mean(cross_val_score(M(h), X, y, cv))
9:   end for; select h* ← argmax s[h]
10: end for
11: return X, y, and cross-validation accuracy of M(h*) for each M

```

D. Classification Modeling and Cross-Validation

Three supervised classifiers are trained on the same processed matrix. For **KNN**, K is swept from 1 to 20 with 10-fold cross-validation and K = 12 is selected. For the **Decision Tree**, maximum depth is swept from 1 to 10 with 10-fold cross-validation and depth 3 attains the highest score. For the **Random Forest**, the number of trees is varied from 10 to 100 in steps of 10 with 5-fold cross-validation and 90 estimators are used. The evaluation metric is mean cross-validation accuracy over the validation folds; the selected configurations are given in Table VI.

E. Interactive Power BI Dashboard

The Heart Disease Analysis Dashboard (“Interactive Patient Health & Disease Insights”) is built on one report page. The coded numeric columns are exposed as readable fields: *Gender* (from sex), *Chest Pain Type* (from cp), *Disease Status* (from target), and *Age Group* (from age). Four measures (Total Patients, Heart Disease Patients, Disease Percentage, Average Age) drive KPI cards, and four slicers filter every visual: clustered columns for gender and chest pain type, a scatter chart of age against cholesterol coloured by disease status, and a stacked column chart by age group.

IV. EXPERIMENTAL EVALUATION & RESULTS

A. Analytical Baseline Comparison

Benchmarking the proposed pipeline against manual spreadsheet review (Table III) shows that automated, cross-validated modelling combined with an interactive dashboard handles multi-attribute patterns, quantified prediction, and repeatable exploration that a manual workflow cannot.

Table III: Framework Benchmark vs. Manual Review Workflow.

Dimension	Manual Review	Proposed Framework
Attribute handling	One or two attributes at a time	All 30 encoded features jointly
Prediction	None	Three cross-validated classifiers
Exploration	Static tables and queries	Slicer-driven interactive dashboard
Scalability	Degrades as records grow	Vectorized Pandas / scikit-learn
Evaluation	None	10-fold and 5-fold cross-validation

B. Overall Portfolio Key Performance Indicators

The dataset contains 303 patients, of whom 165 have heart disease (54.46%) and 138 do not, so the two target classes are reasonably balanced. The average age is 54.37 years (range 29–77). Numeric summaries show wide spread in cholesterol (mean 246.26, maximum 564) and resting blood pressure (mean 131.62, maximum 200), confirming the need for scaling before distance-based modelling.

C. Chest Pain Type, Gender and Age: Segment-Level Discoveries

Chest pain type is among the strongest discriminators of disease status. Patients coded as typical angina form the largest group

(143) yet only 27.27% have heart disease, whereas 82.00% of atypical-angina and 79.31% of non-anginal-pain patients do (Table IV). This ordering is consistent with the positive correlation (0.43) between cp and the target, and it follows the integer coding supplied with the dataset, which may not match clinical labelling conventions.

Table IV: Disease Rate by Chest Pain, Gender and Age Group.

Segment	Patients	Disease Rate
Typical angina (cp = 0)	143	27.27%
Atypical angina (cp = 1)	50	82.00%
Non-anginal pain (cp = 2)	87	79.31%
Asymptomatic (cp = 3)	23	69.57%
Male	207	44.93%
Female	96	75.00%
Age 30–39	15	73.33%
Age 40–49	72	69.44%
Age 50–59	125	52.00%
Age 60–69	80	40.00%
Age 70 and above	10	60.00%

The disease share is higher among female patients (75.00%) than male patients (44.93%), and highest in the younger age bands while lowest at 60–69. These patterns must be read with care:

the sample holds only 96 female patients, the under-30 group has a single record, and the 70+ group has ten, so percentages for small groups are not reliable and the observations describe this sample rather than clinical populations.

D. Clinical Attributes and Correlation Effects

Patients with heart disease in this dataset tend to be younger, reach a higher maximum heart rate, and show lower ST depression than patients without (Table V). Fig. 2 shows that the attributes most positively correlated with the target are chest pain type (0.43), maximum heart rate (0.42), and ST slope (0.35), while exercise-induced angina (−0.44), ST depression (−0.43), number of major vessels (−0.39), and thalassemia measure (−0.34) are negatively correlated. Cholesterol (−0.085) and fasting blood sugar (−0.028) are only weakly correlated with the target here.

Table V: Average Clinical Attributes by Disease Status.

Attribute	No Disease	Heart Disease
age (years)	56.60	52.50
trestbps (resting BP)	134.40	129.30
chol (cholesterol)	251.09	242.23
thalach (max heart rate)	139.10	158.47
oldpeak (ST depression)	1.59	0.58

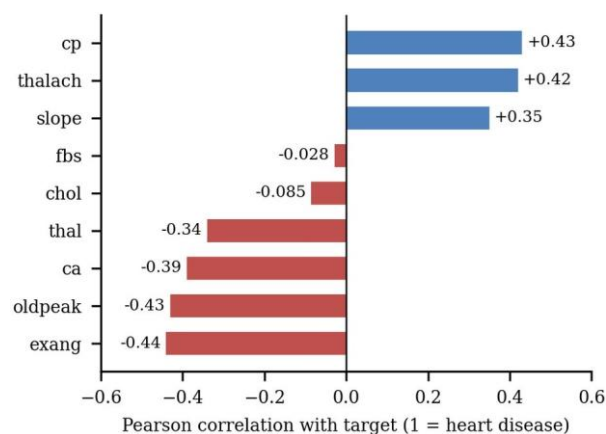


Fig. 2. Correlation of selected attributes with the disease target.

E. Model Comparison

All three classifiers separate the target with reasonable accuracy (Table VI, Fig. 3). KNN (K = 12) achieves the highest cross-validation accuracy at 84.48%, closely followed by the Random Forest (90 estimators) at 84.16%, while the depth-3 Decision Tree reaches 78.51%. The 0.32-point gap between KNN and Random Forest is small and the two models are evaluated with different fold counts (10 vs. 5), so they should be regarded as comparable rather than clearly ranked. The shallow tree performs

several points lower on this dataset.

Table VI: Model Comparison. (Cross-Validation Accuracy)

Model	Selected Parameter	Folds	Accuracy
KNN	K = 12 (tested 1–20)	10	84.48%
Random Forest	90 trees (tested 10–100)	5	84.16%
Decision Tree	Depth = 3 (tested 1–10)	10	78.51%

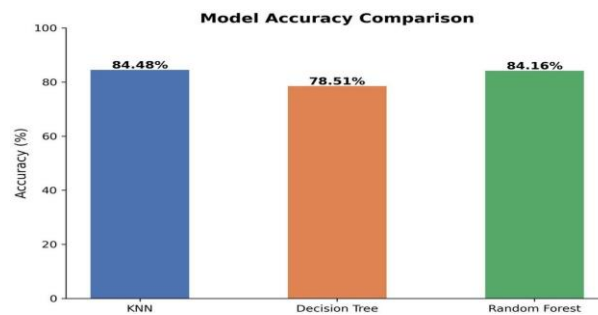


Fig. 3. Cross-validation accuracy of the three classifiers.

F. System Testing and Verification

Twelve functional test cases (TC01–TC12) covered dataset loading (303 rows, 14 columns), missing-value checks, data types (13 integer and one float column), one-hot encoding (30 input columns), standardization of the five numeric attributes, separation of X and y, the three hyperparameter sweeps and evaluations, dashboard KPI values (303 / 165 / 54.46% / 54.37), and the four slicers. All passed. The KNN result was independently re-run on the same processed data and reproduced exactly; the Decision Tree and Random Forest carry no fixed random seed, so their scores can vary slightly between runs. Dashboard gender, chest pain, and age-group counts match counts computed directly from the dataset.

V. DISCUSSION & ANALYTICAL IMPLICATIONS

A. Practical Operational Contrast

Compared with manual spreadsheet inspection, the framework analyses all attributes jointly, produces a quantified prediction of disease presence, and converts a static file into an interactive dashboard that non-programmers can explore. Its outputs are explanatory in terms an analyst already uses (gender, chest pain type, age group), not opaque scores.

B. Four-Point Analytical Takeaways

- 1. Prepare Mixed Data Deliberately:** encode integer-coded categoricals and scale numeric attributes; both choices are decisive for distance-based KNN.
- 2. Tune by Cross-Validation:** on 303 records, sweeping K, depth, and forest size with k-

fold cross-validation uses every record for training and validation.

3. Compare Rather than Presume: a simple KNN and a Random Forest performed comparably (84.48% vs. 84.16%), while a depth-3 tree lagged at 78.51%.

4. Read Segment Rates with Sample Size in Mind: gender and age-group rates reflect sample composition and small cells, and must not be generalized clinically.

C. Limitations

Limitations include: (i) a small dataset of 303 records that may not generalize to other populations; (ii) only cross-validation accuracy is reported, with no precision, recall, F1-score, ROC-AUC, or confusion matrices; (iii) the same cross-validation used to choose hyperparameters also reports accuracy, with no held-out test set, so the figures may be slightly optimistic; (iv) feature scaling is applied to the whole dataset before cross-validation, which lets validation-fold information influence the scaler [9]; (v) the Decision Tree and Random Forest are unseeded and neighbouring hyperparameter scores differ by only a few tenths of a point; and (vi) the system is an analytical notebook and dashboard, not a deployed application, and is not intended for clinical decision-making.

VI. CONCLUSION AND FUTURE WORK

This paper presented an end-to-end, layered framework for heart disease prediction and patient-data analytics. By combining data validation, one-hot encoding and standardization, three cross-validated classifiers, and an interactive Power BI dashboard, the workflow bridges raw patient records and explorable, quantified insight. Evaluated on 303 records (54.46% with heart disease), KNN attains 84.48% cross-validation accuracy, Random Forest 84.16%, and the Decision Tree 78.51%, while segment analysis highlights chest pain type, maximum heart rate, and ST depression as the attributes most associated with disease status. The system is intended for academic and analytical purposes and is not a substitute for professional medical diagnosis. Future extensions will incorporate: (1) larger, more diverse healthcare datasets; (2) precision, recall, F1-score, ROC-AUC, and confusion matrices [16], with a separate hold-out set or nested cross-validation and scaling inside a pipeline to avoid leakage; (3) Logistic Regression, Support Vector Machines, and gradient-boosted models (XGBoost, LightGBM) with systematic tuning such as GridSearchCV; (4) SHAP or feature-importance explanations; and (5) a web interface for patient-level prediction and publication of the dashboard to the Power BI Service with scheduled refresh.

ACKNOWLEDGMENT

The author expresses sincere gratitude to the MCA Department, Chaitanya Bharathi Institute of Technology (Autonomous), Gandipet, Hyderabad, to internship guide Dr. M. Ramchander for guidance and academic support, and to Pantech e Learning Pvt. Ltd., and Mr. M. Bharath, under whose Data Analytics internship this work was carried out.

REFERENCES

1. M. T. Vaishnavi and M. Ramchander, "Heart Disease Prediction," *Technical Internship Report*, MCA Department, Chaitanya Bharathi Institute of Technology (Autonomous), Gandipet, Hyderabad, Telangana, India, 2026.
2. World Health Organization, "Cardiovascular diseases (CVDs)," fact sheet. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
3. R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
4. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.
5. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
6. J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
7. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
8. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.
9. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, art. 15, 2012.
10. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
11. W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 56–61.
12. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.

13. J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.
14. M. L. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021.
15. Microsoft, "Power BI documentation." [Online]. Available: <https://learn.microsoft.com/power-bi/>
16. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.
17. Dua and C. Graff, "UCI Machine Learning Repository: Heart Disease Data Set," Univ. California, Irvine, 2019. [Online]. Available: <https://archive.ics.uci.edu/>