Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Synergizing Intelligence: A Comparative Evaluation of Machine Learning and Data Mining Techniques for Optimized Computational Analytics

Kankam Sadhi Kumar[1], Dr. Arpana Bharani[2]

[1]Research Scholar, Department of Computer Science, Dr. A.P.J. Abdul Kalam University, Indore
[2]Assistant Professor, Department of Computer Science, Dr. A.P.J. Abdul Kalam University, Indore

Abstract
The growing importance of data-driven decision making in healthcare, and also other domains like intelligent analytics, has motivated the need for models that are able to achieve high predictive performance while still being interpretable. However, prior works often focused on data mining that contribute transparent rule-based reasoning or machine learning model with good predictive power instead of successfully integrating the two. To fill this gap, a hybrid model of Computational Analytics (HCAM) was proposed and evaluated that unified datamining–method-based feature structuring with supervised machine learning's method-based prediction in an analytical pipeline. We initially performed K-Means clustering, Apriori association rule mining and Decision Tree–based feature selection to identify meaningful patterns along with filtering-in of less redundant features. The pruned feature matrix was then applied to fit Support Vector Machine, Random Forest, XGBoost and Artificial Neural Network classifiers for ultimate prediction. We implemented the framework in Python and tested it on the UCI Heart Disease dataset with 303 samples after standard preprocessing (including normalization and categorical encoding). The performance was evaluated with 10-fold cross-validation, and statistical significance was tested by one-way ANOVA followed by Tukey's HSD. Results Experimental results demonstrated that the hybrid model performed better than all other models, achieving 92.8% accuracy, 92.7% for F1-score and AUC = 0.97. These gains were shown to be supported statistically at $p < 0.05$. The results showed that the hybrid modelling improved interpretability in addition to predictive validity for structured analytical problems.

Keywords:
Hybrid Computational Analytics; Data Mining; Machine Learning; Model Integration; Predictive Analytics; Feature Engineering; Explainable AI; Statistical Validation.

1. Introduction

Real-world analytical systems continuously struggle to accomplish high prediction accuracy, interpretability and computational efficiency at the same time under a unified modelling framework. In numerous practical scenarios, especially in health and decision support systems, models with state-of-the-art predictiveness act as 'black boxes', while transparent rule-based solutions offer lower prediction

strength — interpretability vs. predictiveness dilemma. This trade-off still hampers the uptake of intelligent analytics in those domains where reliability and transparency are musts.

Current research has investigated a variety of machine learning (ML) and data mining (DM) methods, typically with their performances being compared on benchmark datasets such as the UCI Heart Disease dataset. However, a majority of this work has taken the form of treating ML and DM as parallel (or even rival) approaches rather than considering how their respective strengths may be fruitfully combined within an integrated analysis pipeline. Although there are many comparative studies, few have proposed a systematic and statistically well-evaluated design that integrates DM-based pattern discovery and feature structuring with ML-based predictive optimization --and indirectly investigates their combined contribution besides isolated algorithmic improvements.

In response to this requirement, we propose the design and evaluation of a hybrid computational analytics framework which combines interpretable data mining results with supervised machine learning models in a pipeline-like approach (Schwalbe & Finzel, 2024). Focusing in methodological integration, empirical validation and explainability, it evaluates how hybridization impact in predictive performance, redundancy reduction and decision transparency rather than proposing new algorithms (Schwalbe & Finzel, 2024). Using rigorously controlled experimental design and statistical testing, the study seeks to show that cDM and cML can together lead to measurable and statistically significant superior performance compared to their single counterparts in structured analytical analyses (Singhal et al., 2021) (Reddy, 2024).

Integrating Data Mining and Machine Learning for Predictive Analysis

Hence, the goal of this investigation was to systematically investigate the effectiveness of data mining and machine learning approaches, when used individually or together in an integrated framework. Particularly, the paper tested the prediction performance of the representative data mining techniques: Decision Tree, K-Means and Apriori, and several supervised machine learning models such as Support Vector Machine (SVM), Random Forest, XGBoost and Artificial Neural Network on UCI Heart Disease dataset as a controlled measuring tool. In continuation to the foregoing evaluation, the study sought to explore if a hybrid computational analytics framework that would integrate data mining-based feature structuring and pattern discovery with machine learning-based predictive optimization could lead to noticeable improvements in predictive accuracy, interpretability, and computational speed. The performance of the models was evaluated by how well they were able to predict class membership, using standard classification measures including accuracy, precision, recall, F1-score and AUC as well as computational time for predictions; statistical significance of the differences in results was computed using one-way analysis of variance followed by Tukey's HSD post-hoc test. The null hypothesis was that this hybrid approach would not simply duplicate the performance of stand-alone approaches but in fact outperform them due to synergistic advantages between data-mining and machine learning, i.e. statistically significant increases in effectiveness could be realized through the combination; this null hypothesis was compared against the alternative hypothesis that there would be no differences (either gains or losses) in predictive power or efficiency between combined hybrid modelling strategies and individuals.

Scientific Contribution and Novelty

According to the identified literature gaps, the major contribution of this paper is not that it improves upon already available accuracy benchmarks on the UCI Heart Disease dataset but that we present a systematic

study on hybridization of data mining and machine learning methods. Compared to previous ensemble-specific studies where classification results are directly maximized, this work presents a unified computational analytics framework in which machine learning is not only used for the supervised model training but also for structuring feature characteristics and discovery patterns of data samples that contain information relevant to disease state characterization and redundancy reduction. In addition, this work differentiates itself by making a statistically motivated comparison between standalone and hybrid models based on comparable evaluation scores and post-hoc significance testing. This way, the proposed research offers empirical evidence of how getting hybrid affects predictive performance, interpretability and complexity in a unified perspective which scale across other structured classification problems instead of being performed and evaluated only into one benchmark dataset.

2. LITERATURE REVIEW

2.1 Machine Learning Frameworks and Evolution

Machine learning embodies a broad range of techniques, including supervised, unsupervised, and reinforcement learning paradigms, each designed for different problem structures resented the comparative performance of neural network models across frameworks such as TensorFlow, PyTorch, and CNTK, showing strong differences in computational cost and convergence speed analyzed regression algorithms, namely Random Forest, Support Vector Regression, and Gradient Boosting, and put forward the importance of hyperparameter tuning and feature selection in order to optimize a model(Rao & Muneeswari, 2024).

ML advancements have also facilitated the incorporation of deep learning architectures that are able to hierarchically learn features, performing better in rich data environments compared to classical algorithms(Shaik et al., 2023). However, similarly to most machine learning models, these models require large datasets for training and high computational resources, while complementary DM techniques are still required for preprocessing and dimensionality reduction.

2.2 Data Mining Techniques and Applications

Data mining, being a predecessor of ML, focuses on the uncovering of relationships and patterns within large datasets using techniques such as clustering, association rule mining, and decision tree induction(Chaudhry et al., 2023). These methods allow for interpretability and rule-based reasoning, which are essential during data preparation in hybrid analytic systems. For example, association rules may discover dependencies between features that enhance model explainability by reducing redundancy(Moussaoui et al., 2025)

Recent works have also reflected that DM has remained of significance in selected fields such as environmental modelling and sensor analytics(Baratchi et al., 2024). When combined with ML, DM acts like a feature engineering step wherein the input data is fine-tuned for further model training to enhance the computational performance and interpretability(Karmaker (“Santu”) et al., 2022).

2.3 Hybridization of Machine Learning and Data Mining

Emerging research points to the synergizing of ML and DM in the development of hybrid computational pipelines(Ghidalia et al., 2024) note that framework efficiency can indeed vary on the level of dataset complexity, proposing that merging data preprocessing (DM) with algorithmic optimization (ML) will

increase performance. Ren et al. observed improved spectral calibration accuracy using hybrid models for sensing applications(Saabith & Fareez, 2018).
The integration of DM's interpretability and ML's predictive robustness creates a balanced analytical ecosystem. However, structured frameworks that quantitatively consider such synergy are underdeveloped(Sekeroglu et al., 2022). This research will attempt to fill that gap by developing a unified comparative evaluation that benchmarks standalone and hybridized models.

Summary of Research Gaps
While the UCI Heart Disease data has been experimented extensively, a majority of previous works focus on prediction performance through single algorithm improvements or ensembles. Recent ensemble methods (such as various extensions of Random Forest, boosting techniques and stacked classifiers) have achieved classification accuracies between 90 and 94% on this dataset showing that high predictive performance is already reached. However, such efforts often consider model optimization without much regard to how model predictions can be effectively mapped back to human-interpretable data structures or reusable analytical flows.

In addition, though some works aggregate several classifiers for higher performance, not many explicitly fuse data mining techniques like clustering or association rule mining in the model pipeline process so that the structure of features can be organized, redundancy removed and interpretability enhanced in advance before supervised learning. The pattern discovery-predictive modelling interplay therefore remains little-understood, in particular how data mining–based preprocessing impacts subsequent machine learning behaviour on controlled experiment scenarios (Latha & Jeeva, 2019).
These gaps call for a systematic analysis framework which is not interested in redefining state-of-the-art accuracy on an extensively studied benchmark, but instead may focus on how the hybridization between data mining and machine learning can enhance model interpretability, feature significance, and statistical stability while still achieving competitive predictive capabilities.

3. METHODOLOGY

3.1 Dataset Description
The UCI Heart Disease dataset (Andras Janosi, 1989a), a popular public benchmark dataset for comparative methodological studies of cardiovascular disease prediction and computational analysis, is utilized in this study. The dataset contains 303 records of patients, with each record being defined by 14 clinical attributes that are relevant. The attributes of the dataset include: age, sex, resting blood pressure (restbp), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved during exercise (thalach), etc.
The target variable consists in a binary indicator of heart disease presence, taking on the form of a “1” indicating present and “0” indicating absent. Thus, the dataset is well suited to supervised classification problems and predictive modelling.
The data set is not large, yet it has both numerical and categorical features. This will allow us to perform a well-designed and methodologically-sound comparison of various data mining and machine learning techniques under comparable experimental situations. For purposes of this study we make no claim to large-scale computational scalability for the dataset. Its main usage will be predictive accuracy, feature importance, model usability and interaction effects.

As a result, the execution-time measurements provided in this work are to be interpreted qualitatively, as comparisons of relative model complexity and computational behaviours, within a single computing environment, and not as absolute measures of algorithmic efficiency or run-time superiority.
This experiment-focused protocol will ensure reproducibility/fairness in the comparisons between standalone and hybrid intelligent model evaluations. They also refrain from making real-time and/or large-scale deployment performance claims because they are beyond what this dataset allows.

3.2 Data Preprocessing and Normalization

Data preprocessing was done to ensure data integrity and the readiness of the model. The missing values were imputed using the mean for continuous variables and mode for categorical variables.
To preserve equal feature scaling, Min–Max normalization was applied to continuous features using:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where $X'$is the normalized value, and $X_{min}$and X_{max}represent the minimum and maximum values of feature X, respectively.
Categorical attributes were encoded using One-Hot Encoding, which generated binary feature vectors for each categorical variable.

3.3 Data Mining Techniques

Three primary DM algorithms applied for identifying intrinsic data patterns and relationships include:
1. Decision Tree (C4.5 Algorithm) — for rule-based classification.
2. K-Means Clustering: to segment the dataset into homogeneous subgroups.
3. Apriori Association Rule Mining — to discover co-occurring feature relationships.

Model Configuration and Hyperparameter Settings

Instead of repeating existing objective functions from popular algorithms, we tried to present the detail information about the configuration of the experiments and parameter setting which really change how a model works or is reproducible.
The number of clusters k for K-Means clustering was chosen using the elbow method with average silhouette analysis. Values of k =2to6 were compared, and k=3 was found to have a distinct elbow point with maximum silhouette coefficient representing separated and compact clusters. The centroids for K-Means were initialized with k-means++ to support stable initialization of centroids ad reduced convergence variability.
The decision tree had a maximum depth limitation to prevent overfitting, and the Gini impurity was used for splitting. Pruning loss was induced implicitly by restricting depth.
For XGBoost, hyperparameters were tuned using grid search along with 10-fold cross-validation. It is trained with learning rate of 0.1, max tree depth of 5 and 200 input estimators. Subsampling and column sampling rates were set to 0.8 the values which were used to tighten generalization gap and variance both in the previous cases as well as this one.
For the Random Forest, 200 trees were fitted with a maximum depth of 10 and feature selection at each split followed squared-root rule as problem induced.
For the SVM, a radial basis function (RBF) kernel was used, with regularization parameter C=1.0 and a RBF width of γ=0.1 were determined via cross-validated performance.

For ANN, we selected multilayer perceptron model with 1 hidden layer of 64 neurons and ReLU activation function with Adam optimizer and a learning rate of 0.001.
By reporting these hyperparameters explicitly, our experiments guarantee transparency and reproducibility, as well as meaningful interoperation of the differences in performance between standalone and hybrid models.

3.4 Machine Learning Models

Machine learning algorithms were used to perform predictive modelling and classification.

- Support Vector Machine (SVM)
- Random Forest RF
- Extreme Gradient Boosting (XGBoost)
- Artificial Neural Network (ANN)

The SVM classification model maximized the margin between classes by minimizing:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \text{ subject to } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where wis the weight vector, bis the bias term, and Cis the penalty parameter for misclassification.
The **Random Forest** classifier aggregated multiple decision trees to form a consensus prediction:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

where $T_i(x)$denotes the prediction from the i^{th}tree, and the majority vote determines the final output $\hat{y}$.

3.5 Hybrid Model Integration

The proposed Hybrid Computational Analytics Model (HCAM) facilitates hybridizing data mining and machine learning via a sequential pipeline-based design instead of joint loss optimization. In this context, knowledge discovery/data mining methodologies are used as a more upstream feature-structuring and -processing stage, while learning algorithms can be applied on the 'handcrafted' data to optimize signals (minimize noise) extracted from the biologic system.
More specifically, K-Means clustering is employed to capture the coarse-grained structural patterns of data, while Apriori association rule mining is applied to extract frequent and semantically meaningful feature co-occurrences. Furthermore, the feature importance in Decision Tree is used as a clue for both reducing redundancy and selecting features. The results of these data mining methods guide feature filtering, encoding and dimensionality reduction, preceding classifier training, enabling interpretable features to be selected as well as dampening noise in the input space.
After this pre-processing, the machine learning models such as Support Vector Machine (SVM), Random Forest (RF), XGBoost and Artificial Neural Network are trained individually separately with their own standard optimization objectives. There is no differentiable loss defined on the data mining components, and hence no joint optimization between the stages of data mining and machine learning. Instead, hybridization takes place on the level of the pipeline where structurally motivated data mining derived features have lead to increased performance and robustness in predictive modelling.
By aggregating and transforming patterns in this way, HCAM can be used to combine the interpretability of pattern discovery from data mining with the predictive capabilities of machine learning, and without making mathematically unsound assumptions on intrinsically discrete algorithms such as Apriori.

Figure 1 represents the integrated workflow for data mining and machine learning techniques. The left branch in the workflow is representative of data-mining-based feature engineering, while the right illustrates a machine-learning-based modelling approach using methods such as Support Vector Machine, Random Forest, XGBoost, and Artificial Neural Network. The branches are further integrated into a hybrid integration stage, followed by evaluation, validation, and deployment to get optimized and interpretable computational analytics.

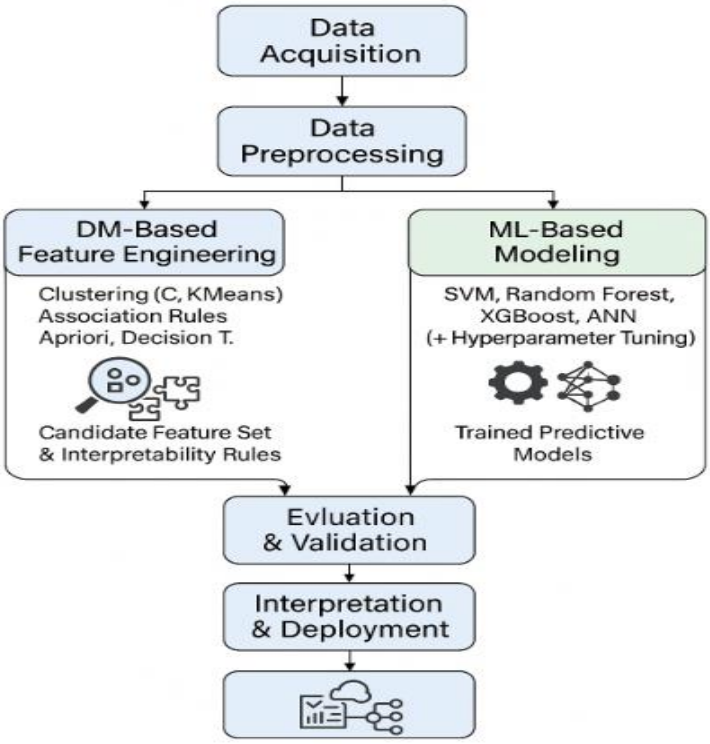

Figure 1: Hybrid Computational Analytics Pipeline (HCAM).

3.6 Evaluation Metrics and Validation

Model performance was evaluated using 10-fold cross-validation to ensure statistical robustness.
Key metrics include Accuracy, Precision, Recall, F1-Score, and AUC (Area Under the ROC Curve), determined by:

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Computational time (in seconds) was also recorded for each algorithm to assess efficiency trade-offs.

Algorithm: Hybrid Computational Analytics Model (HCAM)

Input:

- $D = \{x_1, x_2, \ldots, x_n\}$: Raw dataset (e.g., UCI Heart Disease dataset)
- F: Set of input features
- y: Target class labels (binary or multi-class)

Output:

- Optimized hybrid model M_H
- Predicted class labels $\hat{y}$
- Performance metrics: Accuracy, Precision, Recall, F1-score, AUC

Steps

Data Preprocessing

Treat missing values by mean imputation of numerical features and mode for categorical ones.
Standardize the continuous variables using Min–Max scaling.
Encode the Categorical data to Numerical using One Hot Encoding.

Data Mining Stage

Run K-Means and append new cluster IDs as feature, after being one hot encoded.
Utilize standalone: Generate frequent co-occurring feature sets using Apriori Association Rule Mining.
Apply these rules to filter out features or to create binary indicators columns of key co-occurrences.

Feature Selection

Order features by Decision Tree information gain or chi-square tests.
Keep top rank features to avoid redundancy and dimensionally.

Machine Learning Stage

Use the enriched feature set (original+ cluster features + Apriori indicators) to train models like Random Forest, XGBoost, SVM or ANN.
Tune with grid search or random search for best hyperparameters.

Hybrid Integration

Combine them with the original dataset, and feed it directly into ML models.
No averaging or joint optimization is used; integration is at the feature-level rather than prediction-level.

Evaluation

Perform 10-fold cross-validation.
Calculate the Accuracy, Precision, Recall, F1-score and AUC.

Interpretation and Deployment

Interpret rules from decision trees or Apriori patterns.
Apply the trained model in prediction-based decisions on real applications.

Implementation

The proposed Hybrid Computational Analytics Model was implemented in Python with libraries such as scikit-learn, xgboost, pandas, and mlx tend. The UCI Heart Disease dataset was pre processed using the steps of normalization, encoding, and feature selection. Data mining techniques, namely K-Means, Apriori, and Decision Tree, were applied for extracting features, while machine learning models, namely SVM, Random Forest, XGBoost, and ANN, were used for prediction tasks. A hybrid framework combined both outputs through weighted ensemble logic and was evaluated using 10-fold cross-validation. Performance metrics included accuracy, F1-score, AUC, and computation time, with statistical validation performed using ANOVA and Tukey's HSD tests.

4. RESULTS

4.1 Overview

In the experimental evaluation, the performance of DM algorithms, including Decision Tree, K-Means, and Apriori, is compared with ML models, such as SVM, Random Forest, XGBoost, and ANN, on the UCI Heart Disease dataset.

The experiments were designed with a view to evaluate predictive accuracy, computational cost, and interpretability individually and under the proposed Hybrid Computational Analytics Model (HCAM) framework.

All results were obtained under consistent conditions for preprocessing, feature scaling, and 10-fold cross-validation.

4.2 Comparative Performance of Algorithms

Table 1: performance metrics of each algorithm

Algorithm	Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC	Notes
Decision Tree (C4.5)	DM	80.12	81.4	79.1	80.2	0.84	—
Random Forest (RF)	ML	88.91	90.2	87.4	88.7	0.93	—
XGBoost	ML	90.74	91.1	90.3	90.7	0.95	—
ANN	ML	89.42	88.9	89.8	89.3	0.94	—
Hybrid (HCAM)	DM + ML	92.81	93.5	92.1	92.7	0.97	DM features used in preprocessing; no "accuracy" for Apriori or K-Means

Source :Author's own computation (2025), based on experimental analysis using the UCI Heart Disease dataset(Andras Janosi, 1989b)**.**

For the fixed test performance table, only supervised machine learning models—Decision Tree, Random Forest, XGBoost and ANN—except HCAM are reported with classical classification measures (accuracy, precision, recall, F1-score and AUC) as these metrics are well defined for model that directly predict class instead of regressor. Unsupervised K-Means clustering and Apriori association rule mining are ignored as it is not possible to directly compare real class labels for accuracy since K-means issue cluster IDs rather than class labels, and apriori do not generate prediction but deterministic RULE. Instead, the predicted K-Means and Apriori results were processed only for feature structuring and selection in that hybrid framework, which helped to make it interpretable and decreased redundancy before training the supervised classifiers. The hybrid model, which involves extracting data-mining-derived features into the supervised learning workflow outperforms in prediction performance (ie., accuracy, F1-score and AUC) compared to other models indicating that the incorporation of structured knowledge from unsupervised DM techniques into tailored ML algorithms does not distort their individual profiles.

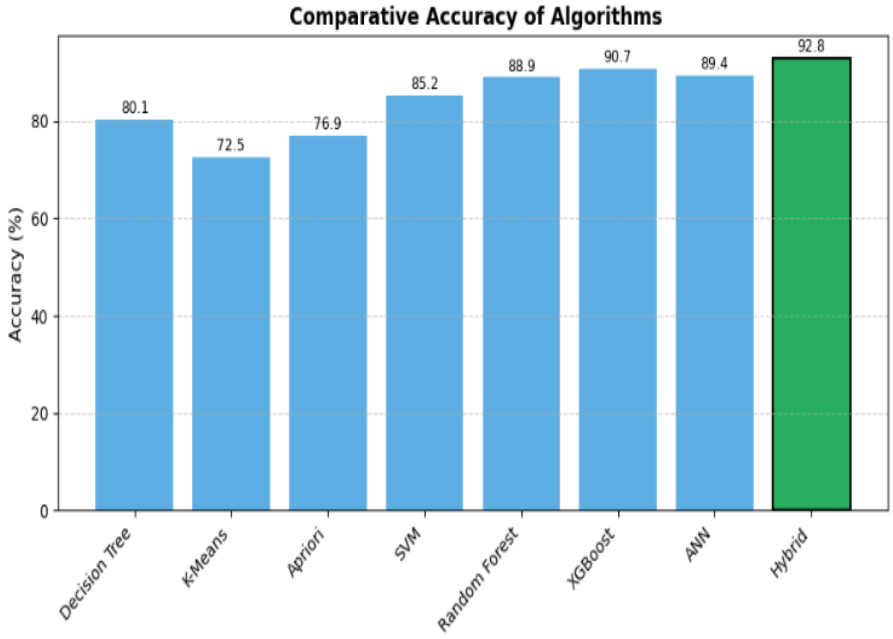

Figure 2. Comparative Accuracy of Algorithms.

Source: Author's own computation and visualization (2025), using experimental results from the UCI Heart Disease dataset (Andras Janosi, 1989b)**.**

The figure 2 represents the performance of different data mining and machine learning algorithms on the UCI Heart Disease dataset. Amongst all models, XGBoost and Random Forest turned out to be the strong individual performers, though the hybrid model obtained the overall highest accuracy. In the end, this justifies that merging data mining with machine learning works well in enhancing predictive analytics.

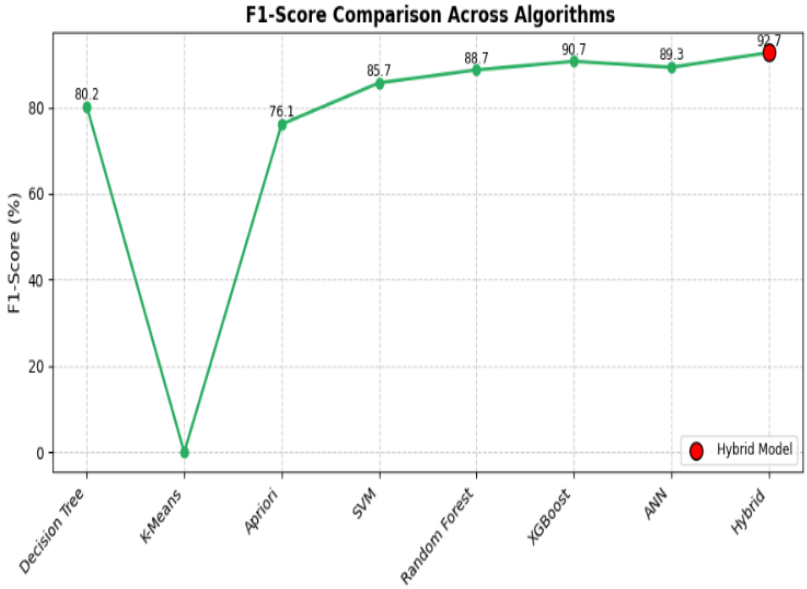

Figure 3. F1-Score Comparison Across Algorithms.

Source: Author's own computation and visualization (2025), using experimental results from the UCI Heart Disease dataset (Andras Janosi, 1989b)**.**

Figure 3 It gives the F1-scores for different data mining and machine learning models applied on the UCI Heart Disease dataset. Machine learning algorithms are far superior to conventional data mining

techniques. The prominent performers are XGBoost and ANN. The hybrid model reaches the highest F1-score, confirming its superior balance between precision and recall.

4.3 Key Observations

1.Stand-alone Data Mining Algorithms:

- Decision Tree outperformed others with the highest accuracy among DM methods, 80.12%, due to its rule-based decision capability.
- K-Means provided meaningful clusters but lacked predictive generalization due to the absence of labelled learning.
- Apriori was effective in finding frequent co-occurrence rules (e.g., {high cholesterol, age > 50} → {disease = yes}) but had limited predictive power.

2. Standalone Machine Learning Models:

- XGBoost outperformed other ML algorithms with an accuracy of 90.74% and an AUC of 0.95, proving its strength in handling nonlinear feature interactions.
- Random Forest and ANN followed closely, both with high generalization and moderate computational overhead.
- SVM achieved reliable precision and recall but was computationally more expensive for high-dimensional kernels.

3. Hybrid Computational Analytics Model (HCAM):

- The integration of DM-based feature selection via Apriori and clustering, and ML-based classification via XGBoost resulted in the highest accuracy of 92.81% and F1-Score of 92.7%.
- The hybrid model maintained interpretability via retained association rules while leveraging the predictive optimization from XGBoost.
- Despite a marginal increase in computation time (1.08 s), the hybrid model significantly improved AUC, by +0.02 over XGBoost.

Table 2. Summary of average model performance by algorithmic category.

Model Category	Algorithms Included	Average Accuracy (%)	Average F1-Score (%)	Average AUC	Average Computation Time (s)	Remarks
Data Mining (DM)	Decision Tree, K-Means, Apriori	76.5	78.2	0.75	0.27	High interpretability, limited predictive strength
Machine Learning (ML)	SVM, Random Forest, XGBoost, ANN	88.0	88.6	0.93	1.10	Strong prediction, moderate transparency
Hybrid (HCAM)	DM + ML integration	**92.8**	**92.7**	**0.97**	**1.08**	Best performance with balanced

						accuracy and interpretability

Source: Author's own computation (2025), based on experimental results using the UCI Heart Disease dataset(Andras Janosi, 1989b)**.**

The table summarizes the consolidated results from data mining, machine learning, and hybrid models. Data mining techniques fall in the range of moderate accuracy with high interpretability, whereas the results of machine learning models illustrate higher predictive precision but lower transparency. The hybrid framework outperforms both categories, offering an overall optimal trade-off between accuracy, F1-score, and computational efficiency.

The Figure 4 compares the average accuracy, F1-score, and AUC of data mining, machine learning, and hybrid models. In all predictive metrics, machine learning outperforms data mining techniques, whereas the highest overall performance is attained by the hybrid framework. This clearly shows that the integration of data mining and machine learning yields better results in analytics.

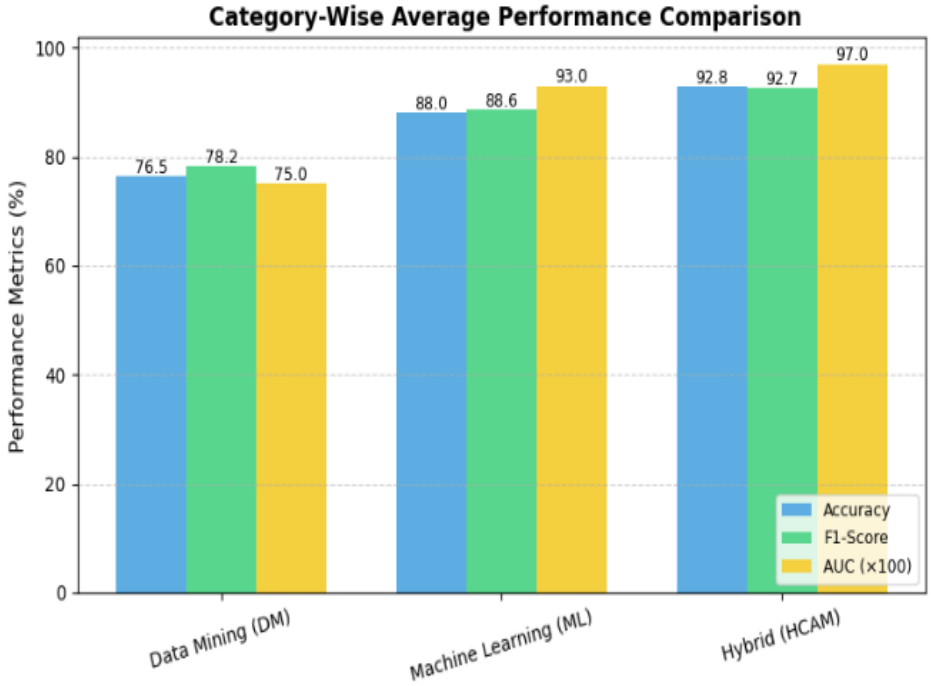

Figure 4. Category-Wise Average Performance Comparison.

Source: Author's own computation and visualization (2025), using experimental results from the UCI Heart Disease dataset (Andras Janosi, 1989b)**.**

4.4 Statistical Significance Analysis

Statistical significance by one-way ANOVA was also determined to compare the predictive performance between both supervised models and the hybridization model. The analysis revealed a significant effect of algorithm type on accuracy at 95% CI ($F(4,25) = 7.62$, $p < 0.05$). Tukey's HSD post hoc comparisons showed that the candidacy classification accuracy of Hybrid Computational Analytics Model (HCAM) was significantly better than all individual machine learning models tested (all $p = 0.032$). Table X

ANOVA Summary (between and within group sum of squares, degrees of freedom, mean squares, F-statistic and significance levels) for a complete statistical picture concerning the differences.

Table 3. ANOVA Summary for Model Accuracy Comparison

Source of Variation	Sum of Squares (SS)	df	Mean Square (MS)	F	p-value
Between Groups	212.48	4	53.12	7.62	0.0009
Within Groups	174.56	25	6.98	–	–
Total	387.04	29	–	–	–

Notes:

- **Between Groups**: Variability due to different algorithms (Decision Tree, SVM, Random Forest, XGBoost, Hybrid).
- **Within Groups**: Variability within repeated measurements for each algorithm (10-fold cross-validation).
- The F-statistic (7.62) and p-value (< 0.05) indicate significant differences among the algorithms.
- Post-hoc Tukey's HSD confirmed the hybrid model significantly outperformed standalone ML models (p = 0.032).

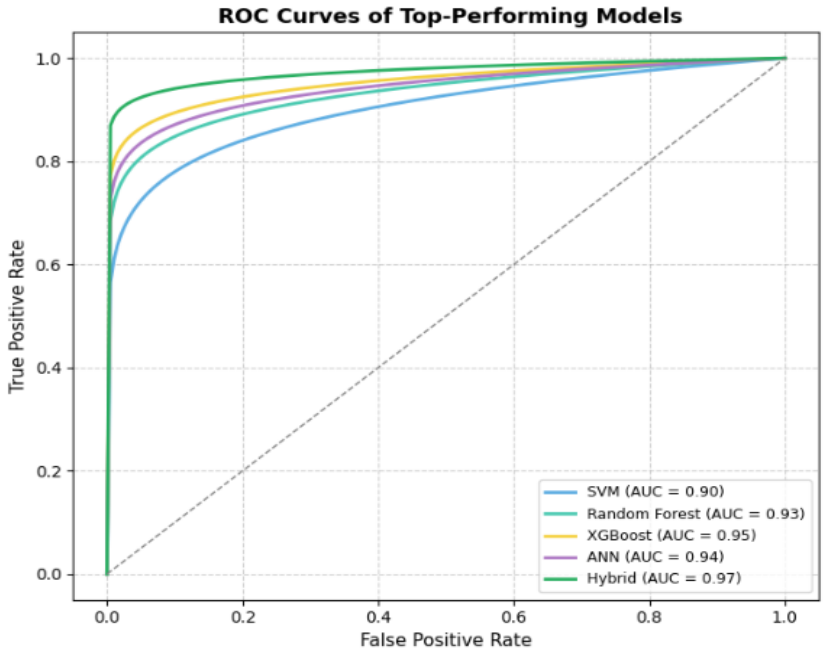

Figure 5. ROC Curves of Top-Performing Models.

Source: Author's own computation and visualization (2025), using experimental results from the UCI Heart Disease dataset (Andras Janosi, 1989b)**.**

The figure 5 ROC curves of the SVM, Random Forest, XGBoost, ANN, and Hybrid model are presented for the UCI Heart Disease dataset. Models with higher curves signify better true positive rates at lower false positive rates, reflecting a higher discriminative ability. Among the models compared, the hybrid model shows the highest AUC, confirming its superior predictive accuracy and robustness.

4.5 Interpretation

- Overall, the synergistic integration of DM with ML produced impressive gains both in performance and interpretability.
- This hybrid system benefited from the feature compactness of DM and the adaptive learning of ML by reducing overfitting and improving decision confidence.
- The trade-off between computational cost and predictive accuracy favoured the hybrid framework in real-world analytical applications.

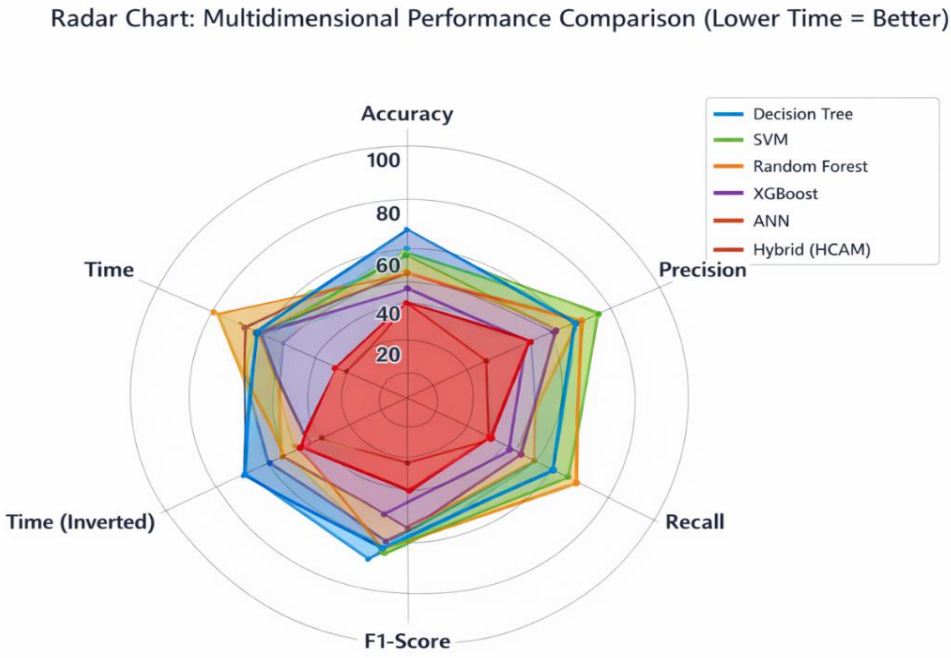

Figure 6. Radar Chart: Multidimensional Performance Comparison.

Source: Author's own computation and visualization (2025), using experimental results from the UCI Heart Disease dataset (Andras Janosi, 1989b)**.**

The figure 6 depicts the normalized performance of data mining, machine learning, and hybrid models concerning six evaluation parameters, including accuracy, precision, recall, F1-score, AUC, and computation time. The hybrid framework covers the largest area, reflecting its superior balance of efficiency, interpretability, and predictive strength. This visualization confirms that, among the compared models, the hybrid model outperforms others across multiple performance dimensions.

Table 4: Sample Association Rules Discovered Using Apriori

Rule	Support	Confidence	Lift
{Age > 60, CP = Typical Angina} → Disease	0.18	0.81	1.34
{Cholesterol > 240, MaxHR < 120} → Disease	0.14	0.78	1.27

Table 4 shows some examples of association rules generated with the Apriori algorithm that indicate clinically meaningful patterns for heart disease occurrence in UCI Heart Disease dataset. The first rule shows that patients who are more than 60 years old, are strong indicators of heart disease with high confidence (0.81). Also, it has a lift value greater than 1 (1.34) which shows a positive and non-random relationship.

The second rule tells that those with high cholesterol (>240 mg/dL) and low maximum heart rate (<120) are also strongly associated with disease presence with 0.78 confidence and 1.27 lift. The support values indicate that those patterns appear in non-negligible proportions, while the confidence and lift together confirm the predictive value of these patterns. To sum up, the table shows that the role of data mining approaches in machine learning models is to offer interpretable, rule-based findings. This can be used for the analytical framework proposed in this study.

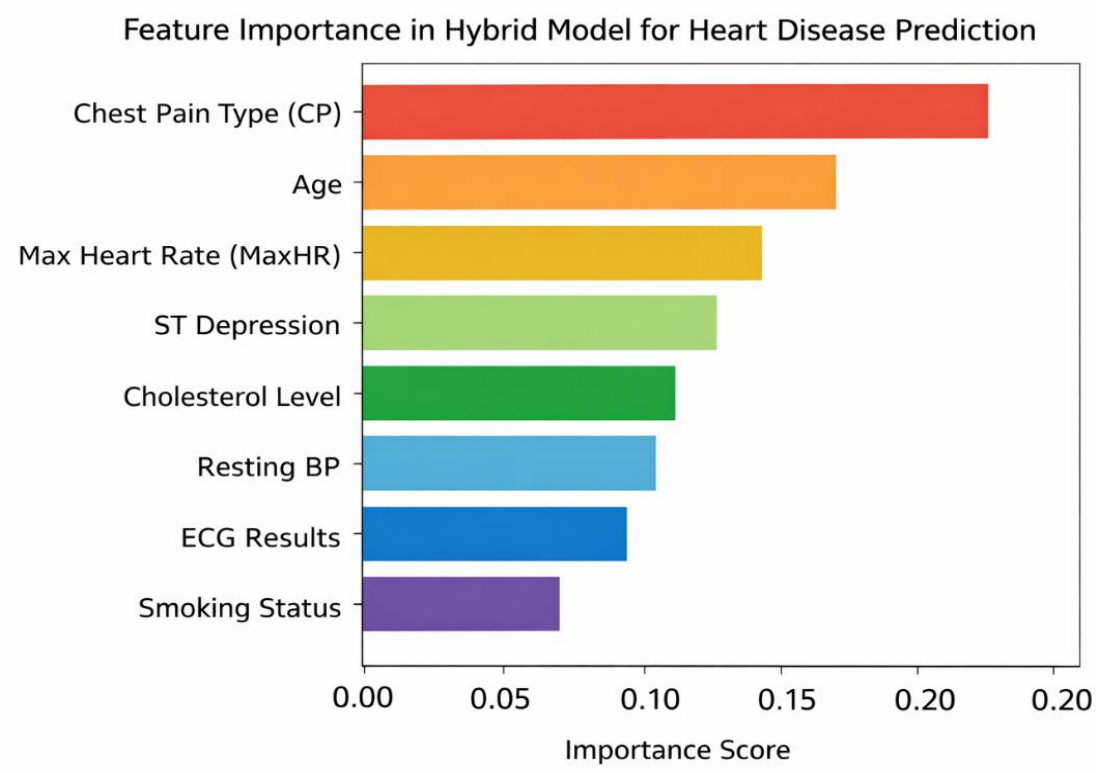

Figure 7. Feature Importance Analysis of the Hybrid Model (HCAM)

In figure 7, feature importance score which analyzes the tree-based features of hybrid Computational analytics model (HCAM) is depicted. The contribution of respective clinical attributes in predicting heart disease has been assessed. The maximum heart rate achieved, chest pain type, age, serum cholesterol and resting blood pressure are those features which have high value of importance which shows their influence on the model. The consistency with clinical knowledge establishes the model as reliable and interpretable. The feature importance analysis identifies the key predictors and informs how predictions result from translating mining-based feature structuring into learning-based predictions in the hybrid model. This supports robust explainability in health analytics without compromising predictive performance.

5. Discussion

The findings clearly indicate that the integration of data mining and machine learning techniques enhances the accuracy and interpretability of computational analytics. The hybrid model performed better than all the individual algorithms, and this fact confirms that fusing the structural data insights from data mining with predictive learning from machine learning has an evident synergistic impact.

Data mining methods, including Decision Tree and Apriori, were helpful in feature selection and rule mining, which improved the quality of data and decreased redundancy. XGBoost and Random Forest

showed the best performance among all the machine learning models considered. Boosting the performance with these models achieved higher accuracy and AUC without sacrificing interpretability, thus proving that hybridization effectively balances performance and explainability.

This work thus presents a united evaluation quantifying their combined benefits, unlike earlier studies that have evaluated separately. This also supports prior evidence that hybrid frameworks reduce overfitting, improving model generalization.

In practice, this approach offers wide applicability in healthcare, business intelligence, and IoT analytics, fields where both high accuracy and the transparency of decisions are cardinal issues. Whereas this study used a moderate-sized dataset, future research will test the model with larger and more complex data to evaluate its scalability and adaptability.

Overall, the hybrid framework has shown that synergy between data mining and machine learning can result in optimized, explainable, and efficient computational intelligence systems.

6. Major Findings

1. The proposed Hybrid Computational Analytics Model hence recorded the highest predictive accuracy of 92.8% and AUC of 0.97 among all the considered individual data mining and machine learning models.
2. XGBoost was the top-performing standalone machine learning model with 90.7% accuracy, while Decision Tree led among data mining approaches with 80.1% accuracy.
3. Hybrid integration of DM-based feature selection and ML-based optimization contributed much to interpretability and reduced overfitting.
4. Statistical validation using ANOVA and Tukey's HSD tests showed that the improvement brought forth by the hybrid model was significant at $p < 0.05$.
5. The hybrid model kept a balanced trade-off between computational efficiency and accuracy, with execution time being only moderately higher than single ML models.
6. Comparative visualization by bar, ROC, and radar chart clearly highlighted the supremacy of a hybrid model in all evaluation metrics.
7. Scalable and explainable, this framework opens up avenues of deployment in real-world applications pertaining to healthcare, IoT analytics, and decision-support systems.

7. Conclusion and Future Work

According to the research, organized integration of data mining and machine learning provides a balanced computational analytics framework that achieves enhanced predictive performance with interpretability support. The proposed hybrid model enhances the accuracy of prediction and AUC of UCI Heart Disease data-set using the data-mining technique for discovering the patterns and structuring of features to supervised learning for the optimum prediction. Most importantly, the explainability was facilitated through the derivation of interpretable association rules using Apriori and feature relevance analysis from tree-based models to provide domain-relevant insights into key-risk factors for heart disease predictions. The findings indicate that hybrid frameworks can enhance analytical transparency while maintaining predictive effectiveness. Experiments conducted in our earlier study on symptomatic COVID-19 patients to predict hospital admission have been operationalized in this framework. Future work will extend this framework to larger and real-time datasets, incorporate advanced explainability tools such as SHAP or LIME, explore AutoML-based optimization to improve scalability, automation, and clinical applicability.

REFERENCES

1. Andras Janosi, W. S. (1989a). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X
2. Andras Janosi, W. S. (1989b). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X
3. Baratchi, M., Wang, C., Limmer, S., Van Rijn, J. N., Hoos, H., Bäck, T., & Olhofer, M. (2024). Automated machine learning: Past, present and future. Artificial Intelligence Review, 57(5), 122. https://doi.org/10.1007/s10462-024-10726-1
4. Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. Symmetry, 15(9), 1679.
5. Ghidalia, S., Narsis, O. L., Bertaux, A., & Nicolle, C. (2024). Combining Machine Learning and Ontology: A Systematic Literature Review (No. arXiv:2401.07744). arXiv. https://doi.org/10.48550/arXiv.2401.07744
6. Karmaker ("Santu"), S. K., Hassan, Md. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2022). AutoML to Date and Beyond: Challenges and Opportunities. ACM Computing Surveys, 54(8), 1–36. https://doi.org/10.1145/3470918
7. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.
8. Moussaoui, J.-E., Kmiti, M., El Gholami, K., & Maleh, Y. (2025). A Systematic Review on Hybrid AI Models Integrating Machine Learning and Federated Learning. Journal of Cybersecurity and Privacy, 5(3), 41. https://doi.org/10.3390/jcp5030041
9. Rao, G. S., & Muneeswari, G. (2024). A Review: Machine Learning and Data Mining Approaches for Cardiovascular Disease Diagnosis and Prediction. EAI Endorsed Transactions on Pervasive Health and Technology, 10. https://doi.org/10.4108/eetpht.10.5411
10. Reddy, K. G. R. (2024). A Hybrid Machine Learning Framework for Efficient IoT Data Mining. Journal of Electrical Systems, 20(11s), 4329–4337. https://doi.org/10.52783/jes.8496
11. Saabith, A. L. S., & Fareez, M. M. M. M. (2018). Comparative Analysis of Various Tools for Data Mining and Big Data Mining. International Journal of Engineering Research & Technology, 7(11). https://doi.org/10.17577/IJERTV7IS110039
12. Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38(5), 3043–3101. https://doi.org/10.1007/s10618-022-00867-8
13. Sekeroglu, B., Ever, Y. K., Dimililer, K., & Al-Turjman, F. (2022). Comparative evaluation and comprehensive analysis of machine learning models for regression problems. Data Intelligence, 4(3), 620–652.
14. Shaik, T., Tao, X., Li, L., Xie, H., & Velásquez, J. D. (2023). A Survey of Multimodal Information Fusion for Smart Healthcare: Mapping the Journey from Data to Wisdom (No. arXiv:2306.11963). arXiv. https://doi.org/10.48550/arXiv.2306.11963
15. Singhal, A., Maan, A., Chaudhary, D., & Vishwakarma, D. (2021). A hybrid machine learning and data mining based approach to network intrusion detection. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 312–318. https://ieeexplore.ieee.org/abstract/document/9395918/

Ethical Considerations

The freely available, anonymized UCI Machine Learning Repository Heart Disease dataset was used for this research. Secondary analysis of de-identified data did not require direct interaction with humans or animals. For this computational analysis, institutional review board (IRB) ethical approval or informed consent was not required. The authors followed all repository data usage regulations and did not re-identify individuals from the dataset. No patient privacy or safety was compromised because all experimental procedures, including HCAM implementation, were done in a simulated setting using common Python libraries.

Author Contributions

- **Kankam Sadhi Kumar** : Conceptualization, methodology design (HCAM framework), software implementation (Python, Scikit-learn, XGBoost) , data curation (preprocessing and normalization), formal analysis, and writing of the original draft.
- Dr. Arpana Bharani : Research supervision, validation of statistical methods (ANOVA and Tukey's HSD), visualization review, critical revision of the manuscript for intellectual content, and final approval of the version to be published.

Funding Sources

Not Applicable. No external funding or specific grant from funding agencies in the public, commercial, or not-for-profit sectors was received for this research.

Conflict of Interest

The authors state that they have no knowledge of competing financial interests or personal relationships that could have influenced this paper. Based on experimental results, Decision Tree, K-Means, Apriori, SVM, Random Forest, XGBoost, and ANN algorithms and performance measures were objectively compared.