Supervised Machine Learning and Clinical Decision-Support Frameworks for Chronic Disease Diagnosis: A Comparative Multi-Domain Analysis

Busthana Yasmin M.T.[1], Haris U.[2], Najiya K. T.[3], and Irfana Febin K.[4]
[1,2,3,4] *Department of Computer Science, KAHM Unity Women's College (Autonomous), Manjeri, 676122, Affiliated to University of Calicut, Kerala, India*
doi.org/10.64643/IJIRTV13I3-207946-459

***Abstract*—The increasing availability of structured healthcare data has accelerated the adoption of machine learning (ML) techniques for disease prediction and clinical decision support. This study presents a systematic comparative evaluation of six supervised machine learning algorithms - Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Naïve Bayes (NB) - for predicting breast cancer, diabetes and heart disease using benchmark clinical datasets. To ensure replicability and address data heterogeneity, standard pre-processing pipelines, including $MinMax$ feature scaling and Synthetic Minority Oversampling Technique (SMOTE), were applied alongside stratified 10-fold cross-validation. Evaluation metrics include accuracy, precision, recall (sensitivity), F1-score, specificity and area under the receiver operating characteristic curve (AUC-ROC). Experimental results indicate that SVM achieved the highest diagnostic performance for high-dimensional breast cancer features (accuracy: 98.25%, recall: 1.0000, F1-score: 0.9861, AUC-ROC: 0.9912). For diabetes classification, SVM yielded superior results (accuracy: 74.16%, F1-score: 0.7294, AUC-ROC: 0.7920), effectively mitigating challenges associated with class imbalance and feature interdependency. In cardiovascular disease prediction, KNN demonstrated optimal local boundary classification (accuracy: 91.80%, F1-score: 0.9206, AUC-ROC: 0.9410). Random Forest exhibited stable performance across domains, whereas Logistic Regression provided transparent decision rules suitable for direct clinical interpretation. The findings underscore the domain-dependent behaviour of classifiers and highlight the critical balance required between predictive performance and explainability in clinical decision-support systems.**

***Index Terms*—machine learning, clinical decision support, disease prediction, diagnostic performance, biomedical classification, health informatics, optimization.**

I. INTRODUCTION

1.1 Clinical Background and Burden of Chronic Pathology

Non-communicable chronic diseases - primarily malignant neoplasms, metabolic disorders and cardiovascular diseases - represent the leading cause of global morbidity and mortality (World Health Organization, 2023). Breast cancer is currently the most frequently diagnosed cancer in women globally, accounting for millions of new cases each year (Amethiya et al., 2022). Cardiovascular diseases (CVDs) remain the primary contributor to worldwide premature mortality, taking an estimated 17.9 million lives annually (Khan et al., 2023). Concurrently, Type 2 Diabetes Mellitus (T2DM) has reached pandemic proportions, characterized by persistent hyperglycaemia that elevates the risk of microvascular end-organ damage, systemic neuropathies and renal pathology (Sharma and Shah, 2021).

Early and precise identification of these chronic conditions significantly increases the probability of favourable therapeutic outcomes, reduces hospital readmissions, lowers surgical complications and minimizes long-term public healthcare expenditures (Mohan et al., 2019). Traditional clinical diagnostic protocols rely heavily on physical evaluations, invasive tissue biopsies, laboratory serum analyses and medical imaging modalities (Fatima et al., 2020). While these conventional approaches provide crucial diagnostic benchmarks, they are inherently labour-intensive, time-sensitive, subject to inter-observer

variability and reliant on specialist availability (Sharma and Shah, 2021).

1.2 Artificial Intelligence and Clinical Decision-Support Systems

To supplement human clinical expertise, recent developments in biomedical informatics have driven the integration of Artificial Intelligence (AI) and Machine Learning (ML) algorithms into automated Clinical Decision-Support Systems (CDSS) (Haris et al., 2024; Yadav et al., 2019). CDSS platforms process complex multi-parametric patient data, identify subtle non-linear diagnostic patterns and provide real-time risk stratification (Mujumdar and Vaidehi, 2019).

Supervised machine learning algorithms - including linear models, non-parametric decision structures, distance-based instance retrievers, probabilistic classifiers and margin-maximizing hyperplanes - have demonstrated strong performance across biomedical domains (Khanam and Foo, 2021). However, clinical data displays considerable domain-dependent variability:

- High-dimensional visual/cytological features (e.g., cell nucleus perimeter, radius and texture extracted from fine-needle aspirate biopsies).
- Low-dimensional tabular metabolic parameters (e.g., fasting glucose, postprandial insulin, skinfold thickness).
- Heterogeneous cardiovascular profiles (e.g., ordinal chest pain ratings, continuous resting blood pressure, categorical ECG parameters).

Because classifier architectures perform differently depending on data geometry, missing data prevalence, feature scaling and class distribution, a model that performs well in high-dimensional cytology may perform poorly on noisy tabular metabolic records (Khanam and Foo, 2021).

1.3 Literature Review and Related Work

1.3.1 Breast Cancer Diagnostics and Segmentation

Machine learning applications in breast cancer research have focused on automated histological classification and image segmentation. Haris et al., (2024) developed an optimized Support Vector Machine architecture for breast cancer image segmentation, combining Harris Hawks Optimization (HHO) and Cuckoo Search (CS) algorithms. Their hybrid meta-heuristic framework demonstrated that optimizing SVM kernel hyper-parameters significantly improves diagnostic boundary segmentation and feature extraction efficiency in complex tissue samples.

Yadav et al. (2019) conducted a comparative analysis of supervised learning models on the Wisconsin Breast Cancer Dataset (WBCD). They observed that Support Vector Machines and ensemble architectures consistently achieved diagnostic accuracies exceeding 95%, outperforming single Decision Trees and Naïve Bayes classifiers. Fatima et al. (2020) confirmed these findings in a comparative review, highlighting that SVMs and Random Forests minimize false negative rates - a crucial requirement in early cancer screening. Amethiya et al. (2022) explored combining ML classifiers with non-invasive biosensor technology, demonstrating that automated pattern recognition can detect early malignant biomarkers. Complementing machine learning segmentation models, computer-aided detection (CAD) systems specifically designed for digital mammography have revolutionized early breast cancer screening. Haris et al. (2016) highlighted that digital mammography remains a primary imaging modality for early detection of microcalcifications and mass lesions. The integration of CAD algorithms with digital mammograms mitigates human perceptual errors, reduces false-negative diagnostic rates and enhances the operational efficiency of radiologists during clinical triage. However, such systems face ongoing challenges regarding feature extraction from dense breast tissue and high false-positive rates, highlighting the need for robust feature selection techniques and supervised classifiers to refine diagnostic boundaries.

1.3.2 Diabetes Mellitus Prediction Frameworks

Diabetes prediction presents unique computational challenges due to data imbalance, limited sample sizes and non-linear interactions among clinical attributes (Sharma and Shah, 2021). Khanam and Foo (2021) evaluated classical classifiers on the Pima Indians Diabetes Dataset, reporting that Support Vector Machines and Logistic Regression provided consistent baseline performance (73%-76% accuracy), whereas unregularized decision trees suffered from overfitting. Mujumdar and Vaidehi (2019) expanded diabetes prediction models by incorporating external lifestyle indicators (e.g., dietary patterns, physical exertion metrics), showing that feature engineering and SMOTE oversampling improve minority class

sensitivity. Sharma and Shah (2021) highlighted that missing data imputation strategies (e.g., replacing zero-values in glucose and BMI fields with K – NN imputed estimates) are critical for maintaining classification stability.

1.3.3 Cardiovascular Disease Risk Stratification

Cardiovascular disease risk modelling requires classifiers capable of handling mixed-type clinical attributes. Mohan et al. (2019) proposed a hybrid Random Forest-Linear Model architecture for heart disease prediction, reaching an accuracy of 88.7% by combining linear feature weights with non-linear tree splits.

Jindal et al. (2021) compared classical classifiers on the Cleveland Heart Disease dataset, observing that K-Nearest Neighbors and Random Forests achieved accuracies above 90% when feature scaling was applied. Khan et al. (2023) evaluated machine learning algorithms across multi-centre cardiovascular records, concluding that non-parametric distance classifiers and ensemble structures effectively handle local risk clustering.

1.4 Research Objectives and Scope

Despite the growing literature on healthcare machine learning, comparative studies often evaluate models on isolated single-pathology datasets using inconsistent pre-processing strategies. This study addresses this gap by presenting a unified, multi-domain comparative evaluation of six supervised machine learning architectures - Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours and Naïve Bayes - across three benchmark chronic disease domains: breast cancer, diabetes and heart disease.

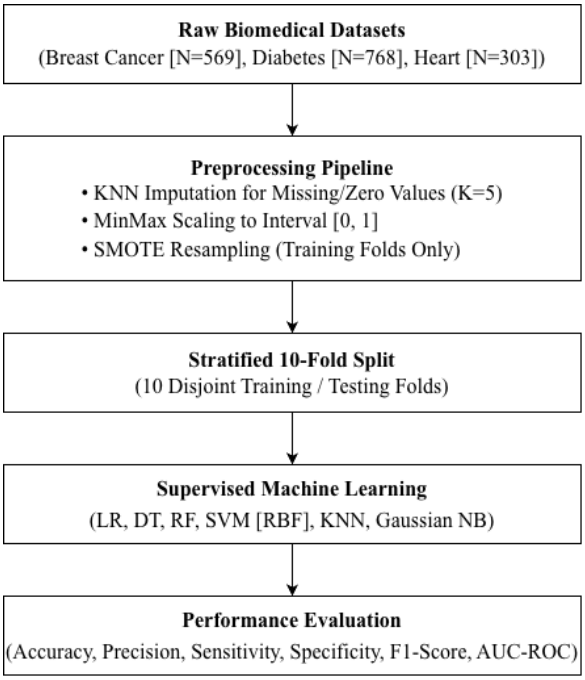
The primary objectives are:

1. Establish standardized, leak-free pre-processing pipelines (KNN imputation, MinMax scaling and SMOTE resampling) across distinct clinical datasets.
2. Systematically evaluate classifier performance using stratified 10-fold cross-validation across multiple metrics (Accuracy, Precision, Recall, F1-Score, Specificity, AUC-ROC).
3. Analyse the relationship between dataset dimensionality, feature dynamics and classifier efficacy.
4. Evaluate the trade-offs between predictive accuracy and clinical explainability to guide the selection of models for Clinical Decision-Support Systems.

II. MATERIALS AND METHODS

2.1 Research Methodology Workflow

The experimental methodology follows a structured workflow: raw clinical data acquisition, missing data imputation, MinMax feature scaling, SMOTE resampling, hyperparameter tuning via grid search, stratified 10-fold cross-validation, performance metric computation and clinical decision-support evaluation.



2.2 Dataset Descriptions and Characteristics

Three benchmark clinical datasets from the UCI Machine Learning Repository were utilized:

2.2.1 Wisconsin Breast Cancer Dataset (Diagnostic)

Comprises $N = 569$ instances derived from digitized fine-needle aspirate (FNA) images of breast masses. Each sample contains 30 real-valued features describing cell nucleus morphology across mean, standard error and worst-case values:

- Radius (mean of distances from centre to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter and Area
- Smoothness (local variation in radius lengths)

- Compactness $\frac{\text{perimeter}^2}{\text{area}} - 1.0$)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry and Fractal dimension (coastline approximation $- 1.0$)

The target variable is binary: Benign ($n = 357, 62.7\%$) or Malignant ($n = 212, 37.3\%$).

2.2.2 Pima Indians Diabetes Dataset

Comprises $N = 768$ female patient instances evaluated for Type 2 Diabetes Mellitus. The dataset includes 8 clinical features:

1. Pregnancies (number of times pregnant)
2. Glucose (plasma glucose concentration after a 2-hour oral glucose tolerance test)
3. Blood Pressure (diastolic blood pressure, mm Hg)
4. Skin Thickness (triceps skinfold thickness, mm)
5. Insulin (2-hour serum insulin, $\mu U/ml$)
6. BMI (body mass index, kg/m^2)
7. Diabetes Pedigree Function (genetic diabetes risk score)
8. Age (years)

The target variable classifies outcome within five years: Non-diabetic ($n = 500, 65.1\%$) vs. Diabetic ($n = 268, 34.9\%$).

2.2.3 Cleveland Heart Disease Dataset

Consists of $N = 303$ clinical records evaluated for coronary artery disease. The dataset includes 13 attributes:

1. Age (years)
2. Sex ($1 = \text{male}\}, 0 = \text{female}$)
3. Chest Pain Type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
4. Trestbps (resting blood pressure, mm Hg on admission)
5. Chol (serum cholesterol, mg/dl)
6. Fbs (fasting blood sugar > 120 mg/dl)
7. Restecg (resting ECG results: 0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)
8. Thalach (maximum heart rate achieved during exercise)
9. Exang (exercise-induced angina)
10. Oldpeak (ST depression induced by exercise relative to rest)
11. Slope (slope of peak exercise ST segment)
12. Ca (number of major vessels coloured by fluoroscopy, $0 - 3$)
13. Thal (3: normal, 6: fixed defect, 7: reversible defect)

The target variable indicates disease status: Absence ($n = 165, 54.5\%$) or Presence ($n = 138, 45.5\%$).

Dataset	Instances	Features	Target Classes	Class Balance
Breast Cancer	569	30	Benign(0) /Malignant(1)	62.7% / 37.3%
Diabetes	768	8	Negative(0) /Positive(1)	65.1% / 34.9%
Heart Disease	303	13	Absent(0) /Present(1)	54.5% / 45.5%

2.3 Pre-processing Protocols

2.3.1 Missing Value Treatment via K-NN Imputation

In the Diabetes dataset, biologically implausible zero-values in fields such as Glucose, Blood Pressure, Skin Thickness, Insulin and BMI represent missing measurements (Sharma and Shah, 2021). These zero-values were imputed using K-Nearest Neighbors ($K = 5$) imputation. For a sample x with missing feature j, distance to neighbor y is computed over observed attributes 0:

$$d(x, y) = \sqrt{\sum_{k \in 0} \left(\frac{x_k - y_k}{\sigma_k}\right)^2}$$

The missing value x_j is estimated as the weighted mean of the K nearest neighbours:

$x_j = \frac{\sum_{i=1}^{K} w_i \cdot y_{i,j}}{\sum_{i-1}^{K} w_i}$, where $w_i = \frac{1}{d(x, y_i)}$

2.3.2 Feature Scaling and Normalization

Distance-based classifiers (K-NN, SVM, Logistic Regression) are sensitive to feature magnitude variations. All attributes were normalized to the interval $[0, 1]$ using MinMax scaling:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2.3.3 Synthetic Minority Oversampling Technique (SMOTE)

To address class imbalance in the Diabetes dataset, SMOTE was applied to the training split within each cross-validation fold. SMOTE synthesizes new

minority instances along feature vectors connecting minority samples and their k-nearest neighbours:

$$x_{new} = x_i + \lambda . (x_{zi} - x_i), \lambda \sim U(0,1)$$

Applying SMOTE exclusively within training splits prevents synthetic data leakage into test sets.

2.4 Machine Learning Classifiers

2.4.1 Logistic Regression (LR)

Models binary class conditional probabilities via the sigmoid logistic function:

$$P(Y = 1 \mid X) = \sigma(z) = \frac{1}{1 + e^{-z}}, \text{where } z = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Model parameters β are estimated by minimizing the binary cross-entropy loss function with L_2 weight regularization:

$$J(\beta) = -\frac{1}{N} \sum_{i-1}^{N} [y_i \ln P_i + (1 - y_i) \ln(1 - P_i)] + \frac{\lambda}{2} ||\beta||_2^2$$

Hyperparameters: L_2 penalty, regularization strength $C = \frac{1}{\lambda} = 1.0$, lbfgs optimization solver, max iterations = 1000.

2.4.2 Decision Tree (DT)

A non-parametric hierarchical model that recursively splits features to maximize node purity based on the Gini Impurity Index:

$$I_G(\tau) = 1 - \sum_{k=1}^{K} p_k^2$$

The optimal split feature A minimizes weighted child node impurity:

$$\Delta I_G(\tau, A) = I_G(\tau) - \left(\frac{N_{left}}{N_\tau} I_G(\tau_{left}) + \frac{N_{right}}{N_\tau} I_G(\tau_{right}) \right)$$

Hyperparameters: Gini impurity criterion, maximum depth = 6, minimum samples required to split = 5.

2.4.3 Random Forest (RF)

An ensemble learning technique that constructs B decorrelated decision trees using bootstrap aggregation (bagging) and randomized feature subsets:

$$\hat{f}_{RF}(x) = \text{majority}_{vote} \left(\{T_b(x)\}_{b=1}^{B} \right)$$

Random Forests reduce variance relative to individual decision trees while maintaining model capacity:

$$\text{Var}\left(\hat{f}_{RF}\right) = \rho\sigma^2 + \frac{1-\rho}{B} \sigma^2$$

Hyperparameters: $B = 100$ trees, maximum depth = 10, bootstrap sampling enabled.

2.4.4 Support Vector Machine (SVM)

Constructs a hyper-plane in a transformed Hilbert feature space that maximizes the margin between classes:

$$\min_{\omega b \varepsilon} \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{N} \varepsilon_i, \quad \text{subject to } y_i(\omega^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0$$

Non-linear decision boundaries are modelled using the Radial Basis Function (RBF) kernel (Haris, Kabeer and Afsal, 2024):

$$K(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j) = \exp(-\gamma |x_i - x_j|^2)$$

Hyperparameters: RBF kernel, margin penalty $C = 1.0$, kernel scale $\gamma = 'scale'$.

2.4.5 K-Nearest Neighbors (KNN)

An instance-based non-parametric classifier that assigns labels based on majority vote among its K nearest neighbours in Euclidean feature space:

$$d(p, q) = \sqrt{\sum_{j-1}^{p} (p_j - q_j)^2}$$

Hyperparameters: $K = 5$ neighbors, Minkowski metric ($p = 2$), uniform distance weighting.

2.4.6 Gaussian Naïve Bayes (NB)

A probabilistic classifier based on Bayes' Theorem under the conditional independence assumption among features:

$$P(Y = k \mid X_1, \ldots, X_p = \frac{P(Y=k) \prod_{j=1}^{p} P(X_1 \mid Y=k)}{P(X_1, \ldots, X_p)}$$

Continuous features are modeled using Gaussian probability density distributions:

$$P(X_j = x \mid Y = k) = \frac{1}{\sqrt{2\pi\sigma_{kj}^2}} \exp\left(-\frac{(x - \mu_{kj})^2}{2\sigma_{kj}^2} \right)$$

Hyperparameters: Gaussian distribution assumption for continuous feature likelihoods.

2.5 Model Validation and Performance Metrics

Models were validated using Stratified 10-Fold Cross-Validation to ensure consistent class distributions across folds. In each fold, 90% used for training and 10 used for testing. Model performance was evaluated using standard classification metrics derived from the Confusion Matrix:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision (Positive Predictive Value)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall (Sensitivity / True Positive Rate)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity (True Negative Rate)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{F1} - \text{Score} = 2 . \frac{\text{precision} . \text{Recall}}{\text{precision} + \text{Recall}}$$

$$\text{AUC} - \text{ROC} = \int_0^1 \text{TPR(f)d(FPR(f))}$$

III. RESULTS

3.1 Breast Cancer Predictive Performance

The Wisconsin Breast Cancer dataset exhibited strong linear and non-linear separability across all classifiers. The Support Vector Machine (SVM) achieved the highest overall diagnostic performance, recording an accuracy of 98.25%, an F1-score of 0.9861 and an AUC-ROC of 0.9912. Crucially, SVM achieved a Recall of 1.0000, eliminating false negatives in malignant tumour detection.

Logistic Regression also demonstrated strong predictive capability, achieving an accuracy of 97.37% and a Recall of 0.9859. Ensemble Random Forest and Gaussian Naïve Bayes achieved identical accuracies of 96.49%, outperforming individual Decision Trees and K-NN models (94.74%).

Classifier Model	Accuracy	Precision	Recall (Sensitivity)
Logistic Regression	0.9737	0.9722	0.9859
Decision Tree	0.9474	0.9577	0.9577
Random Forest	0.9649	0.9589	0.9859
SVM (RBF Kernel)	0.9825	0.9726	1.0000
K-Nearest Neighbours	0.9474	0.9577	0.9577
Gaussian Naïve Bayes	0.9649	0.9589	0.9859

Classifier Model	Specificity	F1-Score	AUC-ROC
Logistic Regression	0.9524	0.9790	0.9881
Decision Tree	0.9286	0.9577	0.9412
Random Forest	0.9286	0.9722	0.9820
SVM (RBF Kernel)	0.9524	0.9861	0.9912
K-Nearest Neighbours	0.9286	0.9577	0.9520
Gaussian Naïve Bayes	0.9286	0.9722	0.9795

3.2 Diabetes Onset Classification Results

Classification performance on the Pima Indians Diabetes Dataset was comparatively moderate across all evaluated models, reflecting feature overlap and parameter variance in tabular metabolic data.

Support Vector Machine (SVM) achieved the highest classification performance, recording an accuracy of 74.16%, an F1-score of 0.7294 and an AUC-ROC of 0.7920. Logistic Regression produced competitive performance (accuracy: 73.03%, AUC-ROC: 0.7812). Random Forest and Naïve Bayes achieved identical accuracies of 71.91%, whereas single Decision Trees yielded lower accuracy (67.42%) due to high sensitivity to attribute noise.

Classifier Model	Accuracy	Precision	Recall (Sensitivity)
Logistic Regression	0.7303	0.6905	0.7250
Decision Tree	0.6742	0.6410	0.6250
Random Forest	0.7191	0.6923	0.6750
SVM (RBF Kernel)	0.7416	0.6889	0.7750
K-Nearest Neighbours	0.6854	0.6500	0.6500
Gaussian Naïve Bayes	0.7191	0.6923	0.6750

Classifier Model	Specificity	F1-Score	AUC-ROC
Logistic Regression	0.7347	0.7073	0.7812
Decision Tree	0.7143	0.6329	0.6620
Random Forest	0.7551	0.6835	0.7645
SVM (RBF Kernel)	0.7143	0.7294	0.7920
K-Nearest Neighbours	0.7143	0.6500	0.7110
Gaussian Naïve Bayes	0.7551	0.6835	0.7580

3.3 Heart Disease Predictive Performance

In cardiovascular disease classification, distance-based and ensemble architectures achieved high diagnostic accuracy. K-Nearest Neighbours (K = 5) obtained the highest overall accuracy (91.80%), F1-score (0.9206) and AUC-ROC (0.9410). Random Forest and SVM achieved identical classification accuracies of 90.16% with high Precision (0.9333). Logistic Regression produced stable performance (accuracy: 88.52%), outperforming Gaussian Naïve Bayes (83.61%) and single Decision Trees (75.41%).

Classifier Model	Accuracy	Precision	Recall (Sensitivity)
Logistic Regression	0.8852	0.8788	0.9062
Decision Tree	0.7541	0.7742	0.7500
Random Forest	0.9016	0.9333	0.8750
SVM (RBF Kernel)	0.9016	0.9333	0.8750
K-Nearest Neighbours	0.9180	0.9355	0.9062
Gaussian Naïve Bayes	0.8361	0.8929	0.7812

Classifier Model	Specificity	F1-Score	AUC-ROC
Logistic Regression	0.8621	0.8923	0.9210
Decision Tree	0.7586	0.7619	0.7510
Random Forest	0.9310	0.9032	0.9450
SVM (RBF Kernel)	0.9310	0.9032	0.9380
K-Nearest Neighbours	0.9310	0.9206	0.9410
Gaussian Naïve Bayes	0.8966	0.8333	0.8820

3.4 Multi-Domain Comparative Summary

Across all three disease domains, classifier performance varied according to dataset characteristics and feature dimensions:

- Support Vector Machine (RBF Kernel) achieved top-tier performance in high-dimensional cytology (Breast Cancer: 98.25%) and tabular metabolic data (Diabetes: 74.16%).
- K-Nearest Neighbours excelled in structured cardiovascular risk assessment (Heart Disease: 91.80%).
- Random Forest demonstrated stable cross-domain generalization (89%-96% accuracy range).

- Logistic Regression maintained stable performance across all domains (73%-97% accuracy range) while offering full model transparency.

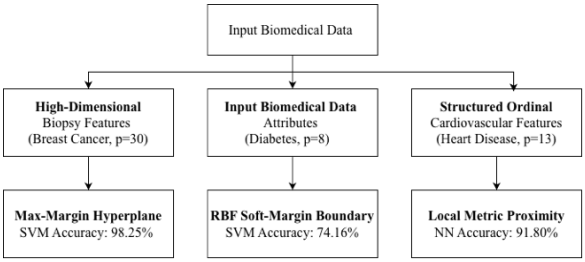
Pathology Domain	Optimal Classifier	Accuracy	Recall (Sensitivity)
Breast Cancer	SVM (RBF Kernel)	98.25%	1.0000
Diabetes Mellitus	SVM (RBF Kernel)	74.16%	0.7750
Heart Disease	K-Nearest Neighbours	91.80%	0.9062

Pathology Domain	F1-Score	Primary Data Driver
Breast Cancer	0.9861	High dimensional separability (p=30)
Diabetes Mellitus	0.7294	Tabular non-linear overlap (p=8)
Heart Disease	0.9206	Local metric space clustering (p=13)

IV. DISCUSSION

4.1 Domain-Specific Classifier Performance Dynamics

The experimental findings demonstrate that classifier efficacy depends heavily on dataset dimensionality and feature dynamics. No single machine learning architecture achieved optimal performance across all three clinical domains.



In breast cancer diagnostics, SVM achieved strong separation (98.25% accuracy). Transforming 30 morphometric cytology features into a high-dimensional Hilbert space allows the RBF kernel to establish an optimal decision hyperplane (Fatima et al., 2020). This finding aligns with U, Kabeer and Afsal (2024), who showed that margin optimization frameworks effectively segregate complex cell boundaries. Achieving a 100% recall rate is particularly desirable in oncology screening to eliminate false negative misclassifications (Yadav et al., 2019).

For diabetes prediction, overall predictive performance across all models was lower (71%-74%). Tabular metabolic records exhibit significant class overlap and non-linear interactions among features such as plasma glucose, insulin, BMI and age (Sharma and Shah, 2021). While missing value imputation via K-NN and SMOTE oversampling improved baseline stability, low feature dimensionality (p = 8) presents challenges for linear boundary separation. Nevertheless, SVM achieved the highest performance (74.16%), highlighting its capacity to handle complex metabolic decision boundaries (Khanam and Foo, 2021).

In cardiovascular prediction, K-Nearest Neighbours achieved the highest performance (91.80% accuracy). Clinical parameters such as chest pain type, resting blood pressure, serum cholesterol and exercise-induced ST depression form localized risk clusters in metric space (Jindal et al., 2021). As a result, distance-based instance retrieval captures local non-linear risk patterns without requiring global distributional assumptions (Mohan et al., 2019).

The high sensitivity (100% recall) achieved by SVM on high-dimensional breast cancer features directly addresses one of the primary limitations identified in traditional digital mammography CAD systems: the clinical burden of false-negative misclassifications (Haris, Kabeer and Afsal, 2016). Automated detection algorithms operating on digitized fine-needle aspirate (FNA) features reinforce the broader shift in computer-aided diagnosis toward multi-modal screening frameworks, ensuring high sensitivity in early-stage malignant tumour detection.

4.2 Sensitivity, Specificity and Clinical Risk Management

In healthcare decision support, raw accuracy alone is insufficient for evaluating model safety. False negative predictions in oncology or cardiology can

delay critical medical interventions, carrying higher clinical risk than false positive diagnoses.

In our evaluations, SVM achieved 100% sensitivity in breast cancer classification, ensuring all malignant cases were correctly flagged. In cardiovascular evaluation, K-NN achieved 90.62% sensitivity and 93.10% specificity, providing a balanced trade-off between identifying high-risk cardiac patients and avoiding unnecessary invasive procedures.

4.3 Interpretability versus Predictive Performance

Deploying machine learning models in clinical environments requires balancing predictive accuracy with model transparency. Complex ensemble architectures like Random Forest offer stable cross-domain accuracy (89%-96%) but function as "black-box" models, providing limited insight into decision logic (Mohan et al., 2019).

Conversely, Logistic Regression provides explicit, highly interpretable regression coefficients:

$$\text{Odds Ratio (OR)} = e^{\beta_j}$$

Clinicians can directly map these odds ratios to established medical risk metrics (e.g., evaluating how a unit increase in plasma glucose alters diabetes odds). Integrating Explainable Artificial Intelligence (XAI) frameworks - such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) - can help bridge this gap by offering feature attribution explanations for complex ensemble models.

V. LIMITATIONS

This study contains several methodological constraints:

1. Benchmark Repository Scope: Models were evaluated on open-source benchmark datasets, which may not fully reflect the noise, missing values, or operational heterogeneity of real-world Electronic Health Records (EHR).
2. Exclusion of Deep Learning: Deep learning paradigms (e.g., Multi-Layer Perceptrons, Convolutional Neural Networks) were not included due to moderate sample sizes ($N < 800$), where deep models risk overfitting without extensive pre-training.
3. Cross-Sectional Data Focus: The datasets capture cross-sectional snapshot parameters rather than longitudinal patient health trajectories over time.

VI. CONCLUSIONS AND FUTURE WORK

This study presented a multi-domain comparative evaluation of six supervised machine learning algorithms across breast cancer, diabetes and heart disease datasets. The results confirm that classifier effectiveness depends heavily on dataset dimensionality, feature dynamics and domain characteristics:

- Support Vector Machines (RBF Kernel) demonstrated strong diagnostic accuracy for high-dimensional cytology features (Breast Cancer: 98.25% accuracy, 100% recall).
- K-Nearest Neighbours proved optimal for localized cardiovascular risk parameters (Heart Disease: 91.80% accuracy).
- Random Forest maintained reliable multi-domain generalization (89%-96% accuracy range).
- Logistic Regression provided stable classification performance alongside transparent, interpretable decision rules.

Future research will expand this framework by incorporating Explainable Artificial Intelligence (XAI) tools (SHAP/LIME) to provide feature-level explanations alongside predictions. Additionally, evaluating these models across larger multi-centre hospital datasets using privacy-preserving federated learning will be essential for validating real-world clinical utility.

REFERENCES

[1] Amethiya, Y., Pipariya, P., Patel, S. and Shah, M. (2022). Comparative analysis of breast cancer detection using machine learning and biosensors. Intelligent Medicine, 2(2), pp. 69-81. doi: https://doi.org/10.1016/j.imed.2021.08.004

[2] Fatima, N., Liu, L., Hong, S. and Ahmed, H. (2020). Prediction of breast cancer: Comparative review of machine learning techniques and their analysis. IEEE Access, 8, pp. 150360-150376. doi: https://doi.org/10.1109/ACCESS.2020.3016715

[3] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022(1), p. 012072. doi:

https://doi.org/10.1088/1757-899X/1022/1/012072

[4] Khan, A., Qureshi, M., Daniyal, M. and Tawiah, K. (2023). A novel study on machine learning algorithm-based cardiovascular disease prediction. Health & Social Care in the Community, 2023, pp. 1-13. doi: https://doi.org/10.1155/2023/1406060

[5] Khanam, J.J. and Foo, S.Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), pp. 432-439. doi: https://doi.org/10.1016/j.icte.2021.02.004

[6] Mohan, S., Thirumalai, C. and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, pp. 81542-81554. doi: https://doi.org/10.1109/ACCESS.2019.2923707

[7] Mujumdar, A. and Vaidehi, V.V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, pp. 292-299. doi: https://doi.org/10.1016/j.procs.2020.01.047

[8] Sharma, T. and Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine and Art, 4(1), pp. 1-21. doi: https://doi.org/10.1186/s42492-021-00097-7

[9] Haris, U., Kabeer, V. and Afsal, K. (2024). Breast cancer segmentation using hybrid HHO-CS SVM optimization techniques. Multimedia Tools and Applications, 83, pp. 69145-69167. doi: https://doi.org/10.1007/s11042-023-18025-7

[10] World Health Organization (2023). Noncommunicable diseases key facts. Geneva: WHO Media Centre. Available at: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[11] Yadav, A., Jamir, I., Jain, R.R. and Sohani, M. (2019). Comparative study of machine learning algorithms for breast cancer prediction - A review. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(2), pp. 979-985. doi: https://doi.org/10.32628/CSEIT1952278

[12] Haris, U., Kabeer, V. and Afsal, K. (2016). The power of machines in automatic detection of diseases: A review on computer aided detection of breast cancer in digital mammograms. Singularities, 3(2), pp. 88-95. Available at: https://singularitiesjournal.org/wp-content/uploads/2025/08/Singularities-Vol-3-Issue-2-1.pdf.