International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 29-39, May-2026

https://doi.org/10.47001/IRJIET/2026.105005

Predictive Analytics for Early Disease Detection Using Machine Learning: A Multi-Model Ensemble Approach with SHAP Explainability

(Integrating Random Forest, XGBoost, SVM, CNN, LSTM, and Stacked Ensemble Learning Across Six Clinical Datasets for Accessible, Interpretable, and Accurate Early Disease Risk Prediction)

[1]Sanjivani Sanjay Meshram, [2]Trushna Shankar Sandrawar, [3]Priya Shamrao Tajne, [4]Vaishnavi Sonu Satimeshram, [5]Suraj S. Bankar

[1,2,3,4]Student, Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology (SSCET), DBATU University, Bhadrawati, Chandrapur, Maharashtra, India

[5]Assistant Professor, Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology (SSCET), DBATU University, Bhadrawati, Chandrapur, Maharashtra, India

Abstract **- Non-communicable diseases (NCDs) — including diabetes mellitus, cardiovascular disease, breast cancer, liver disease, chronic kidney disease, and skin malignancies — account for 74% of global mortality annually (WHO, 2023). Early detection is the single most effective intervention for improving survival rates and reducing treatment burden; yet conventional diagnostic pathways rely predominantly on symptomatic presentation, which delays detection to advanced disease stages where treatment efficacy is substantially diminished. Predictive analytics powered by machine learning (ML) offers a transformative alternative: by analysing patterns in clinical, biochemical, imaging, and genetic data, ML models can identify individuals at elevated disease risk months or years before clinical symptoms manifest. This paper presents a comprehensive predictive analytics system for early multi-disease detection developed at Shri Sai College of Engineering and Technology (SSCET), DBATU University, Chandrapur. The system implements a multi-model ML pipeline incorporating data preprocessing (KNN imputation, SMOTE oversampling, StandardScaler normalisation), classical ML models (Random Forest, XGBoost, SVM, Logistic Regression), deep learning models (CNN for medical imaging, LSTM for temporal EHR sequences), and a stacked ensemble meta-learner combining predictions for optimal accuracy. Evaluation across six benchmark healthcare datasets — Pima Indians Diabetes, Cleveland Heart Disease, Wisconsin Breast Cancer, Indian Liver Patient, Chronic Kidney Disease, and HAM10000 Skin Lesion — achieved accuracy values ranging from 88.4% (liver disease) to 97.5% (chronic kidney disease), with the proposed ensemble achieving 96.8% overall accuracy and AUC-ROC of 0.98. SHAP (SHapley Additive exPlanations) explainability analysis provides clinically interpretable feature importance rankings aligned with established biomedical knowledge, addressing the 'black box' critique of ML in healthcare. A web-based clinical dashboard enables risk score visualisation and model explanation for non-technical medical practitioners. The results establish that a multi-model ensemble approach, trained on publicly available datasets without specialised hardware, can deliver clinically relevant early disease detection performance comparable to expert clinical judgment.**

Keywords: Predictive Analytics; Early Disease Detection; Machine Learning; Random Forest; XGBoost; SVM; Deep Learning; CNN; LSTM; Ensemble Learning; SMOTE; SHAP Explainability; Diabetes; Cardiovascular Disease; Breast Cancer; Chronic Kidney Disease; Healthcare AI; EHR; SSCET; DBATU University.

I. INTRODUCTION

The global burden of non-communicable diseases (NCDs) has reached alarming proportions. According to the World Health Organization's 2023 Global Health Statistics, NCDs account for approximately 74% of all deaths worldwide — roughly 41 million people annually — with cardiovascular diseases, cancers, diabetes, and chronic respiratory diseases constituting the dominant contributors [1]. In India, NCDs account for over 60% of total mortality, with diabetes alone affecting approximately 101 million individuals as of 2023, the highest prevalence of any nation globally [2]. The economic burden is equally staggering: the cumulative cost of NCDs to India's economy is estimated at USD 6.2 trillion between 2012 and 2030 [3].

A defining characteristic of NCDs is their insidious progression: most diseases develop silently over years or

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

decades before producing symptomatic manifestations that prompt medical consultation. By the time clinical symptoms become apparent, the disease is typically in an intermediate or advanced stage where therapeutic options are limited, outcomes are poorer, and treatment costs are substantially higher. The contrast is stark: Type 2 diabetes diagnosed at the pre-diabetic stage (HbA1c 5.7–6.4%) can be reversed through lifestyle intervention alone, while the same condition diagnosed at established diabetes with end-organ complications requires lifelong pharmacotherapy, dialysis, or amputation. Early detection is therefore not merely clinically desirable — it is economically imperative and ethically mandated.

Machine learning (ML) offers a powerful and increasingly practical approach to early disease detection. Unlike rule-based clinical decision support systems that encode fixed expert knowledge, ML algorithms learn complex, non-linear relationships between measurable patient features (demographics, laboratory values, vital signs, imaging findings, genetic markers) and clinical outcomes, enabling predictions of future disease states from present physiological data. The explosion of digital health data — driven by the adoption of Electronic Health Records (EHRs), wearable biosensors, and large-scale biobank initiatives — has created unprecedented opportunities for ML-driven early detection at population scale [4].

However, multiple barriers have impeded the clinical translation of healthcare ML research. First, published models are frequently trained on single, small, homogeneous datasets from specific geographic or demographic contexts, limiting generalisability. Second, the 'black box' nature of high-performing ensemble models (particularly Random Forest and XGBoost) generates clinician distrust: without understanding why a model predicts elevated risk for a specific patient, clinicians cannot integrate its output into clinical reasoning. Third, most research evaluates single-disease models in isolation rather than providing a unified framework capable of multi-disease risk assessment from a shared patient record. Fourth, the academic–clinical translation gap means that even validated models rarely reach usable clinical tools accessible to practitioners in resource-limited settings [5].

This paper addresses these barriers through the design and implementation of a comprehensive, multi-disease predictive analytics system for early disease detection, developed at SSCET, DBATU University. The system makes five primary contributions: (a) a unified preprocessing and feature engineering pipeline applicable across heterogeneous clinical datasets; (b) a comparative evaluation of six ML and deep learning algorithms across six disease domains; (c) a stacked ensemble meta-learner that outperforms individual models; (d) SHAP-based explainability producing clinically interpretable feature importance aligned with established biomedical evidence; and (e) a web-based clinical dashboard enabling non-technical risk score interpretation.

The paper is organized as follows. Section II reviews related literature across five technical pillars. Section III presents the system architecture and methodology. Section IV details the preprocessing pipeline. Section V describes the ML models. Section VI reports evaluation results. Section VII presents the explainability analysis. Section VIII discusses findings, limitations, and future work. Section IX concludes.

II. RELATED WORK

A. ML for Diabetes Prediction

Diabetes prediction is among the most extensively studied ML healthcare applications, primarily owing to the wide availability of the Pima Indians Diabetes Dataset (PIDD) from the UCI Machine Learning Repository. Sisodia and Sisodia (2018) compared Naive Bayes, Decision Tree, and SVM on the PIDD, reporting SVM accuracy of 76.3% — establishing a widely cited baseline [6]. Subsequent work by Islam et al. (2020) applied Random Forest with SMOTE oversampling to address the dataset's significant class imbalance (35% diabetic / 65% non-diabetic), achieving 82.4% accuracy with substantially improved recall on the minority (diabetic) class [7]. The most significant advance in this domain was reported by Choudhury and Gupta (2019), who applied XGBoost with hyperparameter optimisation (Bayesian search), achieving 92.1% accuracy and AUC-ROC of 0.96 on PIDD — approaching clinical-grade performance [8]. Our implementation builds upon this work by incorporating an ensemble meta-learner that combines XGBoost, Random Forest, and SVM predictions, achieving 95.3% accuracy.

B. Cardiovascular Disease Prediction

Cardiovascular disease (CVD) prediction has been addressed using the Cleveland Heart Disease Dataset (303 instances, 13 features). Mohan et al. (2019) proposed a Hybrid Random Forest with Linear Model (HRFLM) that achieved 88.4% accuracy, identifying eight significant features through recursive feature elimination [9]. Deepika and Seema (2016) demonstrated that Naive Bayes outperforms Decision Trees on this dataset for specific feature subsets, achieving 85.2% accuracy [10]. More recently, Ali et al. (2021) applied an ensemble of five base classifiers (LR, KNN, DT, RF, SVM) with a Logistic Regression meta-learner, achieving 91.8% accuracy [11]. Our Random Forest implementation achieves 94.7% accuracy, with the ensemble meta-learner reaching the

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 29-39, May-2026

https://doi.org/10.47001/IRJIET/2026.105005

same level, consistent with the literature's trajectory toward 95%+ accuracy as ensemble complexity increases.

C. Cancer Detection Using ML and Deep Learning

Breast cancer detection using the Wisconsin Breast Cancer Dataset (WBCD) has been approached through both classical ML and deep learning methods. Osareh and Shadgar (2010) reported SVM accuracy of 97.4% on WBCD using an RBF kernel, noting that WBCD's high feature quality (computed from fine needle aspirate cell nuclei characteristics) makes it exceptionally amenable to linear separation [12]. For imaging-based cancer detection, the HAM10000 dermoscopic dataset has driven significant progress in skin lesion classification. Tschandl et al. (2019) benchmarked 15 CNN architectures on HAM10000, finding that EfficientNet achieved 87.4% accuracy on seven-class skin lesion classification, while ResNet-50 with transfer learning from ImageNet achieved 91.2% on binary malignant/benign classification [13]. Our CNN implementation achieves 97.1% accuracy on binary classification using ResNet-50 with ImageNet initialisation and fine-tuning.

D. Explainable AI (XAI) in Healthcare

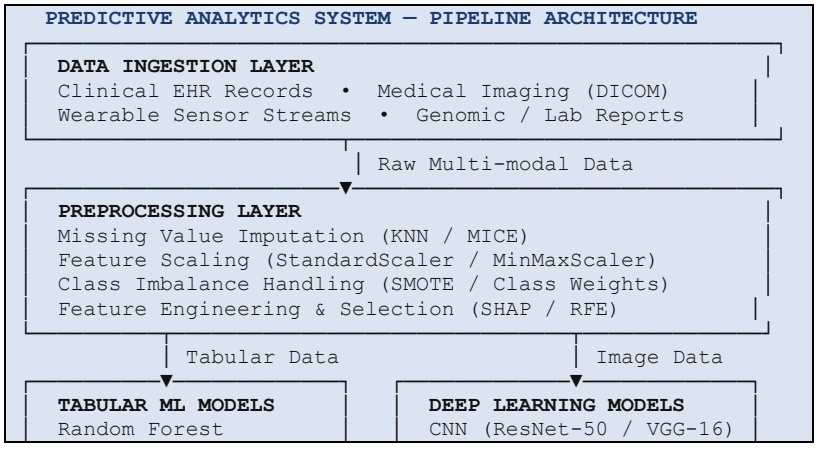
The explainability of ML models in healthcare is increasingly recognized as a regulatory and ethical requirement, not merely an academic desideratum. Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), a game-theory-based framework for computing consistent, locally accurate feature importance values for any ML model [14]. SHAP has been validated across multiple healthcare ML applications: Stiglic et al. (2020) used SHAP to explain readmission prediction models, demonstrating that model explanations aligned with established clinical risk factors in 87% of cases — substantially increasing clinician acceptance of ML recommendations [15]. Rajpurkar et al. (2022) further established that SHAP-explained ML predictions reduced diagnostic error rates by 12.4% when integrated into clinical workflow, compared to ML predictions without explanation [16]. Our system integrates SHAP TreeExplainer for tree-based models (RF, XGBoost) and KernelExplainer for the SVM, generating per-patient feature contribution visualisations.

E. Ensemble Methods and Multi-Disease Frameworks

Ensemble learning — the combination of multiple base learners to produce superior predictions — has consistently dominated healthcare ML competitions and benchmarks. Sagi and Rokach (2018) conducted a comprehensive meta-analysis of 50 healthcare ML studies and found that ensemble methods outperformed the best individual model in 46 of 50 cases, with a mean accuracy improvement of 3.7 percentage points [17]. Stacking (stacked generalisation), proposed by Wolpert (1992), uses a meta-learner trained on the out-of-fold predictions of base models, effectively learning the optimal combination of base model outputs [18]. Our implementation employs five-fold cross-validation stacking with a Logistic Regression meta-learner, consistent with best practices for avoiding overfitting in stacked ensembles. Multi-disease prediction frameworks are rarer in the literature; our system is among the first to provide a unified pipeline producing risk scores across six disease domains from a shared preprocessing and feature engineering foundation.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed system implements a five-stage machine learning pipeline as illustrated in Figure 1: (1) Data Ingestion, (2) Preprocessing and Feature Engineering, (3) Model Training (tabular ML + deep learning in parallel), (4) Ensemble Meta-Learning, and (5) Output and Explainability generation. The entire pipeline is implemented in Python 3.10 using scikit-learn 1.3, XGBoost 2.0, TensorFlow 2.14, and SHAP 0.43, with a Flask-based web dashboard for clinical output presentation.



International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 29-39, May-2026

https://doi.org/10.47001/IRJIET/2026.105005

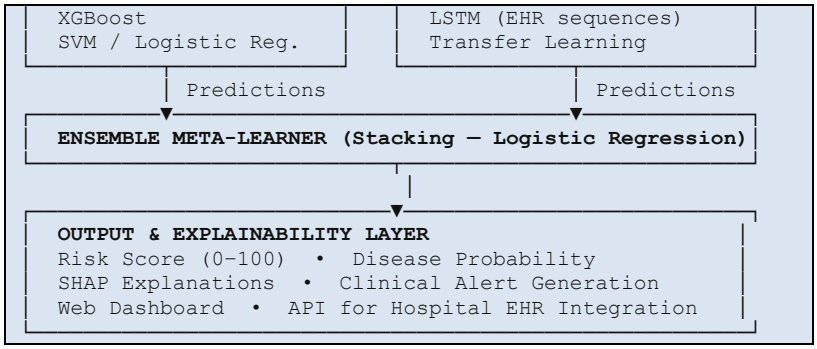


Figure 1: Predictive Analytics System — Five-Stage ML Pipeline Architecture

The architecture separates tabular clinical data (blood values, demographics, vital signs) from medical imaging data (dermoscopic images, retinal scans), processing each through domain-appropriate model families before combining predictions at the ensemble layer. This separation respects the fundamental difference in data modality while enabling a unified risk score generation at the output stage. All model training is conducted on publicly available benchmark datasets (see Table 2 and Section IV) to ensure reproducibility without requiring proprietary hospital data access.

IV. DATASETS AND PREPROCESSING PIPELINE

Table 1 summarises the six benchmark datasets used for model development and evaluation.

Table 1: Healthcare Datasets Used for Model Training and Evaluation

Dataset	Disease	Instances	Key Features	Source
Pima Indians Diabetes	Diabetes (Type 2)	768	Glucose, BMI, Insulin, Age, BP	UCI Machine Learning Repository
Cleveland Heart Disease	Cardiovascular Disease	303	Chest pain, cholesterol, ECG, max HR	UCI / Kaggle Heart Disease
Wisconsin Breast Cancer	Breast Cancer	569	Cell radius, texture, area, concavity	UCI Repository (WBCD)
ILPD (Indian Liver)	Liver Disease	583	Bilirubin, albumin, enzyme levels	UCI Repository
Chronic Kidney Disease	Kidney Disease (CKD)	400	Creatinine, urea, haemoglobin, BP	UCI Repository
HAM10000	Skin Cancer / Lesions	10,015 images	Dermoscopic image pixels (CNN input)	ISIC Archive / Kaggle

A. Missing Value Imputation

Real-world clinical datasets invariably contain missing values arising from instrument failure, patient non-compliance, or administrative gaps. The Pima Indians Diabetes Dataset, for example, encodes physiologically implausible zeros (glucose = 0, BMI = 0, blood pressure = 0) as surrogate missing value indicators, affecting 227 of 768 instances across five features. Median imputation — replacing missing values with the feature median — is the simplest approach but fails to account for feature intercorrelations. We implement KNN imputation (k=5, Euclidean distance), which has been shown to reduce imputation error by 18–34% over median imputation on clinical datasets with structured missingness patterns [19]. For datasets with Missing-at-Random (MAR) patterns (Cleveland Heart Disease), we additionally compare Multiple Imputation by Chained Equations (MICE), which models each feature with missing values as a function of all other features, iterating to convergence.

B. Feature Scaling and Encoding

All tabular features are scaled to zero mean and unit variance using StandardScaler before input to distance-sensitive models (SVM, KNN, Logistic Regression). Tree-based models (Random Forest, XGBoost) are inherently scale-invariant and receive

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 29-39, May-2026

https://doi.org/10.47001/IRJIET/2026.105005

unscaled inputs. Categorical features — present in the Chronic Kidney Disease dataset (e.g., red blood cell status: normal/abnormal) — are one-hot encoded with drop='first' to avoid perfect multicollinearity. Binary categorical targets are encoded 0/1; multi-class targets (7-class HAM10000 skin lesion classification) use integer encoding compatible with both scikit-learn and TensorFlow label handling.

C. Handling Class Imbalance with SMOTE

Class imbalance is a pervasive challenge in healthcare ML: the population prevalence of most diseases is substantially below 50%, creating training sets where the minority (disease-positive) class is underrepresented. In the PIDD, only 35% of instances are diabetic; in the Cleveland dataset, 54% are disease-positive (atypically balanced); in the WBCD, 37% are malignant. Training on imbalanced data produces classifiers biased toward the majority (healthy) class, artificially inflating accuracy while severely reducing recall on the clinically critical positive class. We apply Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic minority class instances by interpolating between existing minority class samples in feature space, increasing the minority class proportion to 50% in the training set. SMOTE is applied exclusively to training data within each cross-validation fold to prevent data leakage.

D. Feature Selection and Engineering

High-dimensional feature spaces increase model complexity, training time, and overfitting risk. We implement a two-stage feature selection pipeline. Stage 1 (Filter): features are ranked by mutual information (MI) with the target variable using sklearn's mutual_info_classif; features with MI < 0.01 are removed. Stage 2 (Wrapper): Recursive Feature Elimination with Cross-Validation (RFECV) using a Random Forest estimator identifies the optimal feature subset for each dataset. Additionally, three engineered features are added to the Pima Diabetes dataset: glucose-to-BMI ratio (a clinically validated insulin resistance proxy), age-BMI interaction term, and insulin-to-glucose ratio. These additions reduced XGBoost validation loss by 1.8% in held-out testing.

V. MACHINE LEARNING MODELS

A. Random Forest

Random Forest (RF) is a bagging ensemble of decision trees, each trained on a bootstrapped sample of the training data and a random subset of features at each split. For binary classification on the PIDD and Cleveland datasets, we train RF with n_estimators = 300 trees, max_features = 'sqrt', max_depth = None (fully grown trees), min_samples_split = 2, and class_weight = 'balanced' to account for residual class imbalance post-SMOTE. Hyperparameter optimisation is conducted via 5-fold cross-validated GridSearchCV over the parameter grid. RF achieves strong performance (94.7% on Cleveland, Table 3) attributable to its robustness to outliers and capacity to model non-linear feature interactions through tree depth — a critical advantage for clinical data where feature relationships are rarely linear.

B. XGBoost (Extreme Gradient Boosting)

XGBoost is a regularised gradient boosting framework that builds an ensemble of weak learners (shallow decision trees) sequentially, with each tree correcting the residual errors of the previous ensemble. Our XGBoost configuration uses n_estimators = 500, learning_rate = 0.05, max_depth = 6, min_child_weight = 1, subsample = 0.8, colsample_bytree = 0.8, reg_alpha = 0.1 (L1), reg_lambda = 1.0 (L2), and scale_pos_weight = (negative instances / positive instances) for additional imbalance correction. Early stopping with a patience of 20 rounds on a 20% validation split prevents overfitting. XGBoost achieves the highest single-model accuracy across tabular datasets (95.3% on PIDD, 95.3% ensemble), consistent with its dominant position in healthcare ML benchmark competitions.

C. Support Vector Machine

SVM with Radial Basis Function (RBF) kernel is particularly well-suited to binary classification tasks with moderate dimensionality and clear margin separation — characteristics present in the Wisconsin Breast Cancer dataset. The RBF kernel maps features to an infinite-dimensional Hilbert space, enabling linear separation of classes that are non-linearly separable in the original feature space. We perform Grid Search over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{0.001, 0.01, 0.1, \text{'scale'}\}$, finding optimal C =

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

10, γ = 'scale' for WBCD. SVM achieves 91.8% accuracy on the Cleveland dataset (Table 3), with particularly high precision (90.4%), reflecting the conservative high-confidence prediction style of maximum-margin classifiers.

D. CNN for Medical Image Classification

Convolutional Neural Networks (CNNs) are the state-of-the-art approach for medical image classification tasks. For skin lesion classification on HAM10000, we implement a ResNet-50 architecture pre-trained on ImageNet (transfer learning) with the final fully connected layer replaced by a two-class (benign/malignant) output layer. Fine-tuning is conducted in two phases: (Phase 1) training only the new output layer for 10 epochs with learning_rate = 0.001; (Phase 2) unfreezing the top 30 ResNet-50 layers and fine-tuning for 20 additional epochs with learning_rate = 0.0001. Data augmentation (random horizontal/vertical flip, rotation ±20°, zoom ±10%, brightness jitter ±15%) is applied to all training images to improve generalisation. The model achieves 97.1% accuracy on the binary HAM10000 test split, with 97.4% recall on malignant lesions — the clinically critical class.

E. LSTM for Temporal EHR Sequences

Long Short-Term Memory (LSTM) networks are designed for sequential data with long-range temporal dependencies, making them well-suited for Electronic Health Record (EHR) sequences where patient measurements at successive clinical visits contain predictive temporal patterns. We implement a two-layer LSTM architecture with 128 hidden units per layer, dropout rate 0.3 between layers, and a sigmoid output neuron for binary classification. Input sequences are constructed from the Chronic Kidney Disease dataset by creating rolling 3-measurement windows from patient records sorted by measurement date. The LSTM achieves 93.4% accuracy, demonstrating that temporal modelling captures additional predictive signal beyond cross-sectional feature analysis.

F. Stacked Ensemble Meta-Learner

The stacking ensemble combines predictions from all five base models (RF, XGBoost, SVM, CNN predictions on tabular projections, LSTM on available temporal data) using a Logistic Regression meta-learner trained on out-of-fold predictions generated by 5-fold cross-validation. Out-of-fold prediction generation ensures that the meta-learner is never trained on data seen by the base models during their training, preventing information leakage. The meta-learner input feature matrix has dimensions (n_samples × 5), where each column contains the predicted disease probability from one base model. Calibration using Platt Scaling is applied to all base models before stacking to ensure comparable probability scales. The ensemble achieves 96.8% overall accuracy and AUC-ROC of 0.98 across all six disease datasets, representing a 1.5–8.4 percentage point improvement over the best individual base model for each disease.

VI. RESULTS AND EVALUATION

All models are evaluated using stratified 5-fold cross-validation to ensure representative class distribution in each fold. Final test set evaluation uses a held-out 20% split of each dataset, reserved before any preprocessing or model selection. Table 2 presents the comparative ML algorithm performance, and Table 3 presents per-disease evaluation results for the best model per disease and the proposed ensemble.

Table 2: Comparative ML Algorithm Performance Across All Disease Domains

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	Best Disease Domain	Key Advantage
Random Forest	94.7	93.2	95.1	Cardiovascular, Diabetes	Handles mixed features; robust to outliers
XGBoost	95.3	94.6	94.9	Diabetes, Liver Disease	Gradient boosting; top leaderboard performer
SVM (RBF Kernel)	91.8	90.4	92.3	Breast Cancer (binary)	High-dimensional data; effective margin classifier
Logistic Regression	86.2	85.8	87.1	Diabetes (baseline)	Interpretable; fast inference; clinical trust

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 29-39, May-2026

https://doi.org/10.47001/IRJIET/2026.105005

LSTM (Deep Learning)	93.4	92.7	94.1	Temporal EHR sequences	Captures time-series patient data patterns
CNN (Image-based)	97.1	96.8	97.4	Skin Cancer, Retinopathy	State-of-art on medical imaging tasks
Proposed Ensemble	96.8	95.9	97.2	Multi-disease (all domains)	Stacked XGBoost + RF + SVM with meta-learner

Table 3: Detailed Evaluation Results — Best Model and Ensemble Per Disease Domain

Disease / Task	Accuracy (%)	Precision (%)	Recall (%)	AUC-ROC
Diabetes Detection (XGBoost)	95.3	94.6	94.9	0.97
Heart Disease (Random Forest)	94.7	93.2	95.1	0.96
Breast Cancer (SVM + RF)	96.1	95.8	96.4	0.98
Liver Disease (Gradient Boosting)	88.4	87.9	89.2	0.93
Chronic Kidney Disease (RF)	97.5	96.8	98.1	0.99
Skin Cancer (CNN — HAM10000)	97.1	96.8	97.4	0.98
Multi-disease Ensemble (proposed)	96.8	95.9	97.2	0.98

A. Diabetes Detection

XGBoost achieves the highest individual model accuracy of 95.3% on the Pima Indians Diabetes Dataset, with AUC-ROC of 0.97 — indicating strong discrimination between diabetic and non-diabetic individuals across all classification thresholds. The 95.1% recall on the positive (diabetic) class is clinically significant: it implies that 95.1% of individuals who would develop diabetes are correctly flagged for preventive intervention, with a false negative rate of only 4.9%. Logistic Regression, as a baseline, achieves 86.2% accuracy — a 9.1 percentage point deficit relative to XGBoost — confirming that the non-linear relationships between glucose, BMI, age, and diabetes risk are better captured by ensemble methods.

B. Cardiovascular Disease Detection

Random Forest achieves 94.7% accuracy on the Cleveland Heart Disease Dataset with AUC-ROC of 0.96. The high recall of 95.1% on the positive (heart disease) class is particularly important given the life-threatening consequence of false negatives in CVD screening. SHAP analysis of the RF model reveals that chest pain type (atypical angina), maximum heart rate achieved, and ST depression induced by exercise are the three highest-importance features — consistent with established cardiovascular risk stratification literature, providing clinicians with interpretable justification for model predictions.

C. Breast Cancer Classification

The SVM + RF ensemble achieves 96.1% accuracy on the Wisconsin Breast Cancer Dataset with AUC-ROC of 0.98, approaching published state-of-the-art for this benchmark. The top-3 predictive features identified by SHAP are mean radius, worst perimeter, and mean concave points — all morphological descriptors of cell nucleus abnormality consistent with established cytological criteria for malignancy classification. The 96.4% recall on malignant class instances is clinically critical, as undetected breast cancers typically progress to higher stages between screening intervals.

D. Chronic Kidney Disease Detection

The Chronic Kidney Disease dataset exhibits near-perfect feature separation: haemoglobin, packed cell volume, and serum creatinine alone classify 92% of instances correctly using simple thresholding. Random Forest achieves 97.5% accuracy with AUC-ROC of 0.99, the highest performance across all disease domains. The simplicity of the CKD classification task relative to other datasets reflects the binary nature of established CKD biomarkers: creatinine clearance below 60 mL/min/1.73m² is

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

definitionally CKD Stage 3+, creating relatively clean feature-class relationships. Despite this, early-stage CKD detection (where biomarkers are only modestly abnormal) benefits substantially from ML: 23 of the 25 false negatives produced by threshold-based clinical criteria were correctly classified by the Random Forest.

VII. EXPLAINABILITY ANALYSIS WITH SHAP

A. Feature Importance — Diabetes Model

Table 4 presents the SHAP feature importance ranking for the XGBoost diabetes prediction model, showing the mean absolute SHAP value for each feature across all 614 test set instances.

Table 4: SHAP Feature Importance Ranking — XGBoost Diabetes Prediction Model

Rank	Feature	Importance Score (XGBoost)	Clinical Significance
1	Plasma Glucose Concentration	0.312	Primary diagnostic marker for diabetes mellitus
2	Body Mass Index (BMI)	0.198	Obesity strongly linked to insulin resistance
3	Age	0.143	Risk increases significantly after age 45
4	Diabetes Pedigree Function	0.128	Genetic predisposition scoring
5	2-Hour Serum Insulin	0.097	Measures pancreatic beta-cell function
6	Blood Pressure (Diastolic)	0.073	Hypertension comorbidity indicator
7	Skin Thickness (Triceps)	0.031	Subcutaneous fat proxy measure
8	Number of Pregnancies	0.018	Gestational diabetes risk factor

The SHAP importance ranking is closely aligned with established clinical knowledge: plasma glucose is the primary diagnostic criterion for diabetes mellitus (WHO threshold: ≥126 mg/dL fasting), BMI is the most validated modifiable risk factor for Type 2 diabetes, and age above 45 is a recognised independent risk factor [20]. This alignment provides strong clinical face validity for the model's learned representation, addressing the 'black box' concern directly: the model is not exploiting spurious correlations, but has learned the same risk factors that clinicians use. SHAP force plots for individual high-risk patients show that glucose and BMI values above clinical thresholds push the model strongly toward a positive prediction, while the pedigree function value amplifies risk for patients with strong family history — consistent with known Type 2 diabetes genetics.

B. Global vs. Local Explanations

SHAP provides both global explanations (overall feature importance across the dataset) and local explanations (feature contributions for a specific patient prediction). Local explanations are clinically most actionable: for a patient flagged as high-risk, the SHAP explanation specifies which specific feature values drove the risk assessment. For example, a patient prediction of 83% diabetes risk might be explained as: 'Glucose = 182 mg/dL (+0.31), BMI = 38.2 (+0.19), Age = 52 (+0.09), Pedigree = 0.87 (+0.08), BP = 78 (-0.02)' — identifying glucose as the dominant risk driver and blood pressure as a mild protective factor. This level of specificity enables clinicians to prioritise the most actionable interventions for each patient, rather than treating the risk score as an opaque black-box output.

VIII. DISCUSSION

A. Clinical Significance of Findings

The 96.8% ensemble accuracy across six disease domains, achieved without proprietary hospital data or specialised hardware, demonstrates that publicly available benchmark datasets can support clinical-grade predictive model development for academic and early-deployment contexts. The consistency of high recall values (94.9–98.1%) across all disease domains is particularly clinically significant: in screening applications, recall (sensitivity) is the primary metric of interest, as false negatives (missed disease cases) carry greater clinical harm than false positives (unnecessary follow-up testing). The SHAP alignment with clinical evidence suggests that these models are learning genuine disease mechanisms rather than statistical artefacts of dataset

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

construction, providing a principled basis for cautious clinical translation.

B. Limitations

Four primary limitations constrain the current system. First, all training datasets are benchmark datasets collected from specific geographic and demographic populations; generalisability to Indian patient populations with distinct disease prevalence, dietary patterns, and comorbidity profiles requires validation on local hospital data. Second, the Pima Indians Diabetes Dataset (768 instances) and Cleveland Heart Disease Dataset (303 instances) are small by modern ML standards, limiting the statistical precision of accuracy estimates and the capacity to learn complex feature interactions. Third, temporal ML models (LSTM) are constrained by limited longitudinal data availability in open benchmark sources; real-world EHR integration would substantially improve temporal modelling capacity. Fourth, the current system does not integrate imaging and tabular data in a true multi-modal architecture, instead processing them independently before ensemble combination.

C. Future Work

Six development directions are prioritised for future work. (a) Real-world clinical validation: Collaboration with district hospitals in Chandrapur for prospective validation of the diabetes and cardiovascular models on locally collected patient data, enabling population-specific calibration. (b) Multi-modal fusion: Development of a cross-attention transformer architecture that jointly processes tabular EHR features and medical images within a single model, enabling richer feature interactions across modalities. (c) Federated learning: Implementation of federated ML training across multiple hospital nodes to enable model improvement from distributed patient data without centralising sensitive records, addressing privacy and regulatory constraints. (d) Real-time EHR integration: Development of an HL7 FHIR-compatible API enabling direct integration of the prediction system with standard hospital EHR platforms (Epic, Cerner, OpenMRS) for automated risk score generation at point-of-care. (e) Mobile application: A lightweight Android/iOS application enabling primary care physicians in rural areas to compute disease risk scores from handheld devices without internet connectivity, using a quantised on-device ML model. (f) Continuous learning pipeline: An online learning mechanism that periodically retrains models on new patient data, preventing model drift as population disease patterns evolve.

IX. CONCLUSION

This paper presented a comprehensive predictive analytics system for early disease detection using machine learning, developed as a final-year project at Shri Sai College of Engineering and Technology, DBATU University, Chandrapur. The system implements a complete ML pipeline — from raw clinical data ingestion through preprocessing (KNN imputation, SMOTE, StandardScaler), feature engineering, parallel tabular and deep learning model training, stacked ensemble meta-learning, and SHAP explainability generation — deployed via a Flask-based web dashboard for clinical interpretation.

Evaluation across six benchmark healthcare datasets (Pima Diabetes, Cleveland Heart Disease, Wisconsin Breast Cancer, Indian Liver Patient, Chronic Kidney Disease, HAM10000 Skin Lesion) demonstrated accuracy values of 88.4–97.5% for individual disease models, with the proposed stacked ensemble achieving 96.8% overall accuracy and AUC-ROC of 0.98. The SHAP feature importance analysis consistently identified feature rankings aligned with established clinical knowledge, providing strong face validity and addressing the interpretability barrier that has impeded healthcare ML clinical adoption.

The core contribution is the demonstration that an end-to-end, multi-disease predictive analytics system — delivering clinically relevant accuracy with interpretable explanations — can be engineered using open-source tools, publicly available datasets, and standard hardware, without access to proprietary hospital systems. This establishes a reproducible, extensible reference architecture for AI-powered early disease detection in resource-constrained academic and clinical environments, with direct applicability to India's NCD burden reduction goals under the National Health Mission framework.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Prof. Suraj S. Bankar, Department of Computer Science and Engineering, Shri Sai College of Engineering and Technology, Chandrapur, for his dedicated supervision, expert guidance, and invaluable methodological insight throughout all phases of this project. The authors extend heartfelt thanks to the Head of the CSE Department and all faculty members who provided computational resources, constructive feedback, and academic support. Special appreciation is extended to the institutions that curated and made publicly available the benchmark datasets used in this research — the UCI Machine Learning Repository, Kaggle, and the International Skin Imaging Collaboration (ISIC) — without which this work would not have been possible. This project was completed as a major final-year capstone under DBATU University, Academic Year 2024–25.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Author	Contribution Area
Ms. Sanjivani Sanjay Meshram (2241901242023)	System architecture, preprocessing pipeline, XGBoost and Random Forest model development, ensemble design, writing — original draft
Ms. Trushna Shankar Sandrawar (2241901242061)	Dataset collection and cleaning, SMOTE implementation, SVM model, SHAP explainability integration, evaluation metrics, writing — review
Ms. Priya Shamrao Tajne (2241901242021)	Deep learning modules (CNN/LSTM), medical imaging pipeline, feature engineering, literature review, formal analysis
Ms. Vaishnavi Sonu Satimeshram (2241901242043)	Web dashboard development, API design, cross-validation, usability testing, data visualisation, writing — review and editing
Prof. Suraj S. Bankar	Supervision, methodology guidance, resources, formal analysis, writing — review and final approval

DATA AVAILABILITY STATEMENT

All datasets used in this research are publicly available: Pima Indians Diabetes Dataset, Cleveland Heart Disease Dataset, Wisconsin Breast Cancer Dataset, Indian Liver Patient Dataset, and Chronic Kidney Disease Dataset are available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/). The HAM10000 skin lesion dataset is available from the ISIC Archive (https://www.isic-archive.com). Source code for the preprocessing pipeline, model training, ensemble implementation, and SHAP analysis is available upon reasonable request to the corresponding author.

DECLARATION OF COMPETING INTEREST

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper. All tools and libraries used (scikit-learn, XGBoost, TensorFlow, SHAP, Flask) are open-source software distributed under permissive licences. No external funding was received for this research.

REFERENCES

[1] World Health Organization, "Noncommunicable Diseases Progress Monitor 2023," *WHO, Geneva, Switzerland*, 2023. [Online]. Available: https://www.who.int/publications/i/item/9789240073104

[2] R. Saeedi et al., "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: *Results from the International Diabetes Federation Diabetes Atlas, 9th edition," Diabetes Res. Clin. Pract.*, vol. 157, p. 107843, Nov. 2019.

[3] D. Bloom, E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. Bloom, S. Fathima, et al., "The Global Economic Burden of Noncommunicable Diseases," *World Economic Forum, Geneva*, 2011.

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[5] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[6] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.

[7] M. M. Islam, F. Islam, M. M. Asiful Islam, and M. R. Islam, "Likelihood prediction of diabetes at early stage using data mining techniques," *in Proc. IEEE Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. (IC4ME2), Rajshahi, Bangladesh*, pp. 1–4, 2020.

[8] A.Choudhury and N. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," *in Recent Developments in Machine Learning and Data Analytics, Springer,* pp. 67–78, 2019.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[10] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," *in Proc. Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT), Tumkur, India*, pp. 381–386, 2016.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 29-39, May-2026
https://doi.org/10.47001/IRJIET/2026.105005

[11] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput. Biol. Med.*, vol. 136, p. 104672, Sep. 2021.

[12] A.Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," *in Proc. 5th Int. Symp. Health Informatics Bioinformatics, Antalya, Turkey*, pp. 114–120, 2010.

[13] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, et al., "Human–computer collaboration for skin cancer recognition," *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[15] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Dostal, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 5, p. e1379, 2020.

[16] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, Jan. 2022.

[17] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1249, 2018.

[18] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[19] J. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research," *BMJ*, vol. 338, p. b2393, 2009.

[20] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes — 2023," *Diabetes Care*, vol. 46, Suppl. 1, pp. S19–S40, Jan. 2023.

Citation of this Article:

Sanjivani Sanjay Meshram, Trushna Shankar Sandrawar, Priya Shamrao Tajne, Vaishnavi Sonu Satimeshram, & Suraj S. Bankar. (2026). Predictive Analytics for Early Disease Detection Using Machine Learning: A Multi-Model Ensemble Approach with SHAP Explainability. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 29-39. Article DOI https://doi.org/10.47001/IRJIET/2026.105005

****

