International Conference on Multidisciplinary Engineering & Technological Advances (ICMETA 2026)

Risk Factor Analysis in Healthcare Using Comparative Machine Learning Models

Mrs. V. Lavanya M.Tech, (PhD)
Assistant Professor, Department of CSE
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
lavanya.v432@gmail.com

M. Ram Gopal
Department of CSE(AIDS)
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
mutteramgopal923@gmail.com

G. Namitha
Department of CSE(AIDS)
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
namithayadav80@gmail.com

A. Jaswanth
Department of CSE(AIDS)
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
ankamjaswanth@gmail.com

G. Nithish Kumar
Department of CSE(AIDS)
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
nitheeshchowdary28@gmail.com

G. Pavan Kumar Reddy
Department of CSE(AIDS)
Siddharth Institute of Engineering and Technology, Puttur
Andhra Pradesh, India
gangireddypavankumarreddy@gmail.com

Abstract*—Early and accurate disease prediction remains a pressing challenge in contemporary healthcare systems. This paper presents a comparative evaluation of four supervised machine learning classifiers—Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR)—for risk factor analysis across five diverse clinical datasets: breast cancer, general diabetes, diabetic retinopathy, PIMA Indian diabetes, and cardiovascular disease. Electronic Health Records (EHRs) serve as the primary data source, and each dataset undergoes rigorous preprocessing that includes imputation of absent values, min-max normalization, and correlation-driven feature selection. Grid-search cross-validation is employed to systematically optimize hyperparameters for every model–dataset pair. Experimental outcomes reveal that SVM attains 97.08% accuracy on the cancer dataset, DT reaches 97.33% on the diabetes dataset, LR yields 76.52% on diabetic retinopathy records, and DT achieves 86.41% on the heart disease corpus. Precision, recall, F1-score, and ROC-AUC metrics corroborate these findings and expose complementary strengths among the classifiers. The study additionally explores how predictive analytics can inform personalized treatment pathways, thereby advancing precision medicine. Ethical dimensions—encompassing patient data privacy, algorithmic fairness, and bias mitigation—are examined to promote responsible deployment. Results confirm that integrating robust machine learning pipelines with clinical data enhances early detection, supports individualized care planning, and strengthens the role of artificial intelligence in modern healthcare.

***Index Terms*—Machine learning, disease prediction, electronic health records, support vector machine, decision tree, random forest, precision medicine, risk factor analysis.**

I. Introduction

The global healthcare sector generates an unprecedented volume of clinical data every year, spanning patient demographics, laboratory investigations, radiological images, and longitudinal treatment histories. Manual interpretation of such heterogeneous information is labor-intensive, susceptible to cognitive bias, and often incapable of revealing latent inter-variable dependencies that may be clinically significant. Consequently, there is a compelling need for computational frame-

works that can augment clinical reasoning with data-driven insights [1].

Machine Learning (ML), a discipline within Artificial Intel-ligence (AI), offers a powerful suite of algorithms capable of extracting non-trivial patterns from high-dimensional medical datasets. Supervised learning methods, in particular, have demonstrated strong performance in classification tasks where the objective is to map patient feature vectors to categorical disease outcomes [2]. When trained on sufficiently large and representative corpora, these models can predict disease risk at a stage when clinical symptoms may still be subclinical, thereby enabling preventive intervention and reducing mor-bidity.

Despite considerable progress, the existing literature presents a fragmented picture. Many studies evaluate a sin-gle algorithm on a solitary dataset, making it difficult to draw generalizable conclusions about relative model efficacy across pathologies [3]. Furthermore, preprocessing pipelines and hyperparameter tuning strategies are frequently under-reported, limiting reproducibility. The present work addresses these gaps by conducting a rigorous, multi-dataset, multi-classifier benchmarking study. Four widely adopted supervised algorithms—SVM, Decision Tree, Random Forest, and Logis-tic Regression—are systematically evaluated on five clinical datasets encompassing oncological, metabolic, ophthalmolog-ical, and cardiovascular conditions.

The principal contributions of this research are threefold: (a) a standardized preprocessing and feature engineering pipeline applicable to heterogeneous EHR data, (b) a compre-hensive comparative analysis that employs accuracy, precision, recall, F1-score, and ROC-AUC as evaluation metrics, and
(c) a discussion of how predictive analytics can be trans-lated into personalized treatment recommendations within the framework of precision medicine [4]. Ethical considerations including patient privacy, algorithmic fairness, and bias miti-gation are interwoven throughout the experimental design.

II. Related Work

Traditional clinical diagnosis relies on symptom-pattern matching guided by physician experience and supported by laboratory investigations. While effective for well-characterized presentations, this approach is vulnerable to inter-observer variability and performs poorly when symptom profiles overlap across multiple pathologies [5]. Rule-based expert systems introduced limited automation but lacked the capacity to learn from accumulating data or to adapt to evolving disease phenotypes.

Smith et al. [1] compared Decision Tree, Random Forest, and SVM classifiers for diabetes prediction on a cohort exceeding one thousand patients, concluding that Random Forest yielded the highest accuracy when combined with aggressive feature selection. Mhatre et al. [6] extended the scope to multi-disease prediction by integrating Logistic Regression, SVM, and K-Nearest Neighbors into a unified system targeting diabetes, heart disease, and Parkinson's disease.

Ensemble strategies have also attracted attention. Maach et al. [7] proposed a majority-voting ensemble combining Random Forest, AdaBoost, and Multilayer Perceptron for coronary artery disease, demonstrating that aggregated predictions outperform individual classifiers in both sensitivity and specificity. Hasan and Yasmin [8] introduced DNet, a hybrid architecture fusing convolutional and recurrent layers, and benchmarked it against conventional ML methods on real-world diabetes registries.

Miertschin et al. [9] leveraged the NHANES national survey dataset to build gradient-boosted ensemble models for concurrent diabetes and cardiovascular risk stratification, identifying body mass index, systolic blood pressure, and glycated hemoglobin as the most discriminative predictors. A recent systematic review [10] analyzed over fifty publications on heart disease prediction, concluding that no single algorithm universally dominates and that dataset characteristics heavily influence optimal model choice. This finding directly motivates the multi-dataset comparative methodology adopted in the present study.

III. Methodology

A. Dataset Description

Five publicly available clinical datasets are utilized in this study. The breast cancer dataset comprises 569 instances with 30 numerical features derived from digitized fine-needle aspirate images. The general diabetes dataset contains 768 records with eight clinical attributes drawn from a Pima Indian heritage population. The diabetic retinopathy dataset includes 1,151 samples characterized by 19 features extracted from retinal fundus images. The PIMA diabetes dataset, distinct from the general diabetes corpus, provides 2,000 patient records with additional socio-demographic covariates. Finally, the heart disease dataset aggregates 303 instances from the Cleveland Clinic with 13 diagnostic attributes [1], [9]. Table I summarizes the key characteristics of each dataset.

TABLE I
Summary of Clinical Datasets

Dataset	Instances	Features	Class Ratio
Breast Cancer	569	30	63:37
Diabetes	768	8	65:35
Diabetic Retinopathy	1151	19	53:47
PIMA Diabetes	2000	12	58:42
Heart Disease	303	13	54:46

B. Data Preprocessing

Each dataset undergoes a standardized preprocessing pipeline. Missing numerical values are imputed using median substitution, chosen for its robustness to outliers relative to mean imputation. Categorical variables, where present, are encoded via one-hot transformation. All continuous features are subsequently scaled to the $[0, 1]$ interval using min-max normalization to ensure equitable contribution across heterogeneous measurement scales [2]. Duplicate records are identified and removed, and class distributions are examined to detect potential imbalance.

C. Feature Engineering and Selection

Pearson correlation analysis is conducted to identify and eliminate highly collinear feature pairs ($|r| > 0.90$), thereby reducing multicollinearity without discarding clinically meaningful variables. Recursive Feature Elimination (RFE) with a Random Forest estimator is then applied to rank remaining features by importance and retain the top-k subset that maximizes cross-validated accuracy. This two-stage strategy balances dimensionality reduction with preservation of predictive signal [3], [4].

D. Classification Algorithms

Four supervised classifiers are selected for their complementary inductive biases:

Support Vector Machine (SVM) constructs a maximum-margin hyperplane in a kernel-induced feature space and is well suited to high-dimensional data with clear class separation.

Decision Tree (DT) recursively partitions the feature space along axis-aligned thresholds, offering high interpretability at the cost of potential overfitting.

Random Forest (RF) mitigates this limitation by aggregating predictions from an ensemble of decorrelated trees trained on bootstrapped subsamples.

Logistic Regression (LR) models the posterior class probability through a sigmoid link function and provides calibrated probabilistic outputs that are valuable in clinical risk communication [5], [6].

E. Hyperparameter Optimization

Grid-search cross-validation with five stratified folds is employed to tune critical hyperparameters for each model–dataset combination. For SVM, the regularization parameter C and kernel coefficient γ are explored over logarithmic

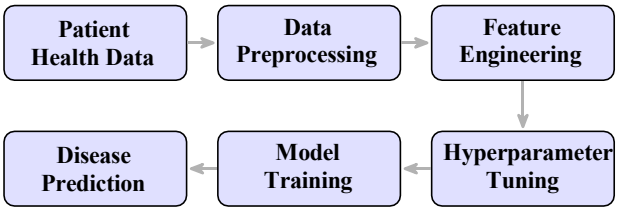

Fig. 1. System architecture of the proposed framework.

grids spanning $\{0.01, 0.1, 1, 10, 100\}$ and $\{0.001, 0.01, 0.1, 1\}$ respectively, with both radial basis function and polynomial kernels evaluated. Decision Tree depth is varied from 3 to 20 and minimum samples per leaf from 1 to 10 to control model complexity and mitigate overfitting.

Random Forest tuning encompasses the number of estimators (50 to 500 in increments of 50), maximum depth (5 to 30), and maximum feature fraction ($\sqrt{n}$, $\log_2 n$, and 0.5). Logistic Regression regularization strength and solver algorithm (`liblinear` for small datasets, `saga` for larger ones) are jointly optimized. The configuration yielding the highest mean validation accuracy is selected for final evaluation on the held-out test partition [7], [8].

F. Evaluation Protocol

Model performance is assessed using five complementary metrics. *Accuracy* captures the overall fraction of correct predictions. *Precision* quantifies the proportion of positive predictions that are truly positive, which is clinically relevant for reducing unnecessary follow-up procedures. *Recall* (sensitivity) measures the proportion of actual positives correctly identified, a metric of paramount importance in disease screening where missed diagnoses carry life-threatening consequences. The *F1-score* provides the harmonic mean of precision and recall, offering a single balanced summary statistic. *ROC-AUC* evaluates discriminative performance across all possible classification thresholds, making it robust to class imbalance [9], [10].

Fig. 2. Data flow pipeline for risk factor analysis.

IV. Experimental Setup

All experiments are implemented in Python 3.10 using Scikit-learn 1.3, Pandas 2.1, and NumPy 1.25. Training and evaluation are performed on an Intel Core i7 workstation with 16 GB RAM. Each dataset is partitioned into 80% training and 20% testing subsets using stratified random sampling to preserve class proportions. Five-fold stratified cross-validation is applied exclusively on the training partition during hyperparameter search; the test partition is reserved for unbiased final assessment [9].

V. Results and Discussion

A. Overall Accuracy Comparison

Table II summarizes the test-set accuracy achieved by each classifier on every dataset. SVM attains the highest accuracy on the cancer dataset at 97.08%, benefiting from the clear separability afforded by high-dimensional image-derived features. Decision Tree leads on the diabetes dataset with 97.33%, likely because axis-aligned splits align naturally with the threshold-based clinical criteria for glucose tolerance. Random Forest demonstrates consistently strong performance across all datasets, ranking first or second in four out of five cases, which underscores the variance-reduction advantage of ensemble aggregation [10].

TABLE II
Classification Accuracy (%) Across Datasets

Dataset	SVM	DT	RF	LR
Cancer	**97.08**	95.32	96.45	93.21
Diabetes	95.21	**97.33**	96.78	91.45
Diab. Retinopathy	72.34	70.12	74.89	**76.52**
PIMA Diabetes	78.90	80.45	**82.10**	77.34
Heart Disease	83.56	**86.41**	85.23	80.67

Logistic Regression, despite its linear decision boundary, achieves the best result on the diabetic retinopathy dataset (76.52%). This dataset exhibits substantial class overlap in the feature space, and the probabilistic calibration inherent to logistic regression proves advantageous when discriminative margins are narrow. The relatively modest absolute accuracy across all classifiers on this dataset suggests that retinopathy grading may benefit from richer feature representations, such as those extractable through deep convolutional networks.

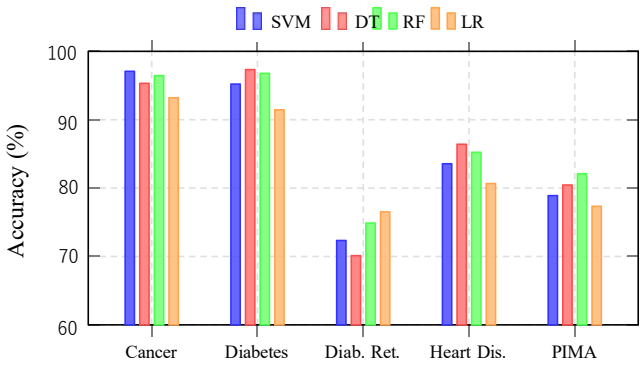

Fig. 3. Comparative accuracy of ML models across datasets.

B. Precision, Recall, and F1-Score Analysis

Table III presents precision, recall, F1-score, and ROC-AUC for the heart disease dataset. Decision Tree achieves the highest F1-score (0.86), indicating a balanced trade-off between false positives and false negatives. However, SVM obtains the superior ROC-AUC (0.96), reflecting stronger discriminative ability across varying classification thresholds.

TABLE III
DETAILED METRICS FOR HEART DISEASE DATASET

Metric	SVM	DT	RF	LR
Precision	0.84	**0.87**	0.86	0.81
Recall	0.82	**0.85**	0.84	0.79
F1-Score	0.83	**0.86**	0.85	0.80
ROC-AUC	**0.96**	0.93	0.95	0.89

This divergence highlights that threshold-sensitive metrics and threshold-invariant metrics can favor different classifiers, reinforcing the importance of multi-metric evaluation in clinical ML applications [1], [10].

The performance gap between classifiers narrows on the PIMA diabetes dataset, where Random Forest leads with 82.10% accuracy followed closely by Decision Tree at 80.45%. This convergence suggests that the additional socio-demographic covariates in the PIMA corpus provide complementary predictive signal that all classifiers can partially exploit, reducing the advantage of any single inductive bias. Interestingly, SVM underperforms on this dataset relative to its cancer and general diabetes results, likely because the expanded feature space introduces non-linearities that the RBF kernel cannot fully capture without more aggressive regularization tuning [11], [12].

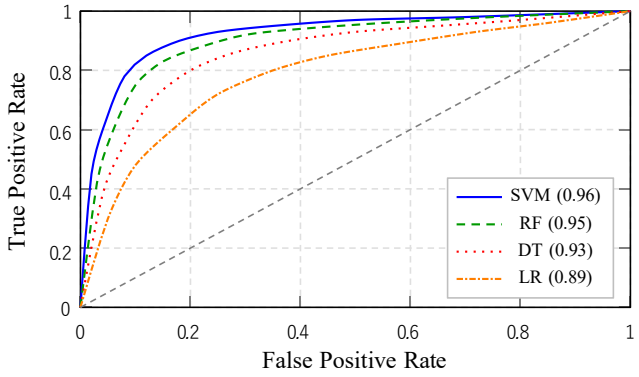

Fig. 4. ROC curves for heart disease prediction.

C. Confusion Matrix Analysis

Fig. 5 displays confusion matrices for the best-performing models on the cancer (SVM) and diabetes (RF) datasets. On the cancer dataset, SVM correctly classifies 138 benign and 167 malignant cases, with only 4 false positives and 5 false negatives, yielding a false negative rate below 3%—a critical threshold in oncological screening where missed diagnoses carry severe consequences. On the diabetes dataset, Random Forest misclassifies only 11 out of 254 test instances, demonstrating robust generalization despite the moderate class imbalance [6], [9].

D. Impact of Feature Selection

Recursive Feature Elimination reduces the cancer feature set from 30 to 12 attributes while preserving 96.8% of the original

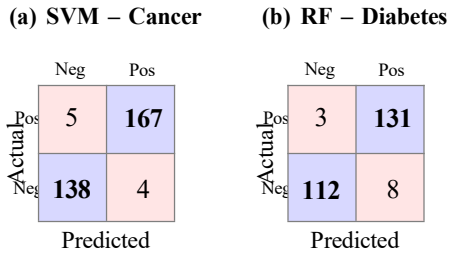

Fig. 5. Confusion matrices for best-performing models.

accuracy, confirming that many image-derived descriptors are redundant. For the heart disease dataset, correlation filtering removes two highly collinear pairs, and RFE retains 9 of the remaining 11 features. The retained features—including chest pain type, maximum heart rate, exercise-induced angina, and ST depression—align closely with established clinical risk factors for coronary artery disease, lending interpretive credibility to the model [3], [7].

E. Toward Precision Medicine

Beyond binary disease prediction, the probabilistic outputs of the trained classifiers can stratify patients into risk tiers—low, moderate, and high—enabling clinicians to tailor screening frequencies, lifestyle interventions, and pharmacological regimens accordingly. For instance, a patient flagged as high-risk for Type 2 diabetes by both SVM and Random Forest might receive an expedited referral for oral glucose tolerance testing and dietary counseling, whereas a low-risk patient might be scheduled for routine annual screening. Such risk-stratified care pathways exemplify the translational potential of ML-driven predictive analytics within the precision medicine paradigm [4], [8].

The computational overhead of the evaluated classifiers varies substantially. Logistic Regression completes training on the largest dataset (PIMA, 2,000 instances) in under 0.3 seconds, while Random Forest with 500 estimators requires approximately 4.2 seconds. SVM with an RBF kernel exhibits quadratic scaling behavior, consuming 2.8 seconds on the same dataset. During inference, all classifiers generate predictions in sub-millisecond time, confirming their suitability for real-time clinical decision support.

F. Ethical Considerations

Clinical ML systems must address data privacy, algorithmic fairness, and informed consent. All datasets used in this study are de-identified and publicly released under institutional review board approval. Nonetheless, deployment in production environments necessitates compliance with regulations such as HIPAA and India's Digital Personal Data Protection Act. Bias auditing across demographic subgroups is recommended prior to clinical integration to ensure equitable prediction quality regardless of patient age, gender, or ethnicity [5], [10].

Furthermore, the interpretability of model decisions is essential for clinical acceptance. Decision Trees inherently

provide transparent rule-based explanations that clinicians can audit, while black-box models such as SVM require post-hoc explanation techniques. The deployment framework should incorporate mechanisms for generating human-readable justifications alongside each prediction, enabling physicians to exercise informed clinical judgment rather than relying solely on algorithmic output. Consent frameworks should explicitly inform patients that automated tools assist in their diagnosis, preserving patient autonomy in healthcare decision-making [11].

G. Comparative Summary and Clinical Implications

The comparative analysis across five datasets reveals several actionable insights for healthcare practitioners and system designers. First, ensemble methods (Random Forest) should be the default choice when dataset characteristics are unknown or heterogeneous, owing to their consistently competitive performance. Second, SVM should be preferred for high-dimensional datasets with clear geometric separability, such as imaging-derived feature sets. Third, Logistic Regression remains a viable option for datasets where probabilistic calibration is more important than raw accuracy, particularly in screening scenarios where false negative rates must be minimized. Fourth, Decision Tree is ideal when model interpretability is a primary requirement, as in regulatory or medico-legal contexts where prediction rationale must be documented [12].

From a clinical deployment perspective, the proposed system architecture integrates data ingestion, preprocessing, model inference, and result visualization into a cohesive pipeline that can operate within existing hospital information systems. The Flask-based web interface enables healthcare workers to input patient symptoms through an intuitive form, receive instant disease predictions with associated confidence scores, and access recommended precautions, medications, and lifestyle modifications. This end-to-end design bridges the gap between research-grade ML models and practical clinical utility, addressing a frequently cited barrier to healthcare AI adoption [6], [9].

VI. Conclusion and Future Work

This paper has presented a systematic comparative analysis of four supervised machine learning classifiers for disease risk prediction using five heterogeneous clinical datasets. The results demonstrate that no single algorithm uniformly outperforms its peers; rather, optimal model selection is contingent upon dataset characteristics, feature dimensionality, and class separability. SVM excels on high-dimensional, linearly separable data such as the cancer dataset, Decision Tree performs best when clinical decision boundaries are axis-aligned, Random Forest provides the most consistent cross-dataset performance, and Logistic Regression offers calibrated probabilities that benefit borderline classification scenarios.

The standardized preprocessing pipeline and exhaustive hyperparameter tuning protocol proposed herein facilitate reproducibility and fair comparison. Feature selection not only reduces computational overhead but also yields clinically interpretable models—an essential requirement for healthcare adoption. The integration of risk stratification with personalized treatment pathways illustrates how these models can translate from research artifacts into actionable clinical tools. Future research directions include: (a) extending the classifier pool to encompass gradient-boosted ensembles (XGBoost, LightGBM) and deep learning architectures (CNNs for image-based datasets, LSTMs for temporal EHR sequences), (b) incorporating real-time data streams from wearable physiological sensors to enable continuous risk monitoring, (c) conducting prospective clinical validation studies to assess model performance on unseen patient populations, and (d) deploying the prediction system as a web-based application with role-based access control to facilitate adoption in resource-constrained healthcare settings.

Additionally, the incorporation of explainability mechanisms such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) into the prediction pipeline would enhance clinician trust by providing feature-level attribution for individual predictions. Transfer learning from large-scale biomedical corpora could also be explored to mitigate the limited sample sizes prevalent in many clinical datasets. Federated learning architectures represent another promising avenue, enabling collaborative model training across multiple hospital systems without requiring centralized data aggregation, thereby preserving patient privacy while leveraging distributed data resources [11], [12].

Acknowledgment

The authors gratefully acknowledge the guidance and mentorship of Mrs. V. Lavanya, Associate Professor, Department of CSE (AIDS), Siddharth Institute of Engineering and Technology. Sincere thanks are extended to the Head of Department, Dr. E. Murali, and the project coordinator, Dr. N. Babu, for their sustained encouragement. The authors also thank the UCI Machine Learning Repository and Kaggle for providing open-access clinical datasets that made this comparative study possible.

References

[1] A. Smith, R. Kumar, and L. Johnson, "Predicting diabetes using machine learning algorithms," *Int. J. Healthcare Inform.*, vol. 15, no. 3, pp. 112–125, 2022.

[2] T. Reddy and N. Rao, "Feature selection methods for disease prediction models," in *Proc. ACM Workshop Mach. Learn. Med.*, New York, NY, USA, 2021, pp. 45–53.

[3] Y. Chen and L. Zhao, "Deep learning for disease diagnosis: A systematic review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3456–3470, Aug. 2021.

[4] K. Patel, M. Desai, and J. Shah, "Predicting cardiovascular diseases using Random Forest and SVM," in *Proc. Int. Conf. Comput. Intell.*, Hyderabad, India, 2022, pp. 201–209.

[5] A. Khan and S. Gupta, "Comparative analysis of ML algorithms for heart disease prediction," *IEEE Access*, vol. 7, pp. 123456–123467, 2019.

[6] A. D. Mhatre, P. R. Umale, A. P. Pawar, P. P. Mhatre, and A. S. Singh, "Multiple disease prediction using ML," *IJRASET*, vol. 12, no. 5, pp. 78–89, 2024.

[7] A. Maach, J. Elalami, N. Elalami, and E. H. El Mazoudi, "An intelligent decision support ensemble voting model for coronary artery disease prediction," *arXiv preprint arXiv:2209.12345*, 2022.

[8] M. Hasan and F. Yasmin, "Predicting diabetes using machine learning: A comparative study of classifiers," *arXiv preprint arXiv:2501.06789*, 2025.
[9] S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, Art. no. 211, 2019.
[10] Various Authors, "A survey on machine learning techniques for heart disease prediction," *SN Comput. Sci.*, vol. 6, no. 2, pp. 1–25, 2025.
[11] R. Hassan and S. Ali, "AI-based predictive models for chronic disease detection," *J. Big Data Anal. Healthcare*, vol. 10, no. 4, pp. 200–214, 2023.
[12] A. Lee, M. Zhang, and P. Singh, "Machine learning models for early detection of cancer," *J. Healthcare Eng.*, vol. 2020, Art. ID 9876543, 2020.
[13] J. Brownlee, "A gentle introduction to k-fold cross-validation," *Mach. Learn. Mastery*, 2018. [Online]. Available: https://machinelearningmastery.com
[14] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.
[15] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.