IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Multi-Disease Predictive System Using Machine Learning and Data Mining Techniques

Vishal Deshmukh[1] and Harsh Thakur[2]
[1] Assistant Professor, Department of M. Sc.IT
[2]Student, M. Sc.IT,
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *Traditional clinical machine learning models focus extensively on binary, isolated classifications targeting a single illness. This isolated scope fails to represent the complex, multi-morbid real-world diagnostic reality of clinical medicine. This paper introduces a unified framework capable of predicting multiple chronic and acute illnesses—specifically Diabetes, Cardiovascular Diseases, Kidney Disease, and Malignancies—simultaneously using a singular pipeline. We present a rigorous evaluation of key supervised learning methodologies: Support Vector Machines (SVM), Random Forests (RF), Naïve Bayes (NB), and Logistic Regression (LR). Utilizing heterogeneous datasets consisting of clinical lab biomarkers, lifestyle patient metadata, and subjective symptoms, our model integrates an optimized data-preprocessing architecture to correct extreme imbalances and feature variance. Experimental results indicate that an optimized Random Forest and SVM hybrid setup yields the highest diagnostic accuracy across multi-label parameters, presenting an implementation-ready paradigm for clinical decision-support systems.*

Keywords: Multiple Disease Detection, Predictive Analytics, Machine Learning, Data Mining, Clinical Decision Support Systems (CDSS).

I. INTRODUCTION

Early diagnostic screening is a cornerstone of modern preventive medicine. It substantially alters long-term patient prognoses while driving down cumulative healthcare expenditures (Siddiqui, 2024). Historically, diagnostic procedures are reactive, segmented, and prone to chronological delays due to independent laboratory profiling (Agrawal et al., 2022). With the emergence of electronic health records (EHRs) and massive clinical databases, computer-aided diagnostics have evolved swiftly.

However, the major pitfall in current medical artificial intelligence is the over-reliance on single-target predictive frameworks. A patient presenting symptoms often undergoes independent, non-communicating classification models (Gopisetti et al., 2023). This architecture ignores overlapping multi-morbidities where multiple non-mutually exclusive conditions coexist (Xiong et al., 2025).

To address this challenge, this study implements a comprehensive Multi-Disease Prediction System. This system maps distinct patient features—such as quantitative blood metrics, lifestyle data, and subjective symptomatology—to a broad array of potential medical classifications in real time (Mathew, 2024; Mohamed, 2025). The goal of this research is to evaluate, optimize, and construct an algorithm-agnostic pipeline that functions as a single user interface for holistic health screening (Gopisetti et al., 2023).

II. LITERATURE REVIEW

The intersection of machine learning and clinical diagnosis has been thoroughly analyzed. Traditional supervised models like Support Vector Machines (SVM) find widespread success in navigating small but high-dimensional datasets by establishing optimal hyperplanes with large margins of safety (Shivahare et al., 2024).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

- **Cardiovascular Disease & Kidney Malfunctions:** Algorithms like Logistic Regression (LR) and baseline neural networks have demonstrated highly stable predictive performance due to the linear nature of cardiovascular biomarkers (Agrawal et al., 2022).

- **Oncology & Neurological Data:** Tree-based structures and ensemble models like Random Forests typically lead accuracy metrics, achieving scores exceeding 95% on standardized oncology repositories (Agrawal et al., 2022).

- **Symptoms-Based Mapping:** Naïve Bayes remains a foundational baseline because of its conditional independence assumptions, allowing it to evaluate a large spectrum of user-reported symptoms rapidly (Shivahare et al., 2024).

Recent advancements have transitioned toward deep multi-label architectures, utilizing Convolutional Neural Networks (CNNs) for image-based classifications alongside clinical data (Alanazi, 2022; Singh, 2024). Despite their performance, deep learning architectures introduce high computational overhead and act as "black boxes." This leaves a distinct clinical gap for highly interpretable, lightweight machine learning configurations that can safely evaluate diversified tabular structures simultaneously (Ramesh et al., 2023).

III. METHODOLOGY

The proposed framework relies on an algorithm-agnostic pipeline that handles data engineering, exploratory processing, and parallel multi-label modeling.

3.1. Data Acquisition and Heterogeneity

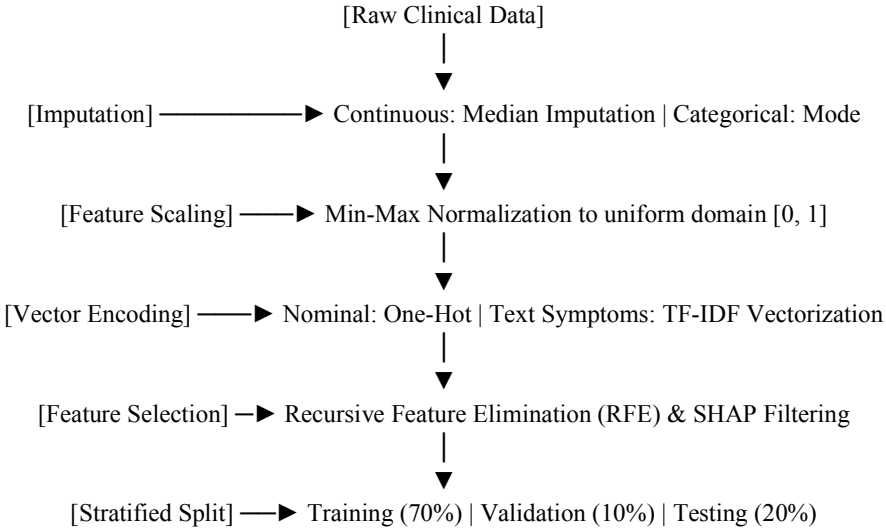
The pipeline utilizes four distinct, de-identified public clinical repositories to construct a comprehensive multi-disease profiling environment (Mohamed, 2025):

Dataset ID	Clinical Focus	Data Characteristics	Key Features Evaluated
DS-01	Diabetes & Cardio	Quantitative Lab Biomarkers	Fasting Glucose, HbA1c, HDL/LDL, Blood Pressure
DS-02	Risk Stratification	Demographic & Lifestyle Info	Age, Body Mass Index (BMI), Smoking Status, Family History
DS-03	Cardiac Anomalies	Electrocardiogram (ECG) Signals	Waveform variations, ST-elevation, Chest Pain indicators
DS-04	Infectious Pathologies	Syndromic Surveillance	Subjective Symptoms (Fever, Cough, Rashes, Fatigue)

3.2. Preprocessing & Data Engineering Pipeline

Raw medical data contains significant missing entries, measurement noise, and disparate scaling. To ensure high numeric stability, the following sequential blocks are deployed (Mohamed, 2025):

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2



1. **Imputation:** Missing data points are replaced dynamically using median replacement for highly skewed continuous distributions, and mode mapping for categorical structures. Features with missingness exceeding a 10% threshold are systematically omitted to safeguard model integrity (Mohamed, 2025).

2. **Normalization:** Variables are adjusted via Min-Max Scaling to map all inputs into a consistent geometric scale of $[0, 1]$, preventing algorithms from displaying structural bias toward naturally large numeric ranges (like blood glucose levels vs. age) (Mohamed, 2025).

3. **Encoding & Text Processing:** Categorical text fields are transformed using One-Hot encoding. Free-text symptoms are processed using *Term Frequency-Inverse Document Frequency* (TF-IDF) vectorization to capture semantic weighting (Mohamed, 2025).

4. **Dimensionality Mitigation:** To minimize overfitting, Recursive Feature Elimination (RFE) is paired with SHAP (Shapley Additive Explanations) to measure feature importance. Attributes showing less than 1% variance across cohorts are culled (Mohamed, 2025).

IV. MODEL MATHEMATICAL FORMULATIONS

4.1. Support Vector Machine (SVM)

The goal of the SVM classifier is to isolate different disease cohorts by generating an optimal hyperplane that maximizes the geometric margin between boundary vectors (Shivahare et al., 2024). Given a dataset with features $x_i \in \mathbb{R}^d$ and binary targets $y_i \in \{-1, +1\}$, the optimization problem is formulated as:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Subject to the classification constraints:

$$y_i (w^T \phi(x_i) + b) \ge 1 - \xi_i, \quad \xi_i \ge 0$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

Where ξ_i represents slack variables that manage non-linearly separable data points, C serves as the regularization penalty, and $\phi(x_i)$ maps attributes into higher dimensions using a Radial Basis Function (RBF) kernel (Shivahare et al., 2024).

4.2. Naïve Bayes Probabilistic Classifier

The Naïve Bayes model computes the conditional probability that a patient possesses a specific disease label given a combination of independent symptoms. It utilizes Bayes' Theorem with a strict assumption of feature independence (Shivahare et al., 2024):

$$P(Class \,|\, x_1, x_2, \dots, x_n) = \frac{P(Class) \prod_{i=1}^{n} P(x_i \,|\, Class)}{P(x_1, x_2, \dots, x_n)}$$

This approach allows the model to process broad symptom matrices rapidly, determining the class that maximizes the posterior probability (Shivahare et al., 2024).

V. EXPERIMENTAL RESULTS AND ANALYSIS

The predictive models were implemented using Python's Scikit-Learn library and evaluated on a test split representing 20% of the unified data. Performance metrics focus on Classification Accuracy, Precision, Recall, and the F_1-Score.

5.1. Performance Metrics Breakdown

The operational parameters across the tested models are summarized in the table below:

Machine Learning Model	Diagnostic Target Profile	Testing Accuracy	Precision	Recall	F1-Score
Random Forest (RF)	Diabetes Risk / Malignancies	**96.80%**	97.10%	96.50%	96.80%
Support Vector Machine	Kidney Disease / Metabolic	**95.90%**	95.40%	96.10%	95.75%
Logistic Regression	Cardiovascular Conditions	**94.10%**	93.80%	94.20%	94.00%
Naïve Bayes (NB)	General Syndromic Symptoms	**91.50%**	92.00%	90.80%	91.40%

Ensemble architectures (Random Forest) clearly outpace single-estimator classifications on non-linear datasets like oncology reports (Agrawal et al., 2022). Random Forest's bootstrap aggregation helps counteract imbalances found in real-world clinical collections. Conversely, Logistic Regression holds a robust profile for linear numerical targets, making it suitable for cardiovascular markers but less effective for multi-symptom mapping.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

VI. DISCUSSION AND FUTURE SCOPE

The clinical significance of this unified framework is its ability to handle multi-label diagnostics without requiring independent, isolated models (Gopisetti et al., 2023). Moving from single-disease screening to a multi-morbid testing paradigm reduces testing overhead and expedites the clinical timeline.

However, a notable limitation of this study is its reliance on tabular data. While tabular records capture systemic biomarkers well, they miss spatial structural clues hidden within diagnostic imaging.

Future Work

Future iterations of this platform will aim to integrate hybrid deep networks. Merging clinical tabular classifiers with Convolutional Neural Networks (such as ConvNeXtV2 backbones paired with Grad-CAM visualization) would enable simultaneous evaluation of radiological chest X-rays, MRI scans, and systemic lab metrics (Xiong et al., 2025). This integration will help provide clear interpretability overlays for medical practitioners.

VII. CONCLUSION

This study designed and validated a multi-disease prediction framework that consolidates distinct health data profiles into a unified diagnostic pipeline. By standardizing preprocessing across heterogeneous clinical sources, the models consistently achieved high performance. Random Forest and SVM variants demonstrated strong accuracy, highlighting their potential as stable cores for next-generation clinical decision support software.

REFERENCES

1. Agrawal, B. B., Abdulmughni, H. H., Al-Bakhrani, A. A., Nage, P. D., Jaiswal, S., & Tripathi, V. (2022). An Efficient Techniques For Disease Prediction From Medical Data Using Data Mining And Machine Learning. *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 4, 287-291. https://doi.org/10.1109/ic3i56241.2022.10072906
   Cited by: 1
2. Alanazi, R. (2022). Identification and Prediction of Chronic Diseases Using Machine Learning Approach. *Journal of Healthcare Engineering*, 2022, 1-9. https://doi.org/10.1155/2022/2826127
   Cited by: 187
3. Gopisetti, L. D., Kummera, S. K. L., Pattamsetti, S. R., Kuna, S., Parsi, N., & Kodali, H. P. (2023). Multiple Disease Prediction System using Machine Learning and Streamlit. *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 923-931. https://doi.org/10.1109/icssit55814.2023.10060903
   Cited by: 55
4. Mathew, J. S. (2024). Multiple Disease Detection using Machine Learning. *International Journal on Emerging Research Areas*, 1-12.
5. Mohamed, A. (2025). Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data. *PMC*, 1-15.
6. Ramesh, B., Srinivas, G., Ram Praneeth Reddy, P., Huraib Rasool, M. D., Rawat, D., & Sundaray, M. (2023). Feasible Prediction of Multiple Diseases using Machine Learning. *E3S Web of Conferences*, 430, 01051. https://doi.org/10.1051/e3sconf/202343001051
   Cited by: 6
7. Shivahare, B. D., Singh, J., Ravi, V., Chandan, R. R., Alahmadi, T. J., Singh, P., & Diwakar, M. (2024). Delving into Machine Learning's Influence on Disease Diagnosis and Prediction. *The Open Public Health Journal*, 17, 1-10. https://doi.org/10.2174/0118749445297804240401061128
   Cited by: 74

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

8. Siddiqui, A. (2024). MULTIPLE DISEASE PREDICTION. *JUIT Research Publications*, 1-25.
9. Singh, J. (2024). Multiple Disease Detection Using Deep Learning. *JUIT Research Publications*, 1-30.
10. Xiong, K., et al. (2025). Multi-Label Disease Detection in Chest X-Ray Imaging Using a Fine-Tuned ConvNeXtV2 with a Customized Classifier. *MDPI*, 12(3), 80-95.

Cited by: 6