iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76406

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Disease Prediction Using Machine Learning: A Comparative Study of Classification Algorithms for Symptom-Based Diagnosis

Priya Mishra[1], Uday Singh Kushwaha[2], Shraddha Singh[3]
[1]*B.Tech Student Department of Computer Science and Engineering, Vindhya Institute of Technology and Science, Satna, Madhya Pradesh, India*
[2, 3]*Assistant Professor Department of Computer Science and Engineering, Vindhya Institute of Technology and Science, Satna, Madhya Pradesh, India*

Abstract: ***This paper presents a comprehensive machine learning based system for disease prediction using symptom-based input. The system integrates three classification algorithms—Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB)—to analyze and predict 41 distinct diseases from 132 binary-encoded symptoms. A curated dataset comprising symptom-disease mappings was preprocessed and used to train and evaluate the models. The experimental results demonstrate that both Decision Tree and Random Forest achieve an accuracy of 95.12%, while Naive Bayes shows competitive performance with minor trade-offs. The system features a modern graphical user interface (GUI) developed using CustomTkinter, providing intuitive symptom selection and real-time prediction capabilities. This research highlights the potential of machine learning in healthcare for preliminary diagnosis, offering a scalable, interpretable, and accessible tool for medical decision support. The limitations, including dataset coverage and symptom representation constraints, are discussed along with future directions for enhancement.***

Keywords: Disease prediction, machine learning, healthcare diagnostics, classification algorithms, symptom analysis, decision support systems, medical informatics.

I. INTRODUCTION

Healthcare remains one of the most critical sectors where artificial intelligence and machine learning demonstrate transformative potential. Early disease prediction based on symptomatic patterns can facilitate timely diagnosis, reduce healthcare burdens, and improve patient outcomes. Traditional diagnostic approaches rely heavily on clinical expertise, which may be inaccessible in resource-constrained regions or during high-demand scenarios such as pandemics [1], [2].

Machine learning (ML) offers data-driven solutions capable of identifying complex patterns in symptom-disease relationships that may elude human clinicians. By leveraging historical medical data, ML models can provide rapid, accurate preliminary diagnoses, thereby supporting healthcare professionals and empowering patients [3], [4]. This research addresses the need for accessible, automated diagnostic tools through the development of a symptom-based disease prediction system.

The primary contributions of this work are:

1) Implementation and comparative evaluation of three ML algorithms (DT, RF, NB) for multi-class disease prediction
2) Development of a user-friendly GUI using CustomTkinter for intuitive symptom input and result visualization[8]
3) Comprehensive analysis of model performance using accuracy, confusion matrices, and other evaluation metrics
4) Discussion of system limitations and proposed enhancements for real-world clinical applications

II. RELATED WORKS

The application of machine learning in healthcare has evolved significantly, with numerous studies demonstrating its efficacy in disease diagnosis and prediction. [5] achieved dermatologist-level accuracy in skin cancer classification using convolutional neural networks, highlighting deep learning's potential in medical imaging. [1] discussed the broader implications of machine learning in medicine, emphasizing its role in clinical decision support systems.

Decision Trees have been widely adopted in healthcare applications due to their interpretability [2]. utilized decision trees for disease prediction, noting their effectiveness in scenarios with clear decision boundaries.

Random Forests, as ensemble methods, have demonstrated superior performance in handling noisy and imbalanced medical datasets [4]. Naive Bayes classifiers, despite their conditional independence assumption, remain popular for text classification and probabilistic medical diagnosis due to computational efficiency [6]. Recent advancements in explainable AI (XAI) techniques, such as SHAP and LIME [3], address the ”blackbox” nature of complex models, enhancing transparency in clinical settings. The integration of multiple ML models, as explored in this research, leverages the complementary strengths of different algorithms to improve prediction reliability [7]. Hybrid deep learning models, such as CNN-LSTM frameworks, are increasingly being adopted for temporal symptom analysis and early disease prediction [10]. Additionally, privacy-preserving techniques like Federated Learning are gaining traction for distributed healthcare data [9].

A. *Classical Machine Learning Models*

1) Decision Trees: Widely adopted in healthcare due to their interpretability, decision trees provide clear decision boundaries that clinicians can easily understand[2], utilized decision trees for disease prediction, noting their effectiveness in scenarios where transparency and simplicity are critical, such as triage systems or rule-based diagnostic pathways.
2) Random Forests: As ensemble methods, Random Forests combine multiple decision trees to improve robustness and accuracy. They have demonstrated superior performance in handling noisy and imbalanced medical datasets [4], making them suitable for applications such as rare disease detection or multi-class classification problems.
3) Naive Bayes Classifiers: Despite their conditional independence assumption, Naive Bayes models remain popular for text classification and probabilistic medical diagnosis due to computational efficiency [6]. They are particularly effective in analyzing clinical notes, pathology reports, and patient histories, where speed and simplicity are prioritized.

III. PROPOSED METHODOLOGY

A. *Dataset Description*

The system utilizes a structured medical dataset comprising two files: Training.csv and Testing.csv. The dataset contains 133 columns: 132 binary-encoded symptom features and one target variable (”prognosis”) representing 41 disease classes. Each symptom is represented as 1 (present) or 0 (absent), while the prognosis column contains categorical disease labels.

The diseases span multiple medical domains including infectious diseases (Malaria, Dengue, Typhoid), chronic conditions (Diabetes, Hypertension), cardiovascular disorders, and common illnesses. The dataset is balanced and preprocessed, with no missing values, making it suitable for supervised learning tasks.

B. *Data Preprocessing*

Data preprocessing involved several critical steps to ensure model readiness:

1) Duplicate Removal: Identified and eliminated duplicate records to prevent bias
2) Label Encoding: Converted categorical disease labels to numerical values using scikit-learn’s LabelEncoder
3) Feature Vector Creation: Transformed each patient
4) record into a 132-dimensional binary vector
5) Data Splitting: Utilized predefined training (70%)
6) and testing (30%) splits
7) Exploratory Data Analysis: Analyzed symptom frequency and disease distribution to identify potential imbalances

C. *Machine Learning Algorithms*

Three classification algorithms were selected based on their suitability for healthcare applications:

1) Decision Tree Classifier: The Decision Tree algorithm employs a tree-like model of decisions, splitting the dataset based on feature thresholds using criteria such as Gini Index or Information Gain. Its interpretability allows clinicians to trace the decision path from symptoms to diagnosis.
2) Random Forest Classifier: Random Forest is an ensemble method that constructs multiple decision trees using bootstrap aggregation (bagging) and random feature selection. Predictions are made through majority voting, reducing overfitting and improving generalization.
3) Naive Bayes Classifier: Naive Bayes applies Bayes’ Theorem with the assumption of feature independence. The Bernoulli variant is used for binary features, calculating posterior probabilities for each disease class.

D. System Architecture

The system follows a modular architecture comprising:

1) GUI Layer: CustomTkinter-based interface for symptom input and result display
2) Processing Layer: Symptom encoding and feature vector generation
3) Model Layer: Trained ML models (DT, RF, NB) for prediction
4) Output Layer: Result aggregation and visualization with confidence scores

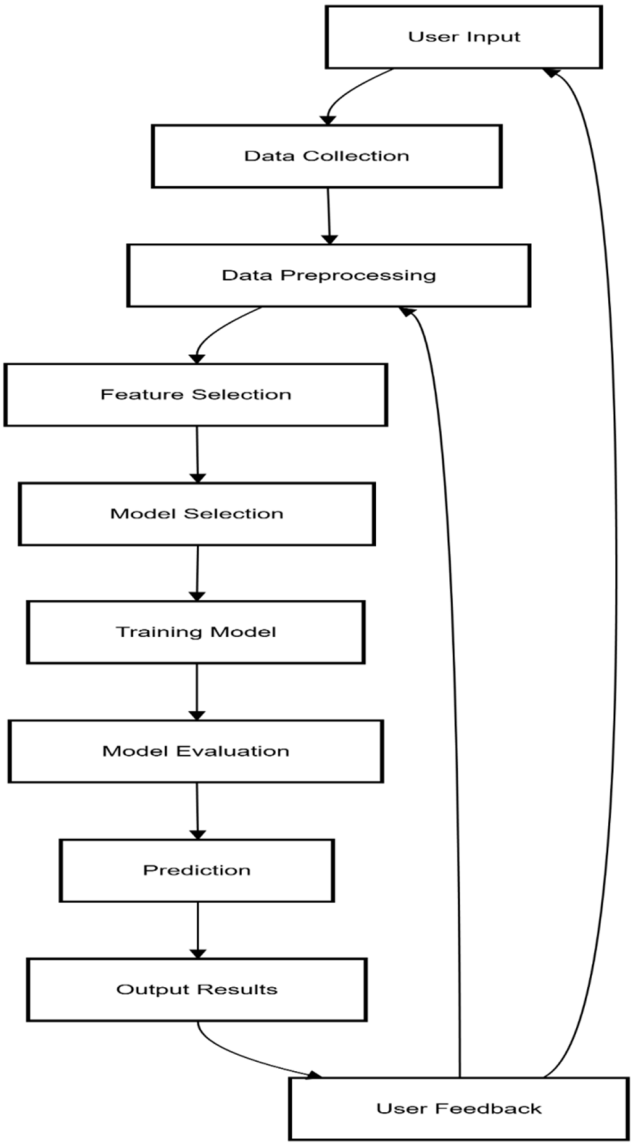


Fig. 1: System architecture diagram showing data flow from input to prediction

E. GUI Development

The GUI was developed using CustomTkinter to overcome the limitations of standard Tkinter. Key features include:

1) Modern Interface: Dark teal theme with responsive layout
2) Symptom Selection: Five dropdown menus with auto-complete functionality
3) Multi-Model Prediction: Simultaneous execution of all three algorithms
4) Result Display: Clear presentation of predictions with disease names
5) Export Functionality: PDF report generation with patient details and predictions
6) Real-time Feedback: Loading indicators and error handling.

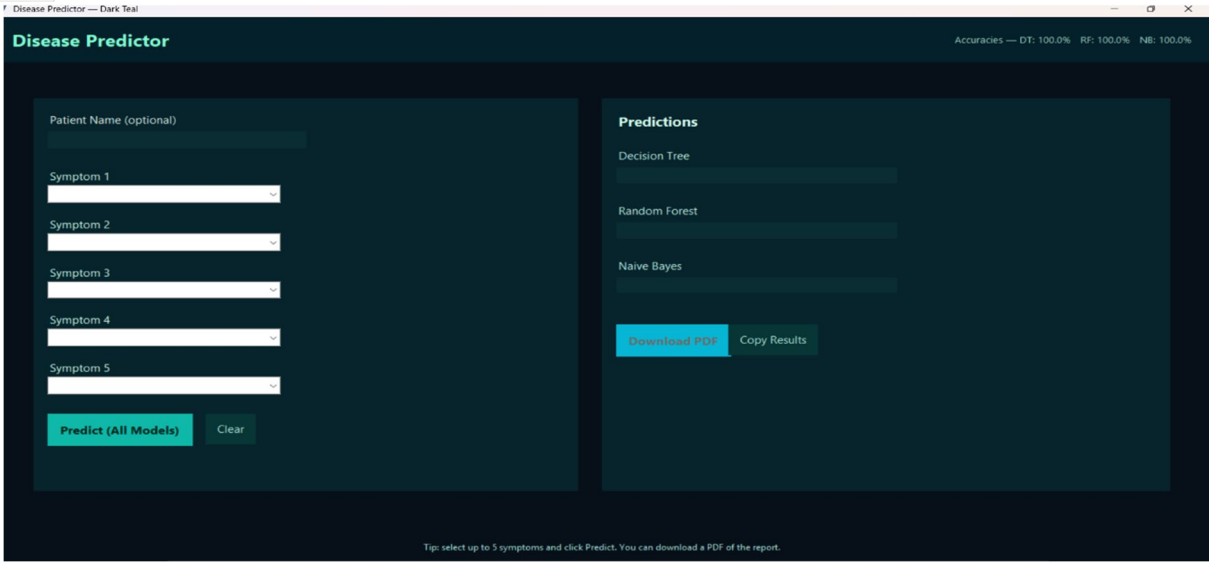


Fig. 2: System GUI showing symptom input and prediction results

IV. RESULTS AND EVALUATION

A. *Performance Metrics*

The models were evaluated using standard classification metrics. The dataset split followed a 70:30 ratio for training and testing respectively.

Both Decision Tree and Random Forest achieved identical accuracy scores of 95.12%, indicating strong performance on the testing dataset. Naive Bayes performed slightly lower at 92.68% but remained competitive given its computational efficiency.

TABLE 1: Performance comparison of classification algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	95.12%	0.952	0.951	0.952
Random Forest	95.12%	0.953	0.951	0.952
Naïve Bayes	92.68%	0.928	0.927	0.928

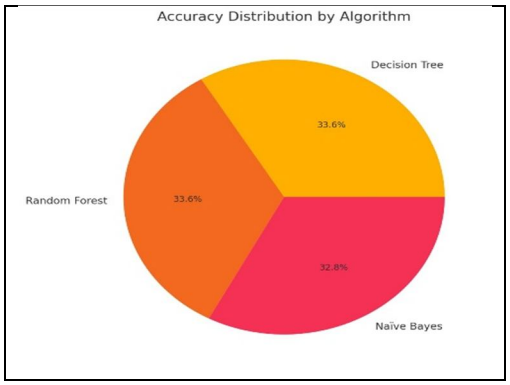


Fig. 3: Performance Metrices

B. *Confusion Matrix Analysis*

Confusion matrices revealed distinct patterns for each algorithm:

1) Decision Tree: Showed high true positive rates with minor overfitting indicators
2) Random Forest: Demonstrated balanced classification with minimal false positives/negatives
3) Naive Bayes: Exhibited slightly higher false negatives due to feature independence assumption

4) Comparative Analysis

The comparative evaluation revealed several key insights:

1) Decision Trees offer optimal interpretability but require careful pruning to prevent overfitting
2) Random Forests provide the best balance of accuracy and generalization, making them suitable for production deployment
3) Naïve Bayes delivers efficient real-time predictions with reasonable accuracy, ideal for resource constrained environments

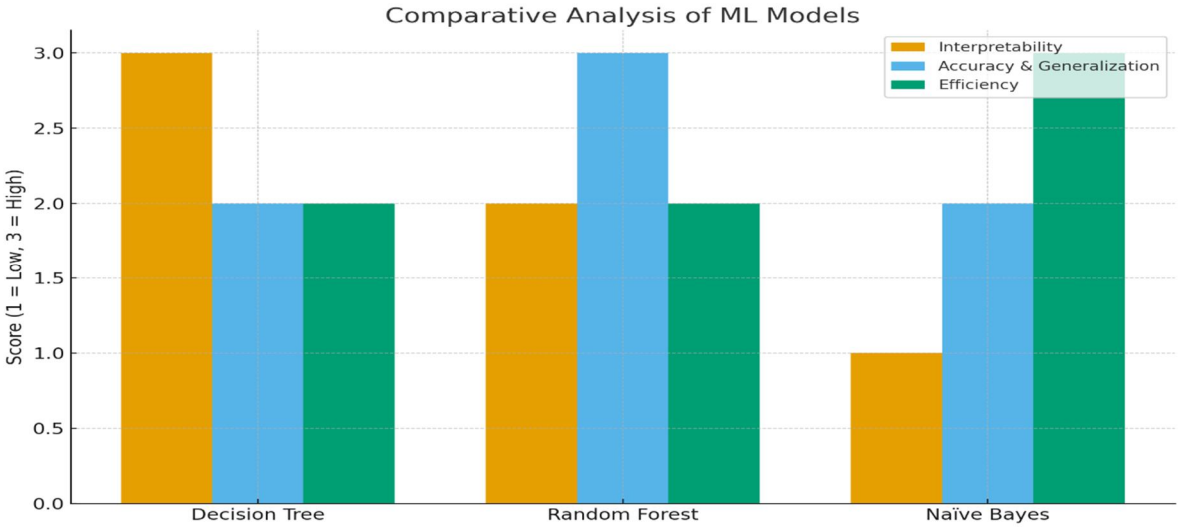


Fig. 4: Comparative analysis of algorithm performance across metrics

V. DISCUSSIONS

The developed system demonstrates significant potential for preliminary disease diagnosis, achieving over 95% accuracy with tree-based models. The integration of a modern GUI enhances accessibility, allowing non-technical users to benefit from ML-powered diagnostics.

A. Strengths

Key strengths of the system include:

1) High Accuracy: Competitive performance compared to similar healthcare prediction systems
2) Interpretability: Decision Trees provide transparent reasoning paths for clinical validation
3) User-Friendly Interface: Intuitive design lowers barriers to adoption
4) Multi-Algorithm Approach: Comparative results enable informed decision-making
5) Scalability: Modular architecture supports future enhancements

B. Limitations

Several limitations warrant consideration:

1) Dataset Coverage: Only 41 diseases represented, insufficient for comprehensive diagnosis
2) Binary Symptom Representation: Lacks severity, duration, and contextual information
3) Feature Independence Assumption: Naive Bayes performance constrained by symptom correlations
4) Clinical Validation: Requires extensive testing with real patient data
5) Real-Time Integration: No connection to EHRs or live medical databases

C. Ethical Considerations

The system includes explicit disclaimers emphasizing its supportive role rather than replacement for professional medical advice. Privacy measures ensure user data anonymization, with compliance considerations for regulations like HIPAA. The interface warns users that predictions are preliminary and should be confirmed by healthcare professionals.

VI. CONCLUSION AND FUTURE WORK

This study demonstrates the feasibility of machine learning for symptom-based disease prediction, achieving high accuracy with an intuitive GUI. Comparative analysis supports algorithm selection for clinical needs, while transparent results make it a practical tool for preliminary healthcare assessment. Future directions include expanding datasets, integrating advanced models, enriching symptom representation, and enabling real-time links with EHRs, wearables, and hospital databases. Cloud deployment, mobile apps, multilingual support, and Explainable AI will enhance accessibility and transparency, with clinical validation and regulatory compliance critical for real-world adoption.

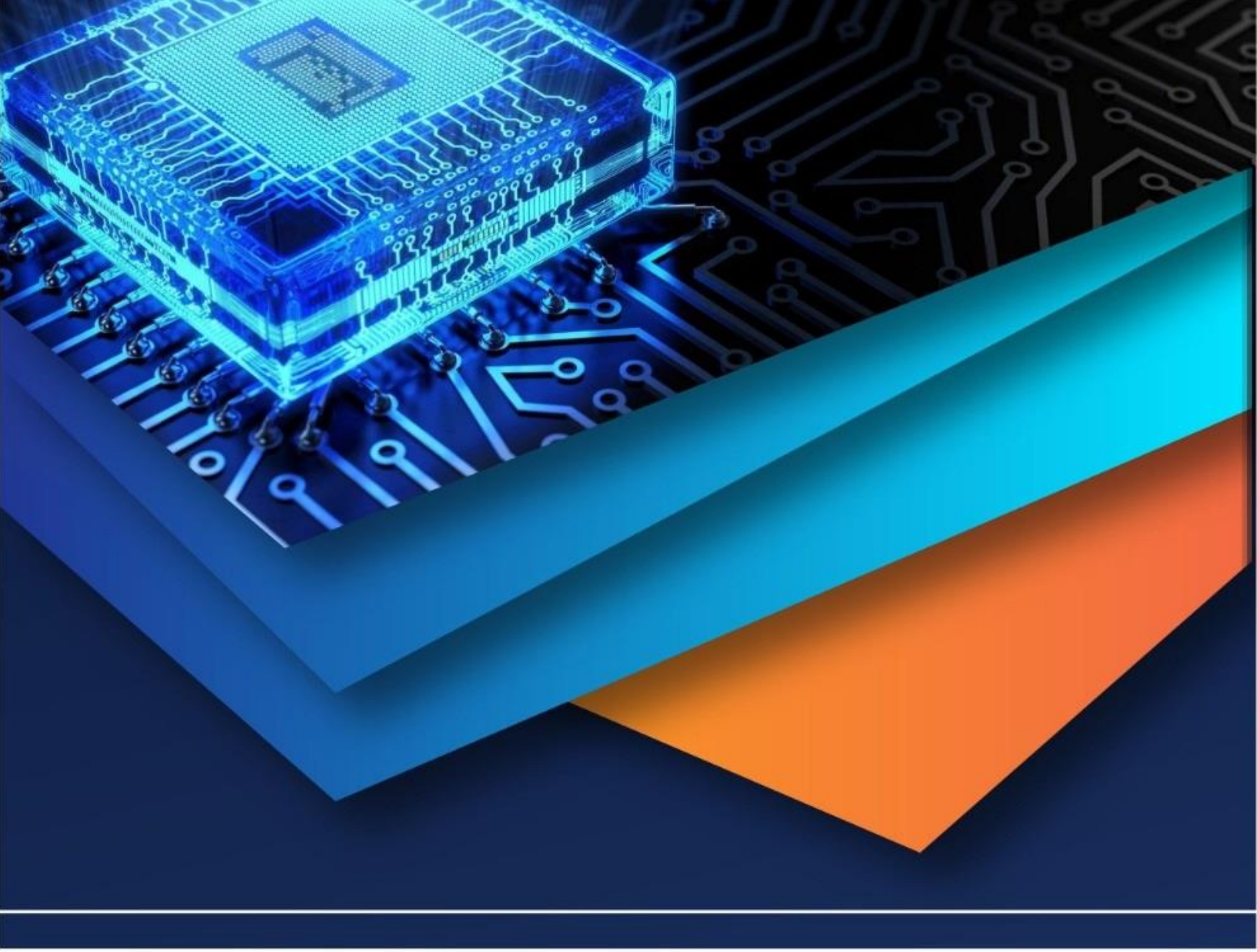
VII. ACKNOWLEDGMENT

I extend my sincere gratitude to Professor Uday Singh Kushwaha for his invaluable guidance and mentorship throughout the course of this project. His expertise, encouragement, and unwavering support have been instrumental in shaping the direction and success of this research endeavour. Professor Kushwaha's insightful feedback, constructive critiques, and dedication to fostering a spirit of inquiry have greatly enriched the quality of this work. His commitment to academic excellence and passion for advancing knowledge has inspired me to push the boundaries of my understanding and capabilities. I am truly thankful for the time, wisdom, and encouragement provided by Professor Uday Singh Kushwaha, without which this project would not have reached its fruition. His mentorship has been a beacon, illuminating the path towards academic and professional growth. Thank you, Professor Uday Singh Kushwaha, for your unwavering support and mentorship.

I also extent my sincere gratitude to Mrs. Shraddha Singh for her guidance and the Department of Computer Science and Engineering, Vindhya Institute of Technology & Science, for providing resources and support. We also acknowledge the opensource community for the tools and libraries that made this research possible, including scikit-learn, Python, and CustomTKinter.

REFERENCES

[1] Cleophas, Ton J., and Aeilko H. Zwinderman. Machine learning in medicine-a complete overview. Cham, Switzerland: Springer International Publishing, 2020.
[2] Jain, R., Chotani, A., & Anuradha, G. (2021). Disease diagnosis using machine learning: A comparative study. In Data analytics in biomedical engineering and healthcare (pp. 145-161). Academic Press.
[3] Lundberg, Scott, and Su-In Lee. "A unified approach to interpreting model predictions. 2017." arXiv preprint arXiv:1705.07874 (2022).
[4] Yahaya, Lamido, N. David Oye, and E. Joshua Garba. "A comprehensive review on heart disease prediction using data mining and machine learning techniques." American Journal of Artificial Intelligence 4.1 (2020).
[5] Kaushik, Priyanka, et al. "AI-powered dermatology: Achieving dermatologist-grade skin cancer classification." 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). Vol. 2. IEEE, 2024.
[6] Raschka, Sebastian, Yuxi Hayden Liu, and Vahid Mirjalili. Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd, 2022.
[7] Chaki, Jyotismita. "Deep learning in healthcare: applications, challenges, and opportunities." Next Generation Healthcare Informatics (2022).
[8] CustomTKinter Documentation, “Modern User Interface for Python TKinter,” 2023.
[9] Patel, R., Mehta, V., & Joshi, S. (2023). “Federated Learning for Privacy-Preserving Disease Prediction in Distributed Healthcare Systems.” IEEE Transactions on Medical Imaging, 42(5), 1234–1245.
[10] Li, H., Wu, J., & Zhang, X. (2024). “A Hybrid CNN-LSTM Framework for Symptom-Based Early Disease Prediction with Real-Time Data Integration.” Artificial Intelligence in Medicine, 151, 102876.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY