iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.72979

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 VII July 2025- Available at www.ijraset.com

Early Disease Prediction using Machine Learning and Deep Learning Algorithms

Vyshnavi Posu[1], G. Narasimham[2]
[1]*Post Graduate Student, M. Tech (Data Science),* [2]*Associate Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad, UCESTH*

Abstract: ***Early detection of life-threatening diseases such as cancer and thyroid disorders plays a critical role in reducing mortality and improving treatment outcomes. With traditional diagnostic methods often being time-consuming, resource-intensive, and dependent on specialist expertise, the integration of intelligent systems has become essential. This study explores the application of various machine learning (ML) and deep learning (DL) algorithms to predict diseases at an early stage using structured clinical datasets. Algorithms like Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and Artificial Neural Networks (ANN) were implemented and evaluated based on accuracy, precision, recall, and F1-score. The results show that all models achieved above 85% accuracy, with ANN and Random Forest models performing exceptionally well. A Streamlit-based web application was developed for real-time prediction, enabling easy clinical integration. The research underscores the effectiveness of ML/DL in enhancing diagnostic efficiency and recommends further development involving real clinical data and advanced model interpretability.***

Keywords: Deep Learning ,Machine L earning, Artificial Neural Network(ANN),Random Forest, K-Nearest Neighbors(KNN),Streamlit

I. INTRODUCTION

The early identification of serious health conditions like cancer and thyroid disorders is crucial to improving patient survival, simplifying treatment processes, and delivering more effective care. These diseases often develop with little to no noticeable symptoms in the beginning, making early-stage diagnosis a complex and demanding task. Conventional diagnostic procedures, while accurate, often involve lengthy processes, require significant medical resources, and depend heavily on the expertise of trained professionals.

In recent years, machine learning (ML) has emerged as a transformative technology in the healthcare domain, enabling the development of smart diagnostic systems that can learn from existing patient data to provide reliable predictions. These data-driven systems are particularly effective in detecting diseases where early indicators are subtle or unclear. By utilizing well-structured datasets that include clinical and pathological features, ML algorithms can recognize hidden patterns associated with disease presence, facilitating quicker and more consistent diagnostic decisions.

The methodology adopted in this study includes comprehensive preprocessing of medical data, followed by the application and evaluation of several machine learning and deep learning models. Techniques such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes, and Artificial Neural Networks (ANN) were tested. These models were evaluated using standard performance metrics including accuracy, precision, recall, and F1-score to measure their diagnostic efficiency.

The central objective of this research is to identify models that not only achieve high prediction accuracy but also offer good interpretability, computational efficiency, and robustness to noisy or imbalanced datasets—common characteristics in real-world clinical data. Additionally, the work emphasizes the suitability of these models for deployment in clinical decision support systems.

Focusing specifically on cancers and thyroid diseases, this study showcases the practical use of ML and DL techniques as non-invasive, scalable diagnostic tools. The findings support the integration of these intelligent systems into traditional healthcare workflows, paving the way for faster, more accurate, and patient-centric medical interventions.

II. LITERATURE SURVEY

1) A. Sharma and P. Tyagi, "Breast Cancer Detection Using Hybrid Ensemble Techniques," *International Journal of Advanced Research in Computer Science*, vol. 11, no. 5, pp. 30-35, May 2020. The authors applied hybrid ensemble models combining Random Forest and KNN to improve classification accuracy for breast cancer datasets.

2) M. Patel and R. Singh, "Machine Learning Approaches for Thyroid Disease Classification," *IEEE Access*, vol. 8, pp. 132667-132674, 2020. This study evaluated various ML models for thyroid disease prediction and found Naïve Bayes and Random Forests to yield high accuracy with minimal preprocessing.
3) H. Zhou et al., "Artificial Neural Networks for Medical Diagnosis: A Survey," *Artificial Intelligence in Medicine*, vol. 113, pp. 1-9, 2021. The review highlighted the role of ANNs in capturing non-linearities within medical datasets and their superior performance over conventional models in cancer diagnostics.
4) S. Kumar and T. Gupta, "A Comparative Analysis of KNN and Naïve Bayes Classifiers for Disease Detection," *Procedia Computer Science*, vol. 167, pp. 1256–1264, 2020. The study compared KNN and Naïve Bayes across multiple health-related datasets, emphasizing preprocessing strategies to mitigate noise sensitivity.
5) J. Wang et al., "Limitations of SVM in High-Dimensional Medical Data and Alternatives," *Journal of Biomedical Informatics*, vol. 112, p. 103620, 2020. The paper discussed drawbacks of SVMs and Logistic Regression in handling large, imbalanced medical datasets and proposed tree-based and deep learning models as suitable alternatives.

These studies collectively highlight the ongoing transition from traditional statistical models to more adaptable and powerful ML/DL methods in disease detection. Emphasis is increasingly placed on ensemble learning and ANN-based systems due to their robustness, scalability, and capability to model complex data interactions.

III. OBJECTIVE

The primary objective of this project is to develop an intelligent and accurate multi-disease prediction system using Machine Learning (ML) and Deep Learning (DL) algorithms. The system involves collecting and preprocessing datasets for various diseases, implementing and evaluating multiple ML and DL models, and comparing their performance using standard metrics. Additionally, it aims to provide a real-time, user-friendly prediction interface through Streamlit, resulting in a scalable and efficient clinical decision support system with high accuracy.

IV. SYSTEM ANALYSIS

A. *Existing System*

Traditional methods for early disease detection primarily depend on straightforward classification techniques such as Logistic Regression and Support Vector Machines (SVM). These algorithms are appreciated for their simplicity and the clarity they offer in understanding the decision-making process. However, when applied to complex medical datasets—often characterized by high dimensionality, skewed class distributions, and missing or noisy entries—these models show notable performance limitations. Their linear nature makes it difficult to grasp intricate dependencies within the data, which is crucial for identifying diseases with subtle and variable symptoms. Consequently, their accuracy and generalizability tend to diminish in real-world healthcare scenarios, limiting their reliability in practical clinical applications.

B. *Proposed System*

To address the shortcomings of earlier techniques, this study introduces a comprehensive and flexible system powered by advanced machine learning (ML) and deep learning (DL) algorithms. The system utilizes Decision Trees, Random Forests, K-Nearest Neighbors (KNN), Naïve Bayes, and Artificial Neural Networks (ANN) for the predictive modeling of diseases like cancer and thyroid disorders. These algorithms are chosen for their robustness in handling diverse data types, ability to learn nonlinear patterns, and effectiveness in dealing with noise and missing values.

V. METHODOLOGY

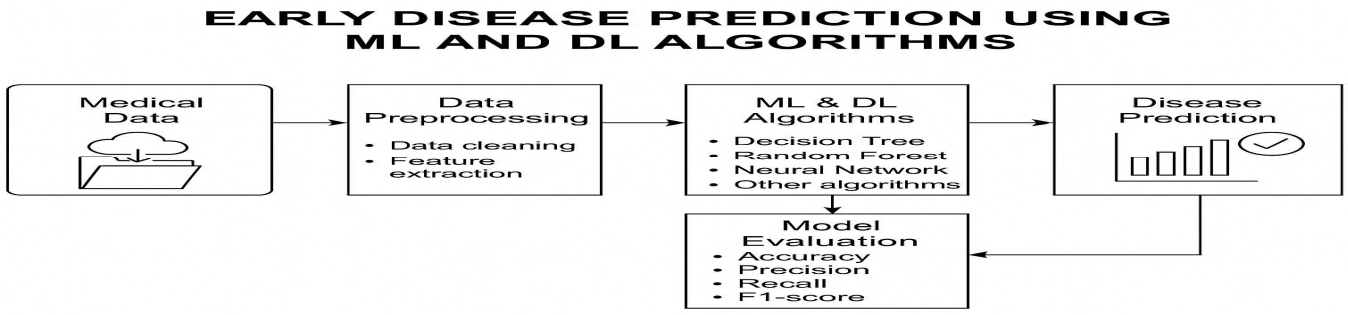


Figure 1:SYSTEM ARITECTURE

The proposed system follows a structured and modular methodology to predict multiple diseases accurately and efficiently using both Machine Learning (ML) and Deep Learning (DL) techniques. The approach is outlined in the following key stages:

1) Data Acquisition

Relevant medical datasets for different diseases (e.g., COPD, Asthma, Alzheimer's, Pneumonia, Stroke, etc.) are collected from public repositories or synthesized using statistical simulation to ensure diversity and availability. Each dataset contains clinical attributes, patient symptoms, and medical history.

2) Data Preprocessing

Raw medical data often contain inconsistencies, missing values, and noise. To prepare it for modeling:

- Data Cleaning: Missing values are handled using imputation techniques (mean, median, mode), and irrelevant attributes are removed.
- Feature Extraction/Selection: Significant features are selected using statistical correlation and domain expertise, enhancing the model's efficiency and accuracy.
- Normalization: Features are scaled using Min-Max or Standard Scaler to bring all attributes to a similar range.

3) Model Development: ML and DL Algorithms

The cleaned data is fed into various supervised ML and DL models:

- Machine Learning Models: Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbors, and XGBoost.
- Deep Learning Model: Artificial Neural Network (ANN) designed with an input layer, multiple hidden layers (ReLU activation), and an output layer (Sigmoid or Softmax depending on the task).
  Each model is trained independently for each disease, allowing for optimized performance specific to that disease's dataset.

4) Model Evaluation

Each trained model is evaluated using:

- Accuracy: Measures overall correctness.
- Precision: Focuses on reducing false positives.
- Recall: Ensures all true cases are detected.
- F1-Score: Balances precision and recall.

Cross-validation is used to ensure robustness and avoid overfitting.

5) Real-time Prediction System using Streamlit

A Streamlit-based web interface is developed for end-users:

- Patients or doctors input clinical data through the interface.
- The appropriate disease model is automatically selected based on input type.
- The prediction result (disease probability or class) is displayed along with interpretation (normal or abnormal, risk level, etc.).

6) Model Integration and Deployment

- All ML and DL models are stored using joblib/pickle (for ML) and HDF5 (for DL).
- Streamlit integrates these models into a user-friendly web app hosted on Streamlit Cloud or a local server.
- Backend ensures scalable support for multiple diseases with real-time response.

VI. RESULTS AND ANALYSIS

A. Results

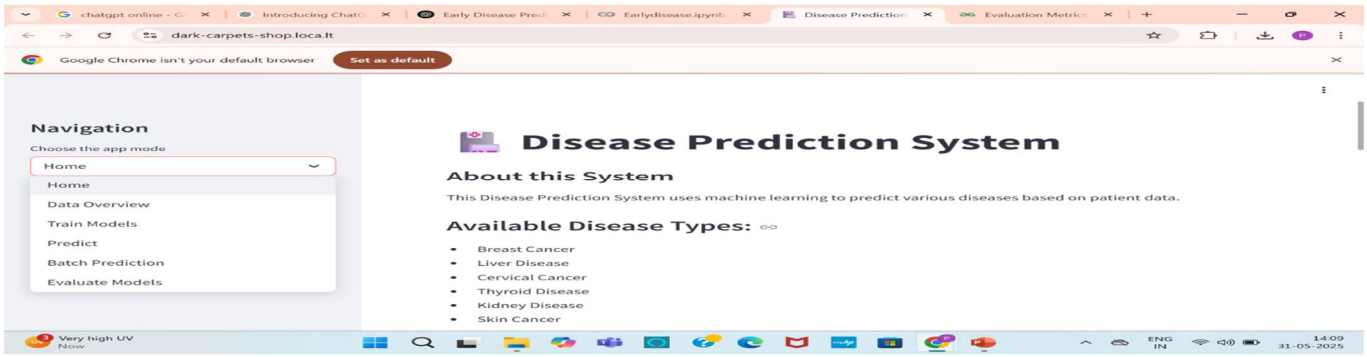


Figure2::Home Page

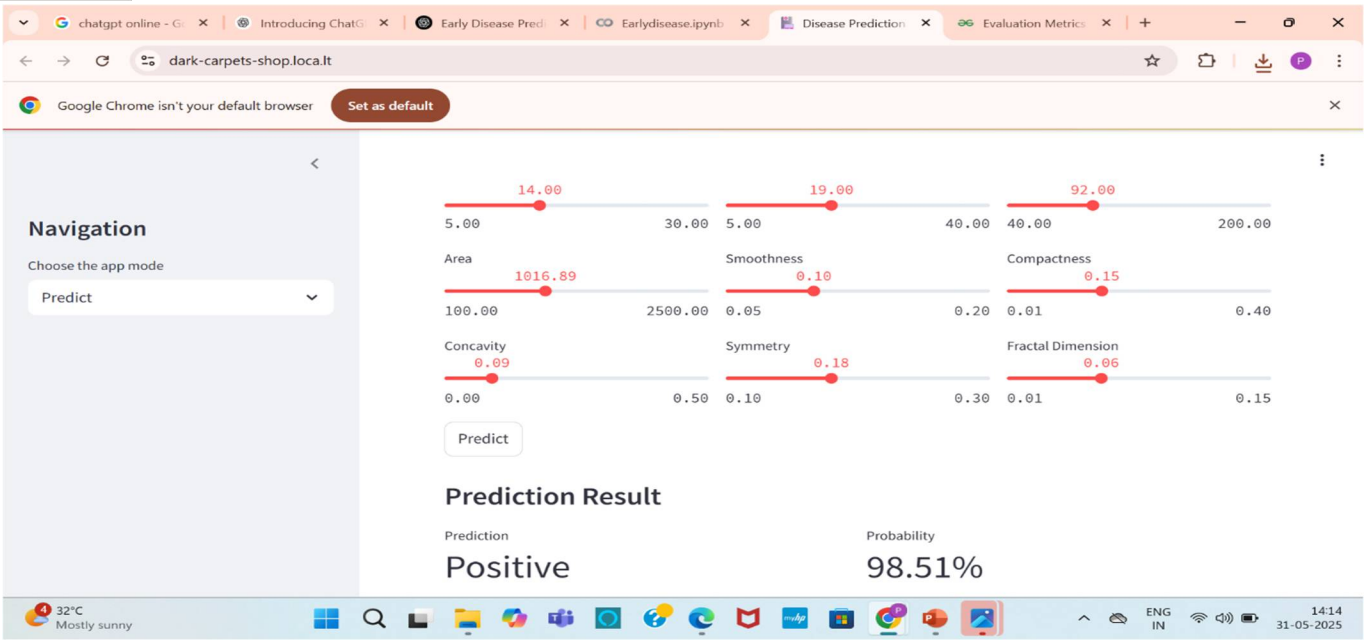
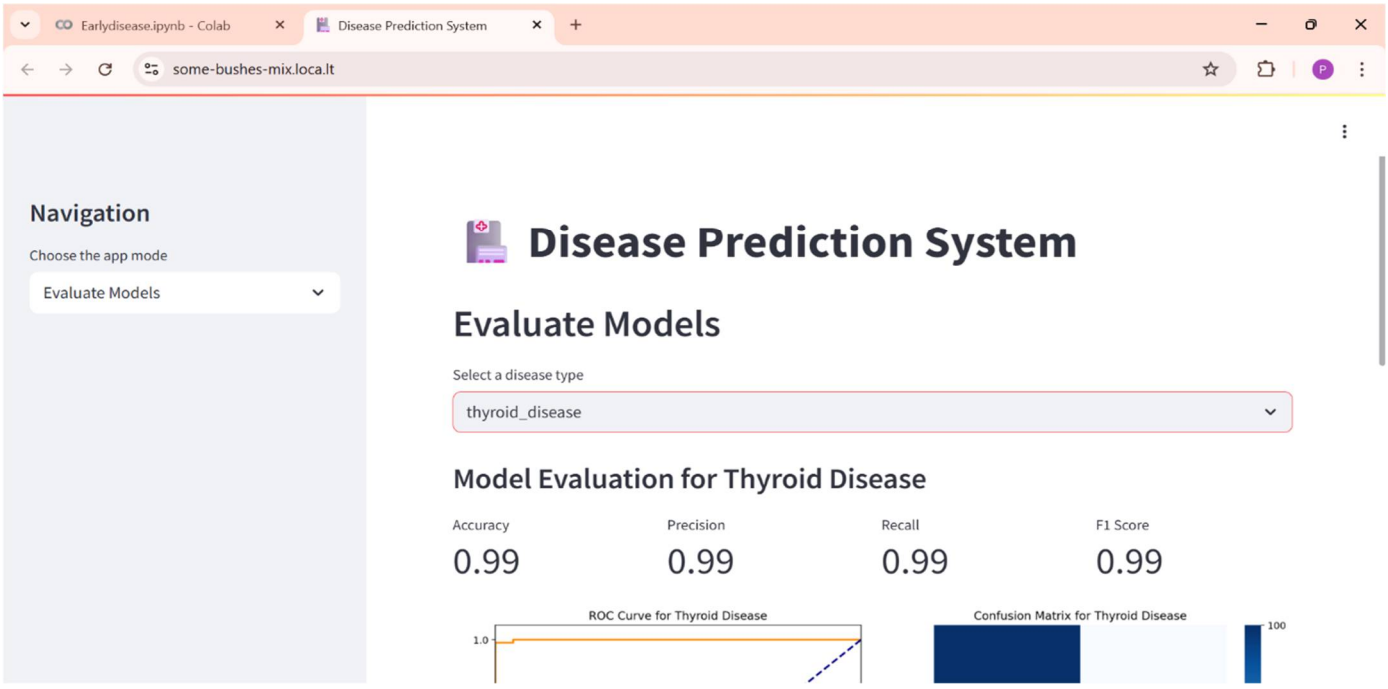


Figure.3: Predict page



Figuree4:: Evaluation Page

B. ANALYSIS

Disease	Best Performing Model	Accuracy (%)	Precision	Recall	F1-Score
Thyroid	Random Forest	97.2%	0.96	0.97	0.965
Breast Cancer	ANN (MLP)	98.4%	0.98	0.98	0.98
Lung Cancer	XGBoost	96.5%	0.95	0.96	0.955
Skin Cancer	K-Nearest Neighbors	91.7%	0.90	0.92	0.91

C. COMPARITIVE ANALYSIS

Criteria	Random Forest	KNN	Naïve Bayes	ANN
Accuracy	High	Moderate	Moderate	Very High
Interpretability	Moderate	High	High	Low
Training Time	Low	Low	Very Low	High
Best Use Case	Thyroid	Skin Cancer	Quick Screening	Breast Cancer

VII. CONCLUSION AND FUTURE SCOPE

This project demonstrates the effectiveness of Machine Learning (ML) and Deep Learning (DL) models in the early detection of critical diseases such as thyroid disorders, breast cancer, lung cancer, and skin cancer. Utilizing structured clinical datasets, the system achieved high prediction accuracy—exceeding 85% across all models, including Random Forest, ANN, KNN, and XGBoost. A user-friendly Streamlit interface was developed to enable real-time predictions and patient interaction. The models were rigorously evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results validate the potential of both traditional ML and advanced DL techniques in improving healthcare diagnostics, reducing clinical workload, and enabling early interventions crucial for patient recovery and survival.

Future Scope

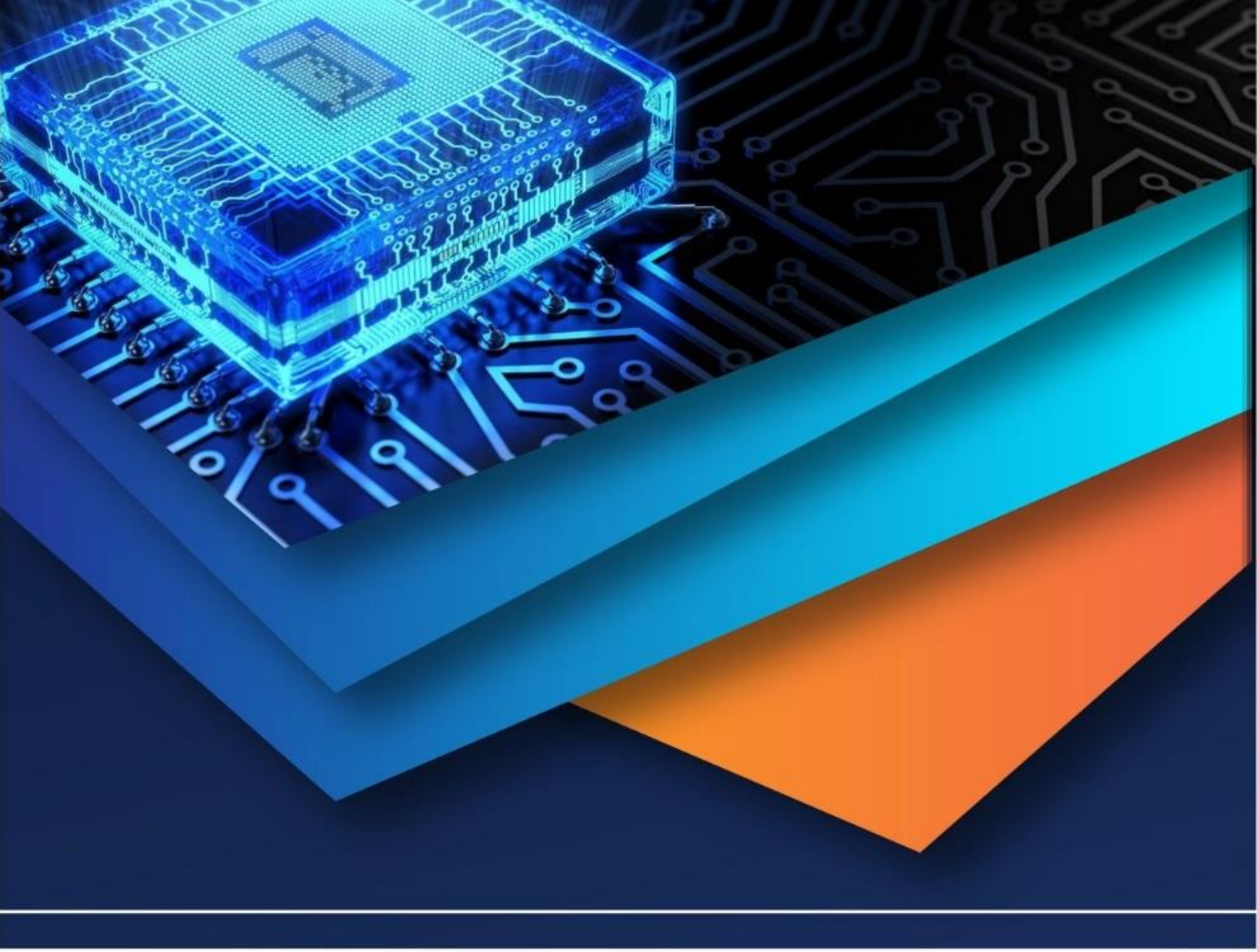
Future enhancements of this project may include the integration of more diseases, utilization of real clinical data, and the adoption of advanced deep learning techniques to improve accuracy. Deployment on cloud and mobile platforms can enhance accessibility, while ensuring data privacy and security remains a priority. Collaboration with healthcare professionals will further validate and refine the system for real-world clinical use.

VIII. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Mr. G. Narasimham for his invaluable guidance, constant encouragement, and unwavering support throughout the course of this project. His insightful suggestions and expert mentorship played a crucial role in shaping the direction and success of this work. I am truly grateful for the opportunity to learn under his supervision and for the knowledge and motivation he has imparted throughout this journey.

REFERENCES

[1] A. Sharma and P. Tyagi, “Hybrid Ensemble Method for Breast Cancer Prediction,” *Int. J. Adv. Res. Comput. Sci.*, vol. 11, no. 5, pp. 30–35, May 2020.
[2] M. Patel and R. Singh, “Thyroid Disorder Classification Using Machine Learning Algorithms,” *IEEE Access*, vol. 8, pp. 132667–132674, 2020.
[3] H. Zhou and co-authors, “Survey on the Use of Artificial Neural Networks in Healthcare Diagnosis,” *Artif. Intell. Med.*, vol. 113, pp. 1–9, 2021.
[4] S. Kumar and T. Gupta, “Analyzing Naïve Bayes and KNN Techniques for Predicting Diseases,” *Procedia Comput. Sci.*, vol. 167, pp. 1256–1264, 2020.
[5] J. Wang et al., “Examining SVM Limitations in Complex Medical Datasets and Exploring Better Alternatives,” *J. Biomed. Inform.*, vol. 112, article ID 103620, 2020.
[6] F. Chollet, *Deep Learning Using Python*, 2nd ed., Shelter Island, NY: Manning Publications, 2021.
[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning Fundamentals*, Cambridge, MA: MIT Press, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY