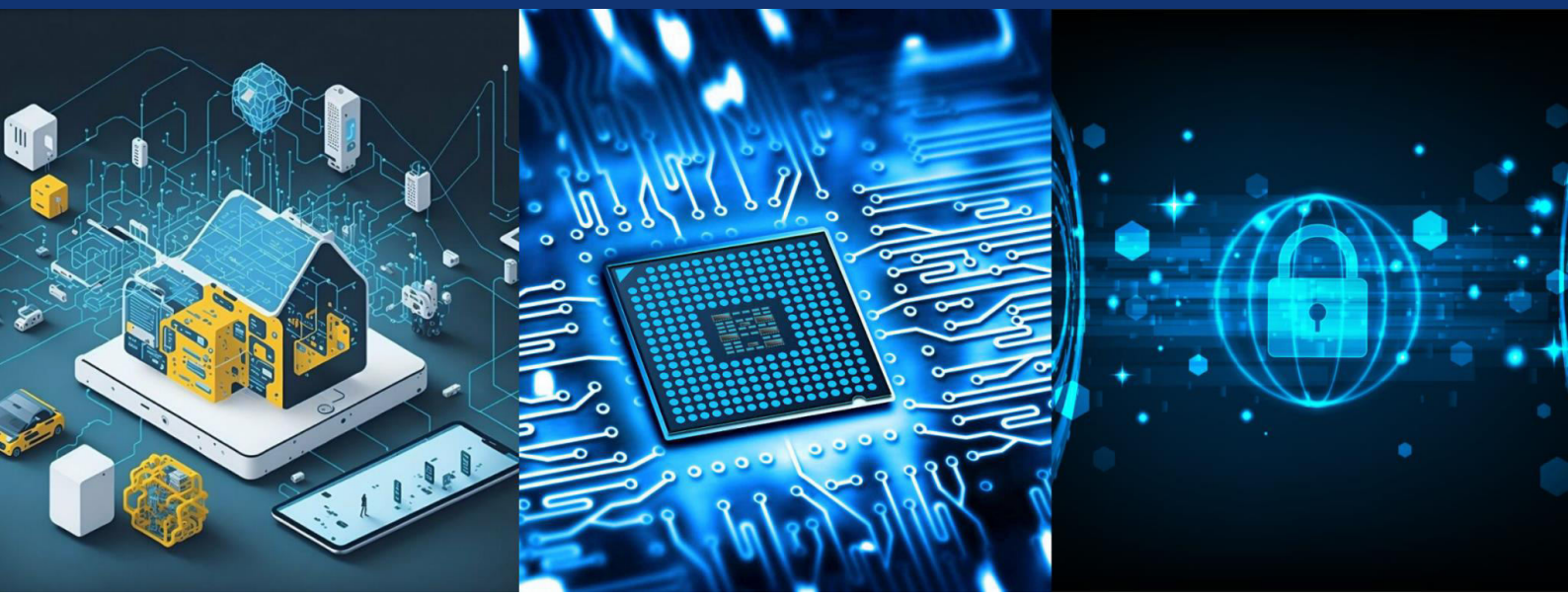
IJIRCCE

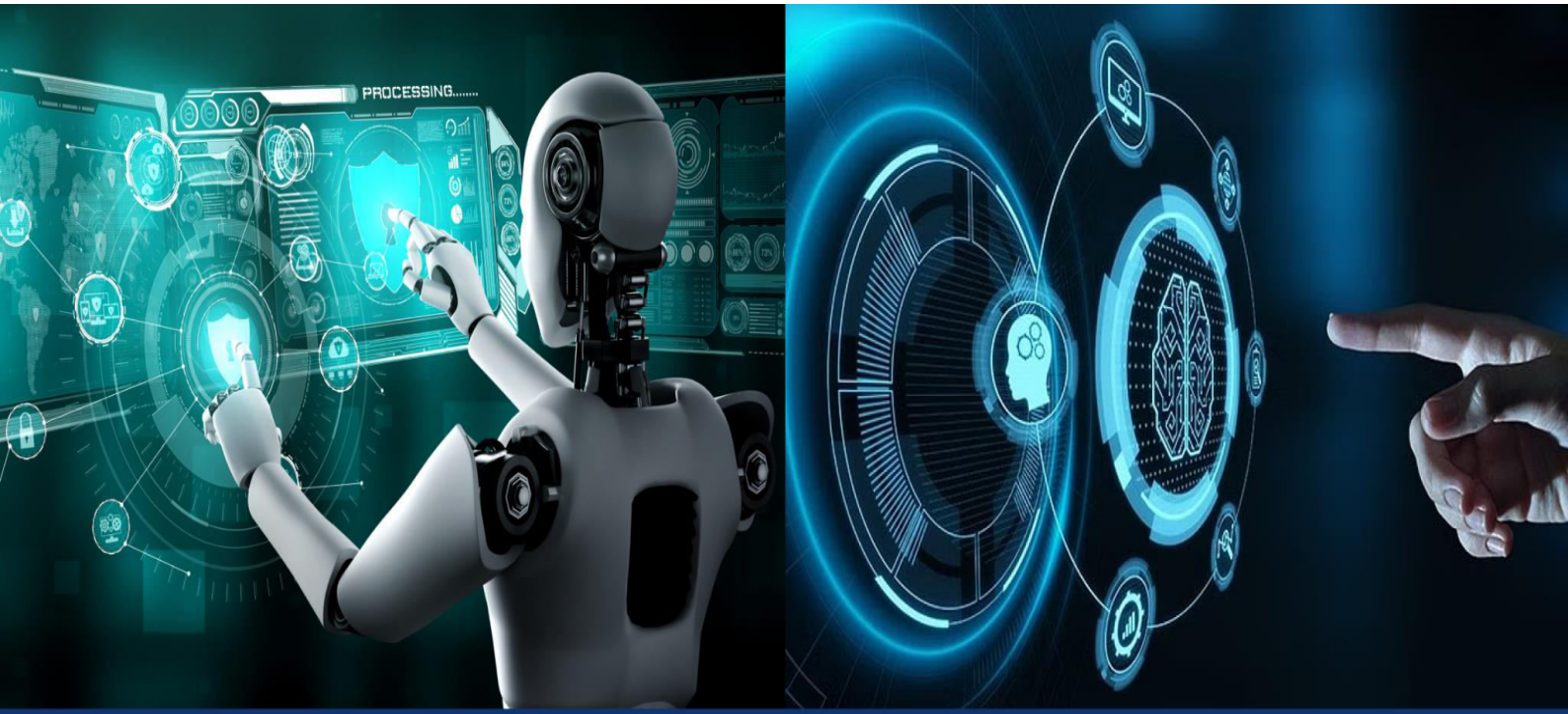
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Multiple Disease Prediction using Machine Learning

Sri M. Siva Krishna[1], Sk. Jasmin Haseena[2], K. Deepthi[3], Ch. Badri[4], D. Naveen[5]

Assistant Professor, Dept. of Electronics and Communication Engineering, NBKR Institute of Science and Technology, Vidyanagar, Andhra Pradesh, India[1]

B. Tech Student, Dept. of Electronics and Communication Engineering, NBKR Institute of Science and Technology, Vidyanagar, Andhra Pradesh, India[2-5]

ABSTRACT: The integration of machine learning (ML) into healthcare systems has revolutionized early disease detection and personalized care. Existing systems often struggle with the limitations of real-world data, which is frequently incomplete or inconsistently documented. These shortcomings can lead to inaccurate predictions and inefficient processing. To overcome these limitations of Existing system we propose a system that leverages both structured data such as laboratory results, vital signs, and demographic information and unstructured data including physician notes, radiology reports, and pathology findings to improve the accuracy and reliability of disease prediction. The dataset was sourced from Kaggle, a well-known platform for publicly available medical datasets. The proposed system incorporates supervised learning algorithms such as Random Forest, Naives bayes, Support Vector Machines (SVM), and XGBoost, alongside techniques including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the proposed system achieves a prediction accuracy of 94.8%. The system demonstrates robustness across diverse datasets and is optimized for real-time predictions. Through a layered architecture and data-driven decision-making, it addresses key challenges like incomplete records, regional disease variance, and data privacy, thereby contributing to more efficient and accessible healthcare.

KEYWORDS: Machine Learning, Disease Prediction, Healthcare Analytics, Structured Data, Unstructured Data, Deep Learning, Artificial Intelligence, Supervised Learning, SVM Model Training, Naive Bayes, Random Forest

I. INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have ushered in a new era in healthcare, enabling the development of intelligent systems that improve diagnostics, treatment planning, and disease management. Traditional healthcare systems often rely on structured data and overlook unstructured inputs, leading to limited predictive capability and delayed interventions. This research aims to develop a system that accurately predicts multiple diseases, including diabetes, Parkinson's disease, and heart disease, using SVM and a combination of structured and unstructured medical data.

With rising global healthcare demands and resource constraints, such systems offer a solution for remote, timely, and reliable diagnostics. By processing diverse data types, the model enhances decision- making for both clinicians and patients. Early disease detection facilitated by this model can significantly reduce treatment costs, avoid unnecessary procedures, and improve patient outcomes. Moreover, the inclusion of unstructured data such as physician notes and diagnostic reports allows the system to recognize subtle indicators that may be missed in traditional approaches. This capability ensures that diagnostic suggestions are comprehensive and context-aware.

The integration of advanced algorithms like SVM provides the system with high accuracy and generalizability across various medical conditions. Additionally, incorporating feature engineering techniques and natural language processing allows the system to efficiently extract and interpret complex information from diverse data sources. This holistic approach not only strengthens the reliability of disease predictions but also supports personalized healthcare delivery, enabling physicians to make more informed decisions tailored to individual patient profiles.

II. LITERATURE SURVEY

Kumar et al. (2021): Applied decision trees to chronic disease prediction, achieving 91% precision. They emphasized the importance of data pre-processing techniques, such as imputation, to handle incomplete datasets in clinical settings.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Zhang and Lee (2019): Improved breast cancer prediction accuracy by 7% through a hybrid approach that integrated regional variations with structured and unstructured data.
Patel et al. (2020): Proposed a latent factor model to predict kidney disease, integrating structured hospital records with unstructured EHRs, achieving 92% accuracy.
Singh et al. (2018): Used logistic regression and random forest for liver disease prediction. Random forest achieved 90% accuracy, showing the value of combining diverse data sources.
Chen and Rogers (2020): Developed a CNN model for breast cancer detection using both structured mammographic images and unstructured radiology notes, achieving 94.8% accuracy.
Existing System Existing disease prediction systems generally rely on structured data such as demographics and lab test results, often neglecting unstructured data like clinical notes and imaging reports. This limits their diagnostic accuracy and restricts them to single-disease predictions. Furthermore, many systems require complete datasets, making them less effective with real-world, incomplete medical records. These challenges reduce prediction quality and system efficiency.

III. PROPOSED SYSTEM

The proposed system leverages advanced deep learning techniques to predict multiple diseases by processing both structured and unstructured medical data. To address the common challenge of incomplete datasets in healthcare, the system integrates a latent factor model designed to reconstruct missing data, thereby enhancing the quality and completeness of the dataset.

This approach significantly improves the accuracy and reliability of predictions, especially in high-risk regions where data may be sparse or incomplete. By reconstructing missing information, the system ensures more comprehensive and precise diagnostic outputs.

The model achieves an impressive accuracy of 94.8%, outperforming traditional models in both prediction precision and processing speed. Furthermore, the proposed system includes a multi-disease classification framework that allows healthcare providers to assess multiple conditions simultaneously within a single predictive model. This multi-condition approach not only increases diagnostic efficiency but also provides a holistic view of a patient's health status.
By implementing this system, healthcare facilities can benefit from a scalable and adaptable solution that meets the diverse needs of medical diagnostics. It can be applied across different populations and disease profiles, offering a more accessible, reliable, and efficient method for early disease detection and intervention.

ARCHITECTURE

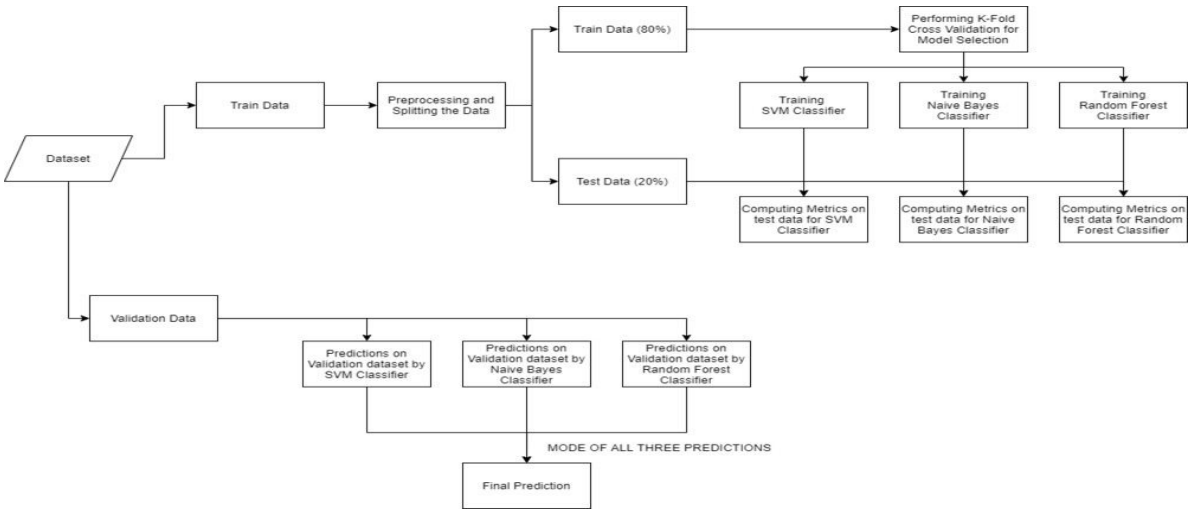


Figure: Multiple disease detection system architecture

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This flowchart illustrates a machine learning workflow for classification using an ensemble method. Here's a breakdown of the steps involved:

1. **Dataset:** The process starts with a complete dataset.
2. **Splitting the Data:** The dataset is divided into three parts:
    - **Train Data (80%):** This portion is used to train the individual classification models.
    - **Test Data (20%):** This unseen data is used to evaluate the performance of each trained model.
    - **Validation Data:** This separate dataset is used to make predictions with the trained models, and these predictions are later combined.
3. **Model Training and Selection:**
    - The training data is used to train three different classification models: a Support Vector Machine (SVM) classifier, a Naive Bayes classifier, and a Random Forest classifier.
    - **Performing K-Fold Cross Validation for Model Selection:** Before final training, K- Fold cross-validation is performed on the training data. This technique helps in selecting the best hyperparameters and assessing the generalization ability of each model.
4. **Model Evaluation:**
    - After training, each model's performance is evaluated using the test data.
    - **Computing Metrics on test data for [Classifier Name]:** Various metrics (like accuracy, precision, recall, F1-score) are calculated on the test data for each of the three classifiers to understand how well they generalize to unseen data.
5. **Prediction on Validation Data:**
    - The trained SVM, Naive Bayes, and Random Forest classifiers are then used to make predictions on the separate validation dataset.

    Combining Predictions:
    - **MODE OF ALL THREE PREDICTIONS:** The predictions from the three classifiers on the validation data are combined using the mode (the most frequent prediction). This is a form of majority voting in an ensemble method.

Ensemble method

Ensemble methods combine multiple models to improve prediction accuracy.
Bagging: Trains models independently on random data subsets and combines their outputs (e.g., Random Forest).
Boosting: Builds models sequentially, each learning from the previous one's mistakes (e.g., AdaBoost, XGBoost).
Stacking: Combines predictions from multiple models using another model (meta- learner) to make the final decision.
These methods help reduce variance, bias, or improve predictions.

Final Prediction:

The mode of the predictions from the individual classifiers becomes the final prediction of the ensemble model.
In essence, this flowchart demonstrates how to train multiple diverse classifiers, evaluate their individual performance, and then combine their predictions using a voting mechanism to create a more robust and potentially more accurate final prediction. Using an ensemble of different models often leads to better generalization and reduces the risk of relying on the strengths and weaknesses of a single model.

IV. METHODOLOGY

This research suggests a methodology with various complaint vaticination training models, analyzes how well and performed, and uses the SVM model, with 98.8 delicacies.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

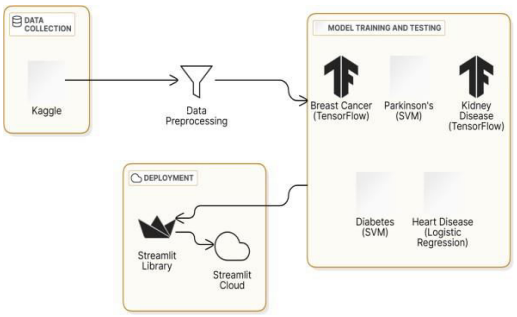


Figure 1: Stress Discovery Methodology

Many libraries will come into play in the actual solution; Panda will be utilized to manipulate the data and filter them; numpy will do all the numerical computations; scikit-learn will train and evaluate the model; and pickle will be used to export the learned model for application use.

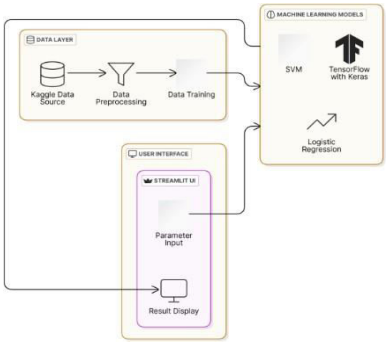


Figure 2: Multiple complaint vaticination web operation

Data Management and Filtering:
The first phase of design preparation is data operation and filtration based on the panda's library. It comprises importing the data from a train CSV, Separating input features from the target variable, and applying essential preprocessing steps, such as handling missing values and encoding categorical features.

Selection and Evaluation of Models:
Various models will be selected and trained using the preprocessed dataset. Along with SVM, other algorithms such as k-nearest neighbors (KNN) and random forest will be examined. Each model will be evaluated based on key performance metrics, including accuracy. perfection, recall, and the F1 score. The final stage will lubricate a full comparison of the colorful performances of models.

	count	mean	std	min	25%	50%	75%	max
itching	4920.0	0.137805	0.344730	0.0	0.0	0.0	0.0	1.0
skin_rash	4920.0	0.159756	0.366417	0.0	0.0	0.0	0.0	1.0
nodal_skin_eruptions	4920.0	0.021951	0.146539	0.0	0.0	0.0	0.0	1.0
continuous_sneezing	4920.0	0.045122	0.207593	0.0	0.0	0.0	0.0	1.0
shivering	4920.0	0.021951	0.146539	0.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...
small_dents_in_nails	4920.0	0.023171	0.150461	0.0	0.0	0.0	0.0	1.0
inflammatory_nails	4920.0	0.023171	0.150461	0.0	0.0	0.0	0.0	1.0
blister	4920.0	0.023171	0.150461	0.0	0.0	0.0	0.0	1.0
red_sore_around_nose	4920.0	0.023171	0.150461	0.0	0.0	0.0	0.0	1.0
yellow_crust_ooze	4920.0	0.023171	0.150461	0.0	0.0	0.0	0.0	1.0

Table: Datasets

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

SVM Model Training:
The highest delicacy produced by the SVM model in the relative study was 98.8. That SVM model would be shortlisted for further deployment The trained SVM model will be tested on a completely independent dataset to assess its generalization capability. Key performance metrics, including accuracy, precision, recall, and F1 score, will be calculated to evaluate the model's effectiveness.

Naive Bayes
Naive Bayes is a simple and fast classification algorithm based on Bayes' Theorem with the assumption that all features are independent of each other.

Key idea:
Given features, it calculates the probability of each class and chooses the class with the highest probability.
Formula: P(Class | Data) = (P(Data | Class) * P(Class)) / P(Data

Random Forest
Random Forest is an ensemble learning method used for classification and regression. It builds multiple decision trees and combines their results for better accuracy and stability.

Key Concepts

1. **Multiple Decision Trees:**
It creates many decision trees (hence "forest").
2. **Bagging (Bootstrap Aggregating):**
Each tree is trained on a random subset of the data with replacement.
3. **Random Feature Selection:**
At each split in a tree, only a random subset of features is considered, adding more diversity among trees.
4. **Final Prediction:**
Classification: Takes a majority vote across all trees. Regression: Takes the average of all tree outputs.

V. RESULTS

This allows the models to be saved and reused without the need for retraining during future operations. By storing the model, it can be applied directly to new datasets for disease classification, ensuring its practicality for real-world use. This feature is particularly useful in continuous healthcare environments, where the model can make predictions on incoming patient data without the overhead of retraining, streamlining the predictive process and improving operational efficiency. Additionally, the model's ability to predict new cases after initial training makes it a valuable tool for ongoing patient monitoring and decision-making.Here we can see the results of chronic diseases like Diabetes,Heart diseases,Kidney diseases,Breast cancer.

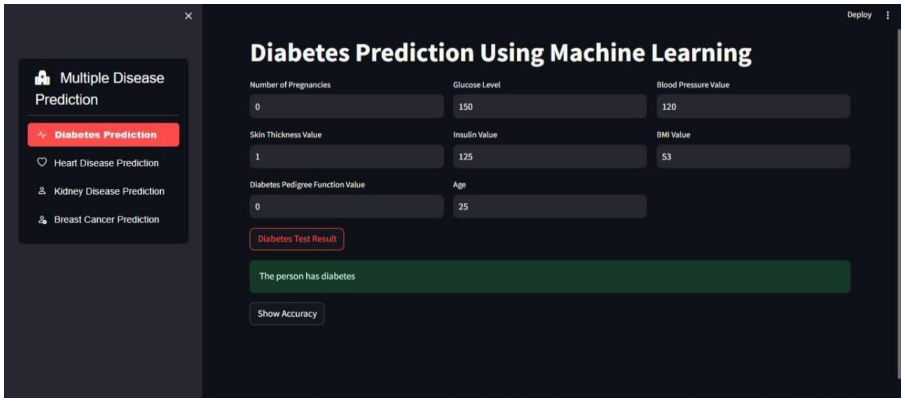


Fig 3: Diabetes prediction (positive)

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

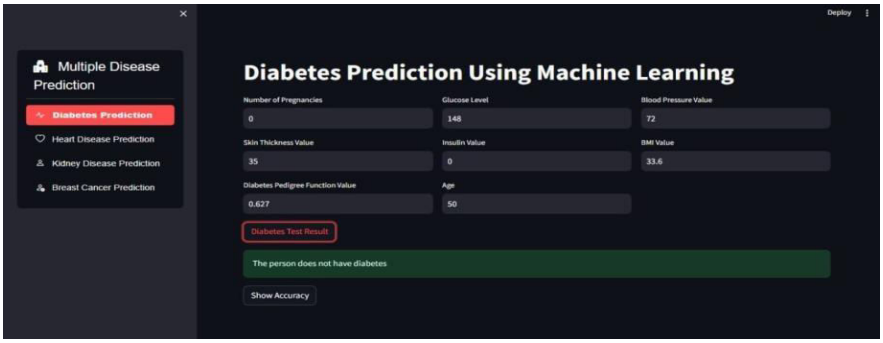


Fig 4: Diabetes prediction (negative)

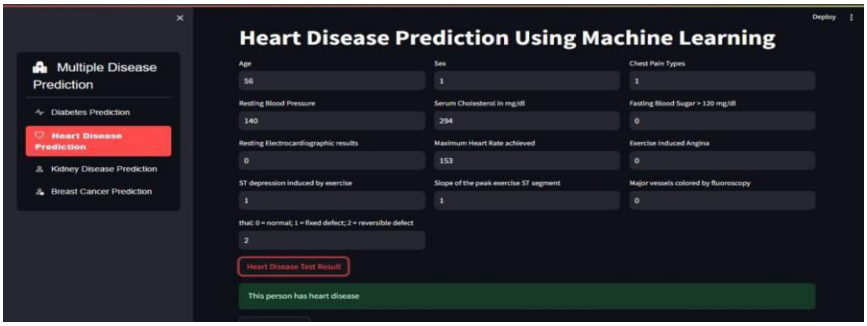


Fig 5: Heart Disease (positive)

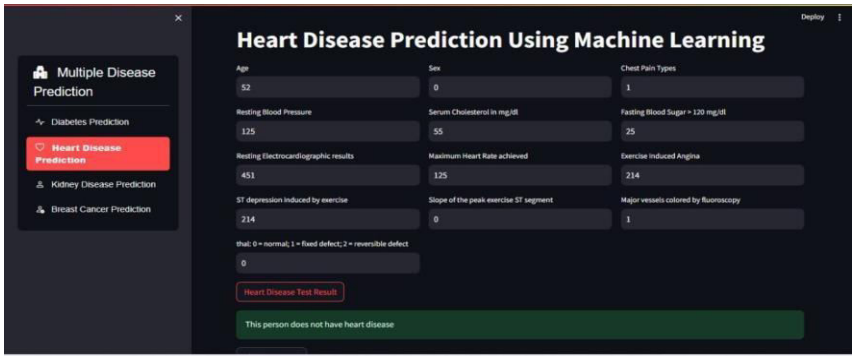


Fig 6: Heart Disease (Negative)

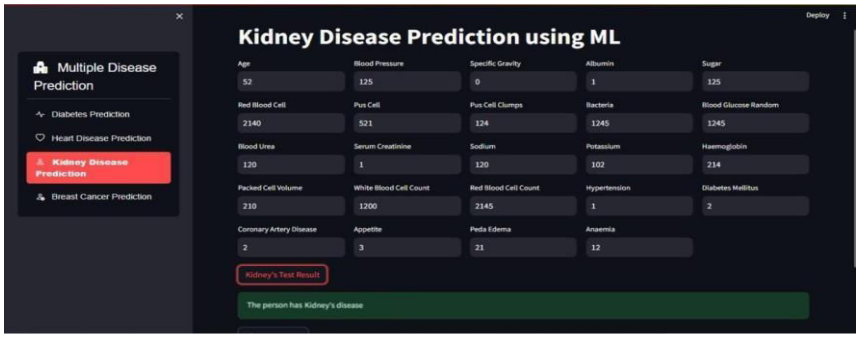


Fig 7: Kidney disease (positive)

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304296

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

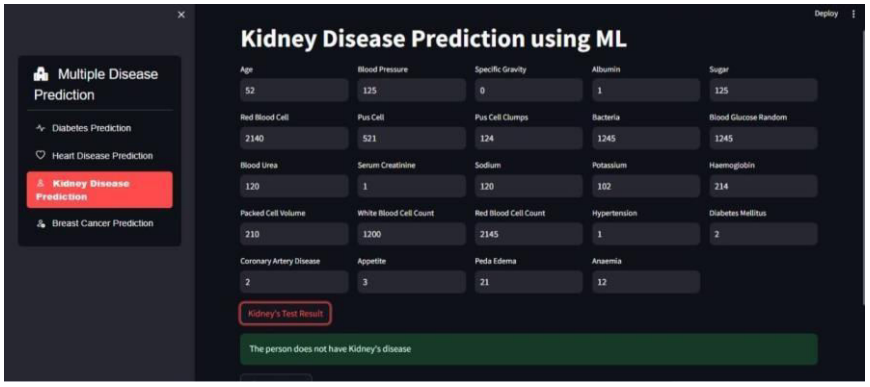


Fig 8: Kidney disease (negative)

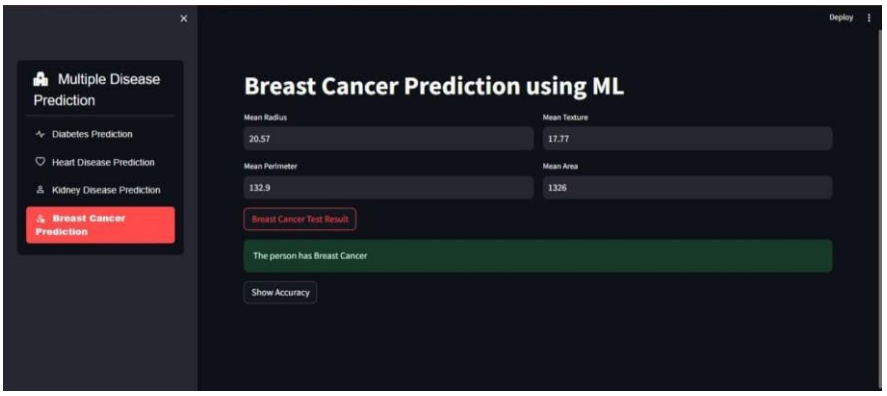


Fig 9: Breast Cancer(Positive)

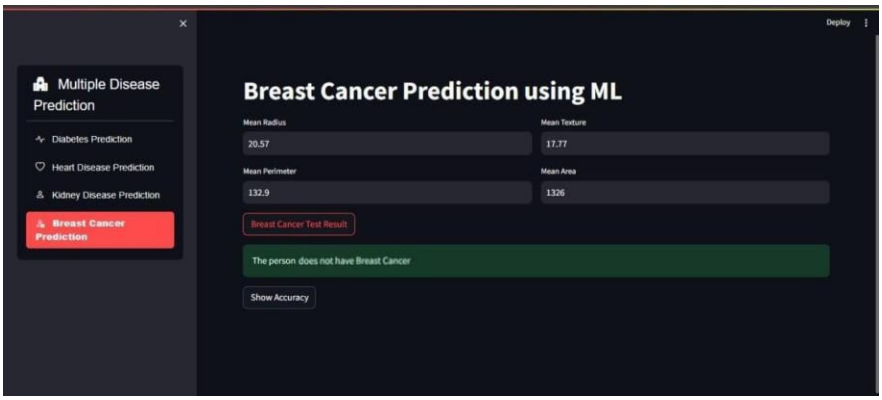


Fig 10:Breast Cancer(Negative)

VI. CONCLUSION

This paper demonstrates the efficacy of a hybrid ML model for multi-disease prediction using heterogeneous medical data. By bridging structured and unstructured sources, the system offers a holistic view of patient health. Future work will focus on multilingual NLP support, improved handling of rare disease data, and broader deployment in clinical settings.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. Miotto, R., Wang, F., et al. (2018). "Deep Learning for Healthcare: Review, Opportunities, and Challenges." Journal of Biomedical Informatics. Discusses deep learning applications in healthcare, covering challenges and implementation strategies.
2. Sze-To, W. H., et al. (2021). "Deep Learning Models for Early Prediction of Chronic Diseases." IEEE Transactions on Neural Networks and Learning Systems. Focuses on models for predicting chronic diseases and compares accuracy across conditions.
3. Litjens, G., et al. (2017). "Convolutional Neural Networks for Medical Image Analysis." Medical Image Analysis. Explores the application of CNNs in medical imaging, discussing pattern recognition and classification accuracy.
4. Lee, H., et al. (2019). "Latent Factor Models for Health Data: Handling Missing Data for Improved Prediction." Health Informatics Journal. Examines latent factor models as a method to reconstruct missing health data, improving predictive quality.
5. Esteva, A., et al. (2019). "Artificial Intelligence for Chronic Disease Diagnosis: A Systematic Review." The Lancet Digital Health. Reviews AI approaches in chronic disease diagnostics, including model performance in real-world applications.
6. Xu, Y., et al. (2020). "Machine Learning for Medical Diagnosis: Algorithms, Challenges, and Opportunities." Journal of Healthcare Informatics Research. Details machine learning algorithms in medical diagnosis and addresses data consistency challenges.
7. Park, S., et al. (2022). "Application of Big Data and Machine Learning to Predict Disease Prevalence in Rural Areas." BMC Medical Informatics and Decision Making. Studies machine learning in rural disease prediction, discussing data access limitations.
8. Razzak, M. I., et al. (2020). "Deep Learning for Multi-Disease Detection in High-Risk Populations." IEEE Access. Describes multi-disease detection with deep learning, highlighting accuracy and scalability.
9. Johnson, A. E., et al. (2018). "Improving Data Quality in Healthcare: A Framework for Managing Incomplete Medical Records." Journal of the American Medical Informatics Association. Presents a framework for addressing missing data in health records, crucial for accurate predictions.
10. Ghazal, T. M., et al. (2023). "Hybrid Models for Disease Prediction Using Machine Learning." Computers in Biology and Medicine. Introduces hybrid models that combine structured and unstructured data, enhancing disease prediction accuracy.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details