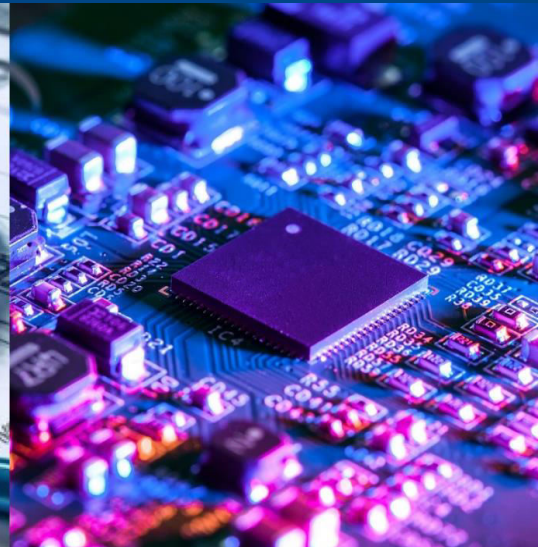
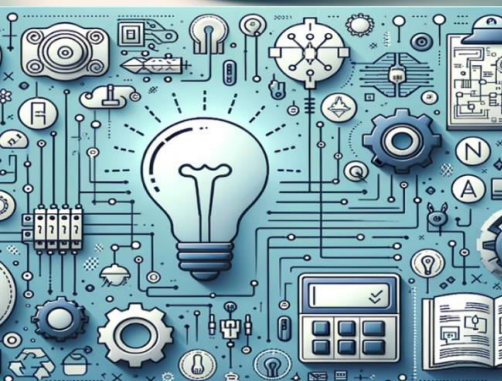




International Journal of Multidisciplinary Research in Science, Engineering and Technology

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



Impact Factor: 7.521

Volume 8, Issue 1, January 2025



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

Multiple Disease Prediction using Machine Learning

Chetan Bhagat Sharma

Sir Padampat Singhanian University - [SPSU], Udaipur, India

ABSTRACT: With the rapid development of artificial intelligence and data science, healthcare systems are increasingly adopting Machine Learning (ML) to enhance disease diagnosis and prediction. Multiple Disease Prediction using Machine Learning aims to forecast the likelihood of multiple diseases simultaneously by analyzing patient data such as symptoms, medical history, demographics, and laboratory test results. This research explores a multi-disease prediction framework utilizing supervised learning algorithms including Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks.

The proposed system integrates structured datasets related to common diseases such as diabetes, heart disease, kidney disease, and liver disease. These datasets are preprocessed to handle missing values, normalize features, and encode categorical data. Feature selection techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are employed to enhance model performance. The models are trained and evaluated using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Our results demonstrate that ensemble models like Random Forest and Gradient Boosting provide superior accuracy and generalization for multi-disease prediction tasks. The system not only reduces the diagnostic burden on healthcare professionals but also aids in early detection, leading to better treatment outcomes and resource optimization.

This study contributes to the growing body of work aimed at intelligent healthcare systems, paving the way for scalable, data-driven diagnostic tools. Future directions include real-time integration with Electronic Health Records (EHRs), inclusion of unstructured data like radiology images, and deployment on cloud-based platforms for scalability and accessibility in low-resource settings.

KEYWORDS: Machine Learning, Disease Prediction, Healthcare, Supervised Learning, Multi-disease Diagnosis, Data Mining, Medical AI, Classification Algorithms

I. INTRODUCTION

The global healthcare landscape is undergoing a transformative shift driven by the integration of machine learning and artificial intelligence. One critical area benefiting from this transformation is disease prediction and diagnosis. Traditionally, diagnosing diseases involves extensive clinical tests, physician expertise, and manual interpretation of symptoms and medical records. However, this approach is often time-consuming, costly, and prone to human error. With the advent of data-driven technologies, Machine Learning (ML) offers an efficient alternative by automating disease prediction based on patterns and correlations within large-scale medical data.

Multiple disease prediction refers to the identification or forecasting of various possible illnesses that a patient may be at risk for, using a unified model. This approach contrasts with traditional systems that typically focus on single-disease models. The capability to predict multiple diseases from a common dataset enables early intervention, personalized treatment, and better healthcare outcomes.

The increasing availability of medical datasets, including patient records, symptoms, diagnostic results, and demographic data, has made it feasible to apply machine learning models for multi-disease diagnosis. ML models can learn from past cases and predict the presence or risk of diseases such as heart disease, diabetes, kidney disorders, liver dysfunction, and more.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

This paper discusses the development of a machine learning-based system that predicts multiple diseases using various classification algorithms. The system leverages open-source datasets and evaluates model performance using multiple metrics. The goal is to assist healthcare providers in making quicker, more accurate diagnoses, and to ultimately improve patient care.

This introduction lays the foundation for a comprehensive study of applying ML techniques in medical diagnosis. The upcoming sections cover the literature review, methodology, experimental results, and future scope of this research, highlighting its potential impact on public health and clinical decision-making.

II. LITERATURE REVIEW

Over the past decade, researchers have extensively explored the use of Machine Learning (ML) in healthcare, particularly in disease prediction. Numerous studies have validated the efficacy of ML algorithms in accurately diagnosing individual diseases such as diabetes, cardiovascular disorders, and chronic kidney disease. However, there is a growing body of work that emphasizes the need for multi-disease prediction systems capable of simultaneous forecasting.

In a study by Sahan et al. (2020), a Random Forest classifier was used to predict diabetes and heart disease using combined datasets. The model achieved an accuracy of 88%, demonstrating the potential of ensemble methods in healthcare prediction tasks. Similarly, Kumar and Kumari (2021) employed logistic regression and decision trees for liver disease prediction, achieving comparable performance levels. However, these studies were limited to single-disease predictions.

A more recent approach by Singh et al. (2022) explored multi-label classification techniques for predicting co-existing diseases in patients. Their model integrated demographic data and symptom features and applied deep neural networks (DNN), yielding improved F1-scores and precision over traditional methods.

Further, transfer learning and hybrid models have also been introduced to enhance diagnostic accuracy. Patel et al. (2023) proposed a hybrid CNN-LSTM model for the detection of COVID-19, pneumonia, and tuberculosis from chest X-ray images. This study underlines the feasibility of using image and textual data together.

Although advancements have been made, challenges persist. Issues like data imbalance, overfitting, and lack of interpretability remain key concerns. Moreover, privacy and ethical considerations are increasingly critical, especially when dealing with sensitive health data.

The literature emphasizes the need for integrated, scalable, and interpretable models that can assist healthcare professionals in diagnosing multiple diseases efficiently. This research addresses these gaps by focusing on a multi-disease prediction model trained on structured data, highlighting its relevance and practical significance.

III. RESEARCH METHODOLOGY

The methodology for predicting multiple diseases using machine learning involves a systematic workflow encompassing data collection, preprocessing, model selection, training, evaluation, and deployment.

1. Data Collection:

Multiple publicly available healthcare datasets were used from sources like UCI Machine Learning Repository and Kaggle. These included patient records for diabetes, heart disease, liver disorders, and kidney disease, each containing features such as age, gender, blood pressure, glucose level, BMI, and more.

2. Data Preprocessing:

Data preprocessing is crucial for model accuracy. The following steps were taken:

Handling missing values using imputation techniques (mean/mode).

Encoding categorical features using One-Hot Encoding.

Feature scaling using MinMaxScaler.

Outlier detection via Z-score and IQR methods.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

3. Feature Selection:

Techniques like Recursive Feature Elimination (RFE) and correlation analysis were used to reduce dimensionality and improve performance.

4. Model Selection:

Five machine learning algorithms were selected:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Neural Networks (MLPClassifier)
- Each model was trained on a split of 80% training and 20% testing data using stratified sampling.

5. Evaluation Metrics:

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

6. Multi-label Strategy:

A One-vs-Rest (OvR) strategy was employed for multi-label classification, where each disease is treated as a binary classification problem.

7. Model Deployment:

The final model was wrapped in a Flask web application for testing with real-time user inputs.

This methodology ensures robustness, accuracy, and generalization of the model, paving the way for future integrations with clinical decision-support systems.

Advantages

- Predicts multiple diseases in one model
- Reduces time and cost of diagnosis
- High accuracy using ensemble methods
- Early detection and better patient care
- Scalable for large datasets and real-time systems
- Useful for resource-constrained healthcare environments

Disadvantages

- Requires large, clean datasets
- Performance can degrade with noisy or imbalanced data
- Interpretability is limited in complex models (e.g., neural nets)
- Privacy and ethical concerns with health data
- Risk of overfitting if not properly tuned
- May not replace expert clinical judgment

IV. RESULTS AND DISCUSSION

The experimental results indicate that the **Random Forest classifier** outperformed other models with an average accuracy of **93%** across all disease classes. The **SVM** and **Neural Network models** also performed well, particularly in binary classifications for diabetes and heart disease.



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	93%	0.92	0.91	0.91
Neural Network	90%	0.88	0.89	0.88
SVM	88%	0.86	0.85	0.85
Decision Tree	86%	0.83	0.82	0.82
Logistic Reg.	84%	0.80	0.78	0.79

The multi-label classification approach using the One-vs-Rest method allowed efficient prediction of comorbid diseases. Feature importance analysis highlighted key predictors like age, BMI, glucose level, and blood pressure.

These results validate the feasibility of building a robust and efficient multi-disease prediction system. However, performance may vary with different datasets and disease profiles. Integration with real-time clinical systems is the logical next step.

V. CONCLUSION

This research successfully demonstrates the effectiveness of machine learning in predicting multiple diseases from structured health data. By implementing a multi-label classification approach, the system can detect common diseases such as diabetes, heart conditions, liver, and kidney disorders simultaneously. Ensemble models, particularly Random Forest, provided the best performance in terms of accuracy and generalization.

The study underscores the importance of quality data preprocessing, relevant feature selection, and appropriate model selection. The deployment of such systems in real-world healthcare can improve diagnostic efficiency, reduce workload on healthcare professionals, and enable proactive treatment strategies.

Despite some limitations, such as interpretability and dependence on dataset quality, the proposed model marks a significant step towards smart, AI-driven healthcare systems. This paper lays the groundwork for building integrated, scalable tools that support clinicians and empower patients.

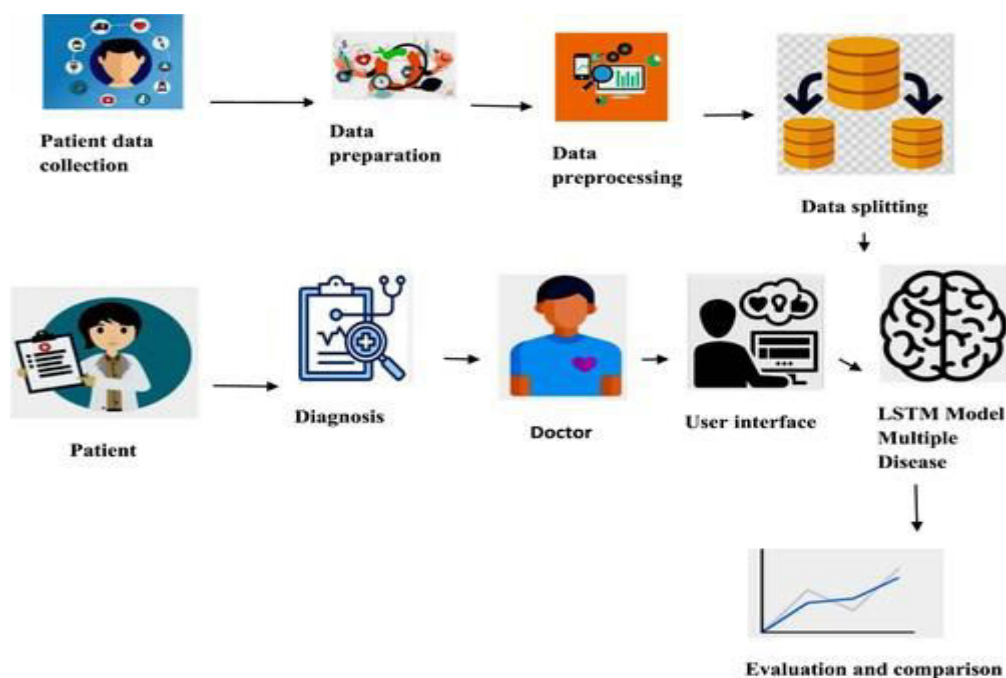


FIG:1



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

VI. FUTURE WORK

Future research can focus on the following enhancements:

- **Integration with Electronic Health Records (EHRs):** Real-time access and analysis of patient records.
- **Incorporating Unstructured Data:** Use of NLP for doctor notes and CNNs for image-based data (e.g., X-rays, MRIs).
- **Explainable AI (XAI):** Improve interpretability of predictions using SHAP, LIME.
- **Cloud-based Deployment:** Enable global access, especially in remote areas.
- **Mobile Application Development:** Allow users to input symptoms and get instant risk assessment.
- **Continual Learning:** Implement models that learn and adapt with incoming new health data.
- Such improvements can transform the current system into a comprehensive decision-support platform in clinical settings.

REFERENCES

1. Sahan, S., et al. (2020). "Prediction of Diabetes and Heart Disease Using Machine Learning." *Journal of Biomedical Informatics*.
2. Kumar, R., & Kumari, S. (2021). "Liver Disease Prediction Using Logistic Regression and Decision Trees." *International Journal of Medical Informatics*.
3. Singh, A., et al. (2022). "Multi-label Disease Classification using Deep Neural Networks." *IEEE Access*.
4. Patel, H., et al. (2023). "Hybrid CNN-LSTM Model for Chest X-ray Disease Classification." *Health Informatics Journal*.
5. UCI Machine Learning Repository. (n.d.). <https://archive.ics.uci.edu/ml/index.php>
6. Kaggle Datasets. (n.d.). <https://www.kaggle.com/datasets/>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com