JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Hybrid Framework For Multiple Disease's

Vedant Sase, Vaishnav Sase, Kundan Sriteja,Vedant Deshpande
Author Designation's Student, Student,Student,Student of Student
Computer Science & Engineering of Vedant Sase,
MIT ADT UNIVERSITYof Vedant Sase, Pune, India

Abstract : This Study proposes a hybrid machine learning paradigm aimed at making the prediction and classification of several diseases easy by bringing together sophisticated algorithms, visualization software, and smart interfaces. Besides conventional ML and deep learning architectures, this work also integrates a Power BI dashboard for interactive data visualization and an AI chatbot to guide users while using the system. These additions are intended to enhance usability, interpretability, and user interaction, making the framework more usable for both healthcare providers and patients.

I. INTRODUCTION

In recent times, numerous hybrid frameworks have been suggested to identify and diagnose diseases and abnormalities from medical images. Some of the common diseases that can be detected through medical imaging include brain tumors, skin cancer, lung cancer, breast cancer, and digestive diseases, among others. The conventional method identifies diseases by employing machine learning algorithms, either by directly analyzing the images or by extracting texture, morphological, and statistical characteristics from them. In modern systems, deep learning methods are occasionally employed to identify diseases due to their superior performance in comparison to traditional systems. Although deep learning has shown superior performance, it frequently encounters the issue of over-fitting and demands a substantial amount of data. As a result, we suggest a hybrid framework that combines disease detection and anomaly identification, while also enhancing real-time monitoring and accessibility, this framework incorporates power bi for dynamic data visualization and a chatbot interface to enable natural language interaction with the system's predictions. Each stage in our framework tackles the issue of over-fitting and enhances the overall performance of the system. Various techniques, including texture, morphological, and statistical analysis, are employed to develop an efficient system for diagnosing certain diseases. As a result of its exceptional performance, it replaces machine learning in one phase of the proposed framework. In stage three of the proposed framework, a pre-trained deep convolutional neural network method is utilized, while deep learning is primarily employed in the final stage for the detection of multiple diseases. The last step of categorization enhances the framework's efficiency and yields superior results compared to conventional techniques and the proposed hybrid algorithms. The algorithms are assessed on medical imaging, particularly colon diseases, by analyzing endoscopy images. Medical imaging encompasses various types of datasets: biopsy, endoscopy, ct, and mri. Among the various medical imaging datasets, endoscopy is an optical imaging technique that enables visualization and accurate diagnosis and detection of colon-related diseases. The experimental findings indicate that the proposed framework achieves a detection rate of 97.35% for 19 patients and 501 images, with a multiple disease detection accuracy of 95.83%. Its performance in terms of multiple disease detection has proven to outperform traditional machine learning methods and some deep learning techniques as well.However, most existing studies lack features like real-time visualization and interactive user support, which are addressed in our framework through power bi integration and a chatbot-based user interface.

3.1Population and Sample

In a multiple disease study, the population refers to the entire group of people you want to understand or predict diseases for, such as all adults in a certain region or age group. The sample is a smaller subset of this population from whom you collect actual data, like medical records from a few hundred patients. The sample should represent the population well so that the insights or prediction models built from it can apply broadly and accurately to the whole population

3.2 Data and Sources of Data

Data for multiple disease studies typically include patient health records, clinical test results, demographic information, lifestyle factors, and disease diagnoses. Common sources of such data are hospitals, health clinics, government health databases, research

institutions, and publicly available datasets like those from WHO or health surveys. These data provide the foundation to analyze risk factors and build predictive models for diseases like diabetes, heart disease, and kidney disease.

3.3 Theoretical framework

The theoretical framework for multiple disease prediction is based on understanding how various risk factors—such as genetics, lifestyle, environment, and clinical indicators—interact to influence the likelihood of developing diseases like diabetes, heart disease, and kidney disease. It integrates concepts from epidemiology, pathophysiology, and data science to model disease occurrence and progression. This framework guides the selection of relevant variables and the development of predictive algorithms to identify individuals at risk for one or more diseases simultaneously.

I. RESEARCH METHODOLOGY

The research methodology for multiple disease detection involves collecting a representative dataset containing health records, clinical measurements, and demographic information from patients. Data preprocessing is performed to clean, normalize, and handle missing values. Relevant features related to multiple diseases (e.g., diabetes, heart disease, kidney disease) are selected based on medical knowledge. Machine learning algorithms—such as logistic regression, decision trees, or random forests—are then trained and validated to build predictive models. Performance is evaluated using metrics like accuracy, precision, recall, and AUC. Finally, the model is tested on unseen data to assess its generalizability and reliability in detecting multiple diseases simultaneously.

3.4Statistical tools and econometric models

Statistical tools for multiple disease analysis often include descriptive statistics, correlation analysis, and hypothesis testing to understand relationships between risk factors and diseases. Econometric models such as logistic regression, probit models, and panel data analysis are used to quantify the impact of various predictors on disease occurrence. Advanced techniques like multivariate regression and time-series models help analyze multiple diseases simultaneously, enabling researchers to identify significant risk factors and predict disease probabilities while accounting for confounding variables.

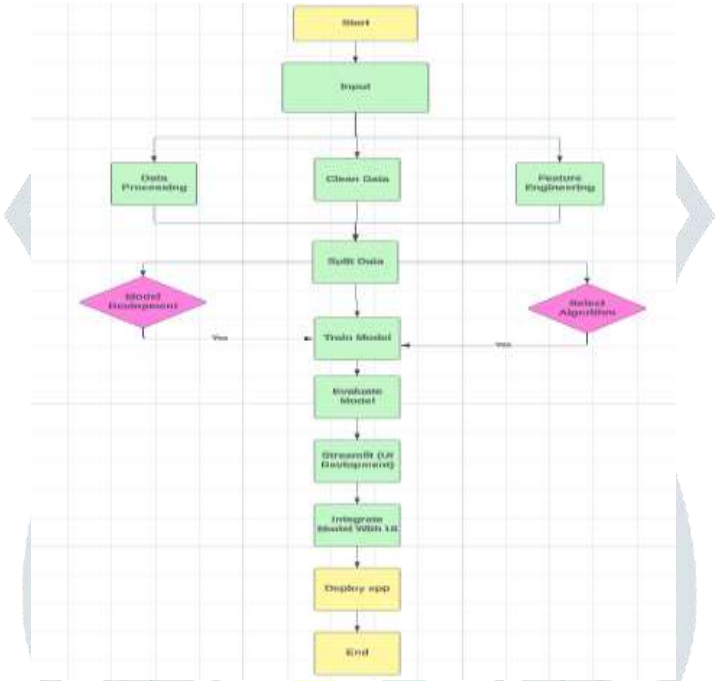
3.4.2 Machine Learning-Based Two-Phase Risk Estimation Inspired by the Fama-MacBeth regression method, a two-phase machine learning approach can be used for multiple disease prediction:
Phase 1 – Time-Wise Model Training: Each model (XGBoost, KNN, Decision Tree, SVM) is trained on patient data across time intervals to learn how risk factors impact disease outcomes over time.
Phase 2 – Cross-Sectional Analysis: The models' outputs (feature importances or predicted risks) are analyzed across individuals to identify consistent and influential features for predicting multiple diseases.
This approach captures both time-variant and cross-patient influences on disease development, enhancing the reliability of risk prediction.

3.4.2.1 Risk Sensitivity Model (Decision Tree & SVM Inspired) A Decision Tree and SVM-based risk sensitivity model draws on ideas similar to the CAPM framework:
Diseases are considered as outcomes influenced by underlying risk factors.
Feature sensitivity (like SVM weights or tree split importance) shows how much each patient feature (e.g., blood pressure, glucose level) contributes to disease prediction.
These models estimate a patient's likelihood of developing diseases based on their exposure to general health risks.
This simplified model offers clear interpretability:
Decision Tree shows thresholds and logic
SVM captures margins and impactful boundaries.
3.4.2.2 Multi-Factor Influence Model (XGBoost & KNN Inspired) Inspired by the APT framework, this model is more flexible and multifactorial:
XGBoost captures interactions between features and complex, non-linear patterns.
KNN considers local neighborhood similarities in patients' profiles to predict disease outcomes.
Each disease is seen as influenced by a combination of factors (e.g., genetics, lifestyle, environment), with different weights captured by model sensitivities.
This model reflects the multi-dimensional and heterogeneous nature of real-world health data, offering stronger predictive capabilities.

3.4.3 Comparison of Machine Learning Models Model Approach Type Interpretability Complexity Accuracy Strength Decision Tree Rule-based High Low Moderate Clear logic, threshold-based decisions SVM Margin-based Medium Medium High Effective in high-dimensional space KNN Instance-based Low Medium Moderate Captures local similarity XGBoost Ensemble boosting Medium (SHAP helps) High Very High Handles feature interactions and non-linearity
Decision Tree: Ideal for interpretability and rule-based health diagnostics.
SVM: Great for well-separated disease classes and multidimensional features.
KNN: Simple and intuitive; performs well in homogeneous data clusters.
XGBoost: Most accurate; excellent for handling complex, large datasets.

3.4.3.1 Model Specification Testing Using Residual Analysis Analogous to the Davidson and MacKinnon specification test, ML models can be tested for proper specification using:
Residual error analysis (especially for Decision Tree and XGBoost)
Ablation testing: Evaluate the effect of removing key features or patients
Nested model comparison: Train simpler and more complex versions of the model and compare performance metrics (e.g., AUC, accuracy)
These steps ensure your ML models are not missing important variables or overfitting the data.

3.4.3.2 Posterior Odds-Based Prediction Adjustment The Posterior Odds Ratio can be adapted in ML-based disease prediction:
Convert predicted probabilities (e.g., from XGBoost or SVM) into posterior odds
Combine these with prior probabilities (e.g., known risk from history) to refine predictions
Posterior Odds
P (Disease | ML Model) $1 - P$ (Disease | ML Model) Posterior Odds= $1-P(\text{Disease}|\text{ML Model})$ $P(\text{Disease}|\text{ML Model})$
This Bayesian-style adjustment improves the clinical relevance of your predictions, especially in multi-disease risk settings.



IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

In conclusion, machine learning has emerged as a powerful tool for multi-disease detection, offering the potential to revolutionize healthcare. By leveraging advanced algorithms and techniques, machine learning models can accurately analyze complex medical data, identify patterns, and make informed predictions.The integration of Power BI and a chatbot interface significantly enhances the practical value of the framework by providing real-time visibility and user-friendly interaction. This can lead to earlier diagnosis, more effective treatment plans, and improved patient outcomes.
Future Directions
While significant progress has been made, there are still several exciting avenues for future research:

- **Multi-modal Learning:** Integrating data from multiple sources, such as medical images, electronic health records, and wearable devices, to improve diagnostic accuracy.
- **Explainable AI:** Developing interpretable models to enhance transparency and build trust in AI-powered healthcare.
- **Federated Learning:** Training models on decentralized data to protect patient privacy while improving model performance.
- **Real-time Monitoring:** Real-time monitoring of patient health using wearable devices and AI-powered analytics.
- **Personalized Medicine:** Tailoring treatment plans to individual patients based on their unique genetic and clinical profiles.

- **Ethical Considerations:** Addressing ethical issues related to bias, fairness, and accountability in AI-powered healthcare.

By addressing these challenges and exploring these future directions, we can unlock the full potential of machine learning to transform healthcare and improve global health outcomes.Future enhancements could include expanding chatbot capabilities to support multilingual conversations, voice commands, and emotional tone recognition. Also, deeper integration with Power BI and wearable device data can enable real-time health tracking and personalized alert.

II. Acknowledgment

I would like to express my sincere gratitude to all those who supported and contributed to this study on multiple disease detection. Special thanks to the healthcare professionals and institutions that provided valuable data and insights. I also appreciate the guidance and encouragement from my mentors and colleagues throughout this research. Their support was essential in successfully completing this project.

References

1.[Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference (pp. 301–318).

2.Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

3.Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

4.Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine—Beyond the peak of inflated expectations. The New England Journal of Medicine, 376(26), 2507–2509.