iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: VI Month of publication: June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83868

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 14 Issue VI June 2026- Available at www.ijraset.com

A Review of Machine Learning Techniques for Symptom-Based Disease Prediction

Prafulla B.Bhand, Aniruddha N.Rane, Ayush K.Khairnar, Soham S.Phatak. Punam A.Shinde, Dr.Emmanuel M.
Information Technology Pune Institute of Computer Technology Pune, India

Abstract: Symptom-based disease prediction is crucial in healthcare. It addresses the need for early diagnoses and effective treatment plans. Traditional diagnostic methods rely on human expertise, which can beslow and error-prone. Theuse of Machine Learning and Deep Learning techniques has greatly improved automated disease prediction systems. These systems provide faster and more accurate predictions. Still, there are ongoing challenges. These include issues with incomplete symptom data, improving prediction accuracy for rare conditions, and ensuring that models are easy to understand. This survey paper reviews current research in symptom-based disease prediction models. It focuses on Machine Learning algorithms, hybrid approaches, and data preprocessing methods. It also discusses recent advances in neural networks and Natural Language Processing for symptom analysis. Additionally, the paper proposes a framework that uses multi-modal data fusion to provide more reliable predictions. The review highlights technological innovations, challenges, and potential future applications in healthcare.

Index Terms: Machine Learning, Deep Learning, Symptom-Based Prediction, Explainable AI, Data Imputation, Rare Dis-eases, Multi-modal Data Fusion

I. INTRODUCTION

In healthcare, early and accurate disease diagnosis is vital “for effective treatment and better patient outcomes. Traditional diagnostic methods often depend on human expertise, which can be slow and inconsistent among different doctors. As medical conditions become more complex, manual diagnosis becomes even tougher, especially when symptoms overlap among various illnesses. To overcome these challenges, auto-mated disease prediction systems that use Machine Learning andArtificialIntelligence havegainedtraction. Thesesystems analyze patient symptoms and generate predictions based on data-driven methods. However, some key issues still need solutions:

- IncompleteandNoisySymptomData:Medicaldatasets often have missing or inconsistent symptom information. This can harm the accuracy of prediction models.
- Managing Rare Conditions: Many prediction systems havedifficultyaccuratelypredictingrareconditionsdueto insufficient data and symptom overlap with more common conditions.
- InterpretabilityofPredictions:Inhealthcare,automated systems must provide not only accurate predictions but also understandable reasoning for the forecasts. This is essential for gaining the trust of medical professionals.

II. LITERATURE SURVEY

A. Machine Learning for Symptom-Based Disease Prediction

Machine Learning is widely used for symptom-based dis-ease prediction to enhance diagnostic accuracy and speed. TraditionalMachineLearningmethods,likedecisiontreesand Random Forests, classify diseases using symptom data. For instance, a review highlighted effective algorithms such as Support Vector Machines and k-Nearest Neighbours [4], [5]. However,traditionalmethodsoftenstrugglewithunstabledata and complex conditions involving multiple symptoms.

Deep Learning techniques have led to more powerful mod-els for disease prediction. Deep Neural Networks and Con-volutional Neural Networks have made significant strides in analyzing multi-dimensional symptom data. Research showed that CNNs could identify complex patterns in symptom clus-ters that traditional methods missed [7], [8]. Despite these improvements, Deep Learning models typically need large datasets and can overfit when used with small or imbalanced data.

B. Handling Incomplete and Noisy Symptom Data

Medical datasets often have missing values or noisy data, significantlyhurtingdiseasepredictionmodel performance.

A proposed solution suggested using imputation techniques along with Machine Learning models to manage missing clinicaldata[9].

Itwasfoundthatcombiningk-meanscluster-ing with imputation improved prediction accuracy. A hybrid method that combined imputation and anomaly detection was alsointroduced.Thissystemusedautoencoderstorecon-struct missing symptom data and identify outliers that could affectaccuracy.Whilethisapproachimprovesdataquality, challengesstillexist whenscalingupto largedatasets from multiple sources.

C. Rare Disease Prediction

Predictingrareconditionsischallengingduetoalackof training data. Traditional Machine Learning models often misclassify rare diseases because they are underrepresented in datasets.Onesystemproposedanensemblelearningapproach that balanced the dataset by oversampling rare disease cases [2]. While effective in improving detection, it still struggles with generalization to new, unseen cases. Recently, transfer learning techniques have been introduced to tackle the data scarcity problem in rare disease prediction. One study used transferlearningwithpre-trainedmodelsfromlargemedi-cal datasets to improve rare disease classification. However, this method's effectiveness depends on accessing large, well-annotated medical datasets.

D. Explainability and Interpretability in Disease Prediction Models

Explainabilityiscrucialinhealthcare,asprofessionalsmust trust and understand the predictions from automated systems. Black-box models like deep neural networks are often hard to interpret, making them less favorable for clinical use. One prediction framework integrated decision trees with neural networks to enhance interpretability. This model provided clearer insights but sacrificed some prediction accuracy.

In recent years, Explainable AI techniques like LIME and SHAPhavegainedpopularityformakingcomplexmod-elseasiertounderstand[6],[10].SHAPwasappliedtoa symptom-basedpredictionmodel,allowingclinicianstosee the contribution of each symptom to the final prediction [1]. However,thisapproach,whiletransparent,provedcostlyin computational resources for real-time use in clinical settings. Moderndiseasepredictionsystemsincreasinglycombine datafromvarious sourceswiths ymptomdata. Thisincludes patienthistory,geneticinformation,andimagingdata.Inte-gratingthesedifferenttypesofdataposessignificantchal-lenges. Oneframe workdevelopedamulti-modallearningsys-temthatcombinedtext-basedsymptomdatawithimagingdata toimprovepredictions. ThissystemusedConvolutionalNeural Networks for image analysis and Recurrent Neural Networks for text data. It showed improvements in prediction accuracy, especiallyforcomplexconditions,butrequiredsignificant computingres ources. Another teamproposedaframeworkthat integratedElectronicHealthRecordswithsymptomdatato enhancepredictionmodels. The irresults confirmed headvan-tagesofmulti-modaldatafusionforimprovingdiagnostics, althoughitfacedchallengesinprocessinghigh-dimensional data in real time.

III. METHODOLOGIES STUDIED

This section reviews established techniques and their lim-itations, laying the groundwork for the paper's proposed framework.

A. Traditional Machine Learning Techniques

TraditionalMachineLearningmethodswereusedtoclassify diseases based on symptom data.

- Methods Used: These included standard algorithms like decision trees, Support Vector Machines, and k-Nearest Neighbours.
- Limitations:Thesemethodsfacedchallenges,especially with unbalanced datasets, where some diseases are much rarer than others, and with complex diseases that have overlapping symptoms across various conditions.

B. Data Imputation for Incomplete Symptom Data

To address the common problem of poor data quality in medicalrecords,weexploredtechniquesformanagingmissing or noisy data.

- Techniques Used: We applied methods such as k-means clustering and autoencoders for data imputation, which fills in missing values, and for anomaly detection, which identifies or corrects noisy outliers.
- Limitations: Although these imputation methods im-proved the quality of individual datasets, their effec-tiveness drops when scaling up to handle large, diverse datasets that come from many different sources and institutions.
- Noise Reduction: The framework also uses noise reduc-tion techniques to clean the data, making the final input more dependable for predictions.

IV. CONCLUSION

This survey paper has systematically reviewed the current researchonautomatedsymptom-baseddiseaseprediction.It highlights significant improvements in machine learning algorithms,datapreprocessingtechniques,andeffortstomake models more understandable. The system proposed in this review directly addresses critical challenges noted in the literature, including:

- Effectivelymanagingincompletesymptomdata.
- Increasingtheaccuracyofpredictionsforrarediseases.
- Ensuringtheclarityandtransparencyofpredictions.

By using data-driven methods, the proposed framework greatly improves the accuracy and reliability of automated disease prediction, offering valuable support to healthcare workers. While some challenges remain, such as integrating multimodal data and optimizing real-time prediction perfor-mance, the system shows promise for addressing the growing complexities of automated clinical diagnosis.

V. FUTURE SCOPE

Looking forward, future work should focus on several key areas to expand the system's usefulness and strength:

- Data Expansion: We could enhance capabilities to man-age larger and more varied datasets, including patient histories, genetic data, and real-time health monitoring information.
- Adaptability and Explainable AI (XAI): Research should focus on improving the model's capacity to adapt to new health conditions and on integrating better ex-plainable AI techniques to boost its clinical value.
- Scalability and Security:Incorporatingcloud-basedand federated learning approaches can significantly improve scalability across different healthcare settings while en-suring patient data privacy and maintaining model accu-racy.

REFERENCES

[1] P. Guleria, P. N. Srinivasu, S. Ahmed, N. Almusallam, and F. K.Alarfaj, "XAI framework for cardiovascular disease prediction usingclassification techniques," Electronics, vol. 11, no. 24, Art. no. 4086,2022.

[2] M. Wojtara, E. Rana, T. Rahman, P. Khanna, and H. Singh, "Artificialintelligence in rare disease diagnosis and treatment," Clin. Transl. Sci.,vol. 16, no. 11, pp. 2106–2111, 2023.

[3] M. M. Ahsan and Z. Siddique, "Machine learning-based heart diseasediagnosis: A systematic literature review," Artif. Intell. Med., vol. 128,Art. no. 102289, 2022.

[4] N. G. Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of artificialintelligence techniques in disease diagnosis and prediction," Discov.Artif. Intell., vol. 3, no. 1, Art. no. 5, 2023.

[5] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-learning-baseddisease diagnosis: A comprehensive review," Healthcare (Basel), vol.10, no. 3, Art. no. 541, Mar. 2022.

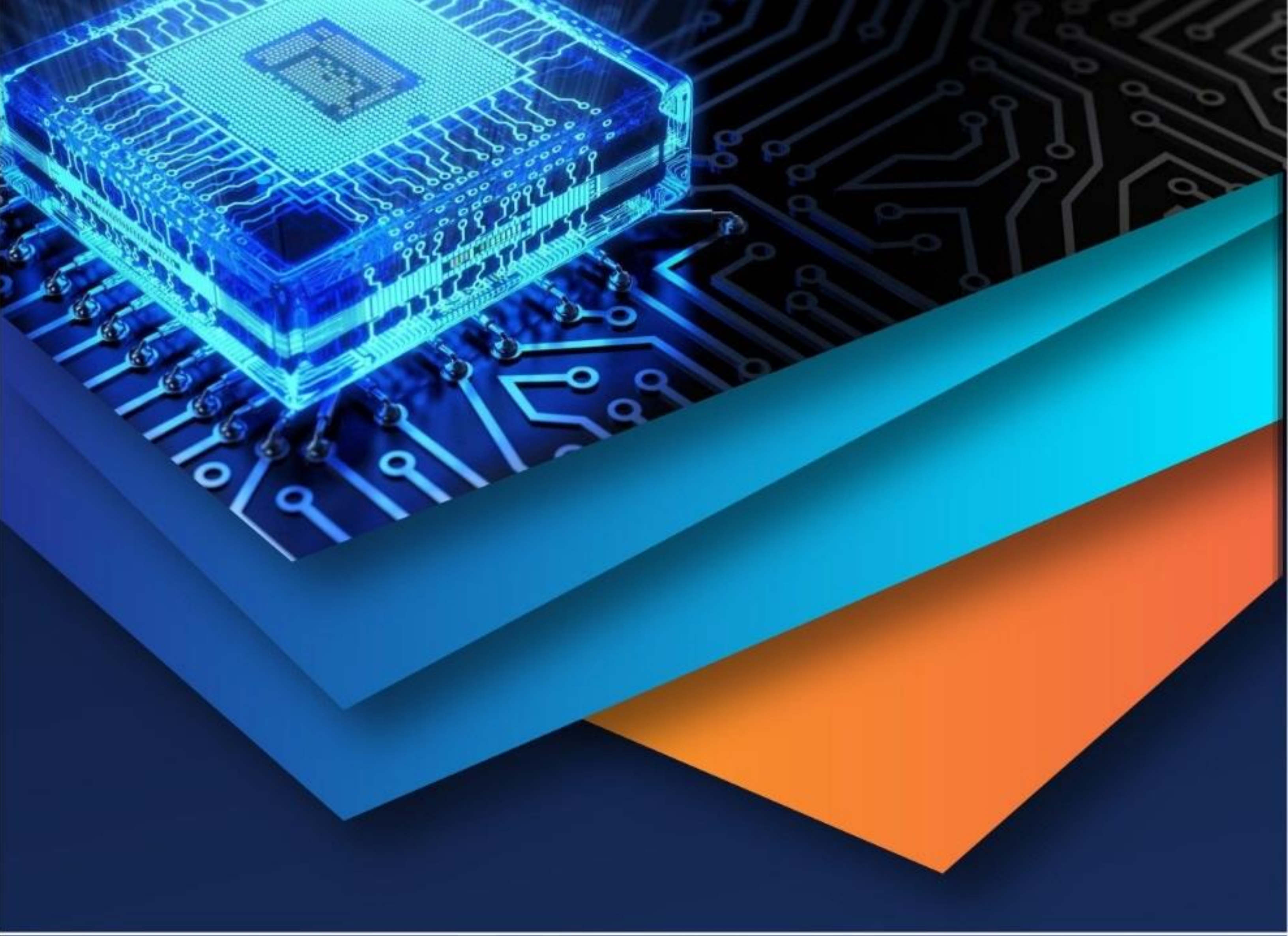
[6] G.Stiglic,P.Kocbek,N.Fijacko,M.Zitnik,K.Verbert,andL.Cilar, "Interpretability of machine learning-based prediction models inhealthcare,"WileyInterdiscip.Rev.DataMiningKnowl.Discovery,vol.10, no. 5, Art. no. e1379, 2020.

[7] G.Litjensetal.,"Asurveyondeeplearninginmedicalimageanalysis,"Med. Image Anal., vol. 42, pp. 60–88, 2017.

[8] T. Bhanuteja, K. V. N. Kumar, K. S. Poornachand, C. Ashish, and P.Anudeep, "Symptoms-based multiple disease prediction model usingmachine learning approach," Int. J. Innov. Technol. Explor. Eng. (IJI-TEE), 2021.

[9] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, "Missing datain clinical research: a tutorial on multiple imputation," Can. J. Cardiol.,vol. 37, no. 9, pp. 1322–1331, 2021.

[10] R. Alkhanbouli, H. M. A. Almadhaani, F. Alhosani, and M. C. E.Simsekler, "The role of explainable artificial intelligence in diseaseprediction:asystematicliteraturereviewandfutureresearchdirections,"BMC Med. Inform. Decis. Making, vol. 25, no. 1, Art. no. 110, 2025.



10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)