www.ijarp.com ISSN 2456-9992 Page: 01-09

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 06

HEALTH RISK PREDICTION USING BIG HEALTH DATA

***Dr. Shrinivas Sonkar, Prachi Auti, Saloni Bhingare, Shweta Date, Prajkta Boraste**

Department of Computer Engineering Amrutvahini College of Engineering, Sangamner, India.

Article Received: 13 May 2026, Article Revised: 3 June 2026, Published on: 23 June 2026

***Corresponding Author: Dr. Shrinivas Sonkar**

Department of Computer Engineering Amrutvahini College of Engineering, Sangamner, India.

Doi: https://doi-doi.org/101555/ijarp.2499

ABSTRACT

The rapid digitization of healthcare has led to the creation of extensive and diverse medical datasets generated from electronic health records, laboratory information systems, diagnostic imaging platforms, and wearable health technologies. Although such data holds substantial potential toenhance clinical decision-making, conventional diagnostic practices continue to rely largely on manual evaluation and are inefficient in handling complex, high-dimensional medical information. Consequently, the early detection of chronic diseases such as lung cancer and Chronic Kidney Disease (CKD) remains a significant global challenge. To address this issue, this study presents a machine learning–based health risk prediction system that utilizes Big Health Data for accurate disease risk assessment. The proposed framework analyzes clinical, demographic, and lifestyle features using Logistic Regression, Random Forest, and XGBoost algo-rithms. By automating feature extraction and pattern recognition, the system enhances predictive accuracy, reduces diagnostic inconsistency, and supports timely clinical intervention. Exper-imental findings indicate that ensemble and gradient boosting techniques achieve superior predictive performance compared to traditional diagnostic approaches.

INDEX TERMS: Big Health Data, Machine Learning Models, Health Risk Prediction, Chronic Kidney Disease, Lung Cancer Detection, Random Forest, XGBoost, Clinical Decision Support Systems.

I. INTRODUCTION

Advancesindigitaltechnologyhavesignificantlyalteredthe structure and functioning of modern healthcare systems [1], [2], [5]. The integration of Electronic Health Records (EHRs),

telehealth services, cloud-enabled diagnostic tools, and wear-able monitoring devices has made it possible to collect med-ical information continuously and at scale [8], [21]. This extensive and heterogeneous collection of healthcare informa-tion—commonlyreferredtoasBigHealthData—includeslab-oratory measurements, medical images, demographic profiles, lifestyle-related attributes, and long-term patient records [8]–[10]. The growing availability of such data creates new pos-sibilities for intelligent, data-centric healthcare analysis and decision support [16], [21].

Despite this data abundance, diagnostic processes in many healthcare settings remain largely manual and clinician-dependent [3], [4]. Conventional diagnostic approaches re-quire substantial time and effort, are vulnerable to subjective judgment, and lack the capability to efficiently analyze com-plex, high-dimensional datasets [8], [21]. These constraints significantlylimittheeffectivenessofearlydiseasedetec-tion, especially for conditions that develop gradually and exhibitminimalsymptomsintheirinitialstages[1],[6].

As a consequence, many serious illnesses are identified only after reaching advancedstages, when treatment becomes more expensive and less effective [2], [6].

Chronic Kidney Disease (CKD) and lung cancer are promi-nent examples of this challenge [1], [10], [11]. CKD often progresses unnoticed during its early phases and is frequently diagnosed only after irreversible kidney damage has occurred, resulting in End-Stage Renal Disease (ESRD) and long-term economic and clinical burden [2]–[4]. Likewise, lung cancer continues to account for a high proportion of cancer-related deaths worldwide, primarily due to late diagnosis and limited earlyscreeningeffectiveness[10]–[12].Theserealitiesexpose a fundamental limitation in current healthcare systems: the absence of reliable, automated, and scalable mechanisms for early disease risk assessment [8], [21].

Machine Learning (ML) has gained increasing attention asa viable solution to overcome these limitations [8]–[10]. ML techniquesenableautomatedprocessingoflarge-scalemedical datasets and can uncover complex, non-linear relationships among clinical variables [15], [16], [21]. Such capabilities support critical healthcare tasks including disease prediction, patient risk stratification, and clinical decision assistance [8], [10], [16]. Although prior research has investigated the use of individual ML models for medical prediction, many existing studiesfocusonasingledisease,relyonlimiteddatasets, or fail to provide comparative evaluation across multiple algorithms [9], [10], [12]. Additionally, there remains a need forintegratedpredictiveframeworksthateffectivelyutilizeBig Health Data while maintaining accuracy, interpretability, and scalability [16], [21].

Inresponsetothesechallenges,thispaperaimstodevelopa machine learning–based health risk

prediction framework for early detection of CKD and lung-related disorders using Big Health Data [8]–[10]. The proposed system employs Logistic Regression, Random Forest, and XGBoost models to analyzea comprehensive set of clinical, demographic, and lifestyle features [9], [10], [12]. By combining linear models with ensemble and gradient boosting techniques, the framework seeks to enhance predictive performance, reduce diagnostic inconsistency, and improve robustness when handling high-dimensional healthcare data [15], [16].

Thekeycontributionofthisstudyisthedevelopmentof a unified and comparative predictive framework for chronic disease risk assessment [9], [10], [16]. The results demon-stratethatensembleandboosting-basedmodelsoffersupe- rior performance over conventional diagnostic methods while supporting consistent and data-driven clinical decisions [15], [16],[21].Overall,thisworkcontributestotheadvancementof intelligenthealthcaresystemsbyprovidingascalableapproach for early disease detection, reduced diagnostic burden, and improved patient outcomes [8], [21].

II. METHODOLOGY

A. DatasetDescription

The datasets used in this work were collected from publicly available Kaggle sources and are intended for healthcare risk prediction [5], [8], [10]. Two datasets were employed: one related to chronic kidney disease and another focused on lung disease [1], [10].

The kidney disease dataset consists of 50,001 records withatotalof21attributes,including20medicalfeaturesand one target variable [5], [8]. The features include demographic information, laboratory test values, and clinical conditions such as blood pressure, biochemical measurements, blood-related parameters, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia [5], [21]. The target label specifies whether the patient is affected by chronic kidney disease [1], [5].

The lung disease dataset contains 50,001 samples with 9 columns in total, where 8 columns represent input featuresand one column represents the output label [10], [11]. These featuresdescribelifestylefactors,environmentalexposure,and respiratory symptoms, including age, body mass index, pollu-tion level, air quality index, smoking habits, cough, shortness of breath, chest pain, fatigue, and family history [11], [12]. The target variable indicates the presence or absence of lung disease [10], [12].

B. Data Preprocessing

Data preprocessing was carried out to improve the qualityof the datasets and prepare the data for machine learning modeltraining[8],[21].Initially,thedatasetswerecheckedfor missing values, duplicate records, and inconsistent entries [5], [8]. Duplicate and irrelevant records were removed to avoid bias in the prediction results [5].

Missingvaluespresentinnumericalattributeswerehandled usingmeanormedianimputationtechniques,whilecategorical attributes were filled using the most frequent value [8], [21]. This ensured that valuable data was not lost during prepro-cessing [8].

Categorical variables were converted into numerical form using label encoding methods [8]. Numerical features were normalized using Min–Max normalization to scale all values between0and1,whichhelpsimprovemodelconvergenceand performance [21].

These preprocessing steps ensured that the data was clean, consistent, and suitable for effective machine learning–based disease prediction [8], [21].

C. Feature Selection and Analysis

Feature selection was performed to identify the most rele-vant medical attributes contributing to disease prediction and to reduce model complexity [8], [9], [21]. Statistical correla-tion analysis and model-based feature importance techniques were applied [8], [9].

For the kidney disease dataset, laboratory indicators suchas serum creatinine, blood urea, hemoglobin, blood pressure, diabetes mellitus, and hypertension were identified as highly influentialfeatures[5],[8].Theseparametersareclinicallyas-sociated with kidney dysfunction and disease progression [1], [2].

For the lung disease dataset, lifestyle and environmental factors including smoking habits, air quality index (AQI), pollution exposure, shortness of breath, chest pain, and family history were found to have significant impact on prediction outcomes [10]–[12]. Selecting these features improved model interpretability and predictive accuracy [11], [16].

D. Proposed System Architecture

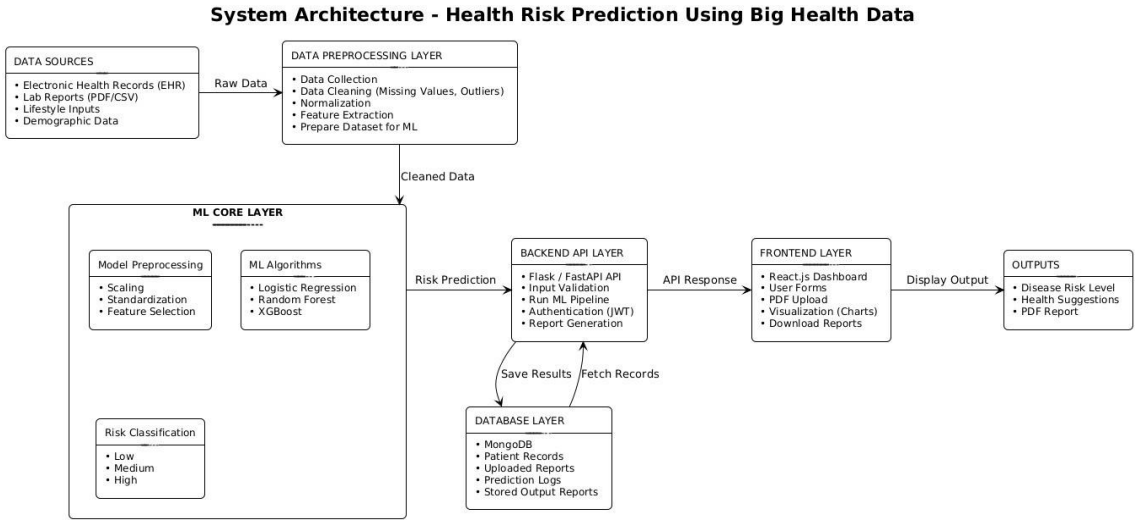

Fig.1.Overall Architecture of the Health Risk Prediction System.

Theproposedsystemfollowsalayeredarchitecture(Fig-ure 1) for efficient health risk prediction. Data from EHRs,lab reports, lifestyle inputs, and demographics is first pre-processed to handle missing values, normalize features, and select relevant attributes [8], [21]. The ML core layer applies Logistic Regression, Random Forest, and XGBoost for risk prediction and classifies patients into Low, Medium, or High risk [9], [15], [16]. The backend API layer handles input validation, runs the ML pipeline, and generates reports [16]. Patient records and prediction logs are stored in the database layer, while the frontend layer provides an interface for data input,visualization,andreportdownload[10],[16].Thismod-ular design ensures scalability, accuracy, and timely clinical decision support.

III. MACHINELEARNINGMODELS

A. LogisticRegression

LogisticRegressionisalinearclassificationalgorithmcom-monly used in medical prediction tasks [9], [22]. It estimates the probability of disease occurrence using a sigmoid function and serves as a baseline model due to its simplicity and interpretability [9], [21].

B. Random Forest

Random Forest is an ensemble learning method that com-bines multiple decision trees to improve classification perfor-mance [16], [23]. It reduces overfitting and handles nonlinear relationships effectively, making it suitable for healthcare datasets with mixed features [16], [21].

C. XGBoost

XGBoostisagradientboostingalgorithmknownforitshigh accuracy and scalability [15], [16]. It

applies regularization techniquesandefficienttreeconstruction,allowingittohandle largeandcomplexhealthcaredatasets[15].Duetoitssuperior performance, XGBoost was selected as the final model [16], [21].

IV. Experimental Results and Performance Analysis

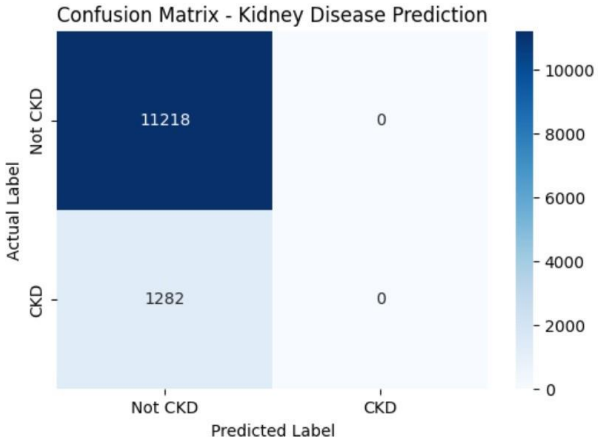

Fig.2.ConfusionMatrixofXGBoostModelforKidneyDisease.

A. Experimental Setup

All experiments were conducted using Python in a Jupyter Notebook environment [9], [10]. Libraries such as NumPy, Pandas, Scikit-learn, and XGBoost were used for implemen-tation [9], [15], [16]. The datasets were divided into training and testing sets using an 80:20 split [8]. Model performance was evaluated using accuracy, precision, recall, and confusion matrix analysis [9], [16].

B. Performance Comparison

TABLEI PerformanceComparisonforKidneyDiseaseDataset.

Model	Accuracy	Precision	Recall
LogisticRegression	0.88	0.89	1.00
RandomForest	0.8594	0.90	1.00
XGBoost	**0.8974**	0.90	1.00

TABLEII PerformanceComparisonforLungDiseaseDataset.

Model	Accuracy	Precision	Recall
LogisticRegression	0.8041	0.79	0.80
RandomForest	0.7943	0.79	0.79
XGBoost	**0.8226**	0.81	0.82

XGBoost achieves the highest accuracy on both Kidney (0.8974) and Lung (0.8226) disease datasets, outperforming LogisticRegressionandRandomForest[15],[16].Allmodels show strong recall, but XGBoost provides the best overall performance for these classification

tasks [16], [21].

C. ConfusionMatrixAnalysis

The confusion matrices illustrate the classification performanceoftheXGBoostmodelforbothkidneyandlungdisease prediction[15],[16].Forthekidneydiseasedataset,themodel correctlyclassified11,218instancesasnon-CKD,while1,282 CKD cases were misclassified as non-CKD, indicating a class imbalance [5], [8]. For the lung disease dataset, the model demonstratedbalancedperformancebycorrectlyidentifying

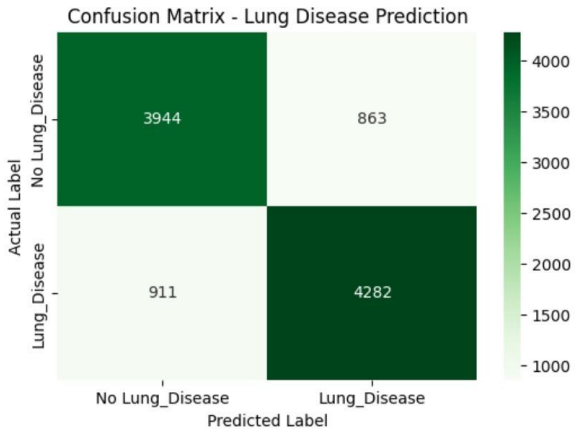

Fig.3.ConfusionMatrixofXGBoostModelforLungDisease.

3,944non-diseasecasesand4,282lungdiseasecases,with863 false positives and 911 false negatives, reflecting reasonable sensitivity and specificity [10], [12], [16].

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. Shrinivas Sonkar, Department of Computer Engineering, Amrutvahini College of Engineering, for his constant guidance,encouragement,andvaluablesuggestionsthroughoutthe courseofthisresearchwork.Hisexpertiseinmachinelearning anddata-drivenhealthcaresystemssignificantlycontributedto shaping the direction and technical depth of this study. His timely advice, constructive feedback, and continuous motiva-tion helped the authors overcome various challenges during data preprocessing, feature selection, model development, and performance evaluation.

The authors also extend their appreciation to the Depart-ment of Computer Engineering, Amrutvahini College of En-gineering,forprovidingthenecessaryinfrastructure,academic resources, and a supportive research environment required to successfully complete this

work.

Furthermore, the authors acknowledge the researchers and organizations that made publicly available healthcare datasets through platforms such as Kaggle, which enabled comprehensive experimentation and evaluation of the proposed machine learning models.

The authors are grateful to their peers and colleagues for theirhelpfuldiscussionsandcooperation,whichcontributedto refiningthemethodologyandanalysis.Finally,heartfeltthanks are extended to their families for their unwavering support, patience, and encouragement throughout the duration of this research.

REFERENCE

1. M.Khaliliaetal.,"Predictingdiseaseriskusingmachinelearningonhealthcaredata,"*BMCMedicalInformaticsandDecisionMaking*,2011.
2. B.TabaeiandW.Herman,"Diabetespredictionmodels:Areview,"*Diabetes Care*, vol. 27, no. 1, pp. 1–13, 2004.
3. T.-S.Leeetal.,"Decisiontreemodelsforhealthriskprediction,"*Healthcare Informatics Research*, 2010.
4. I.Lemieuxetal.,"Triglyceridesaskeyindicatorsinmetabolicdisor-ders," *Circulation Research*, 2001.
5. A.Worachartcheewanetal.,"Decisiontreeanalysisformetabolicsyndrome prediction," *BMC Medical Research Methodology*, 2019.
6. N. R. Hill, S. T. Fatoba, J. L. Oke, J. A. Hirst, et al., "Global prevalenceofchronickidneydisease:Asystematicreviewandmeta-analysis,"*PLoSONE*, vol. 11, no. 7, p. e0158765, 2016.
7. W.-C. Yang and S.-J. Hwang, "Incidence, prevalence and mortalitytrends of dialysis end-stage renal disease in Taiwan," *Nephrology Dial-ysis Transplantation*, vol. 23, no. 12, pp. 3977–3982, 2008.
8. C.-H. Tang, H.-H. Chen, M.-J. Wu, et al., "Out-of-pocket costs andproductivitylossesinhaemodialysisandperitonealdialysis,"*BMJOpen*,vol. 9, no. 3, p. e023062, 2019.
9. H.-H.Wu,C.-F.Kuo,I.-J.Li,etal.,"Familyaggregationandheritabilityof end-stage renal disease in Taiwan," *American Journal of KidneyDiseases*, vol. 70, no. 5, pp. 619–626, 2017.
10. C.-C. Shih, C.-J. Lu, G.-D. Chen, and C.-C. Chang, "Risk prediction forearly chronic kidney disease," *International Journal of EnvironmentalResearch and Public Health*, vol. 17, no. 14, p. 4973, 2020.

11. C.-P. Wen, T.-Y. D. Cheng, M.-K. Tsai, et al., "All-cause mortalityattributable to chronic kidney disease," *The Lancet*, vol. 371, no. 9631, pp.2173–2182,2008.
12. R. T. Gansevoort, R. Correa-Rotter, B. R. Hemmelgarn, et al., "Chronickidney disease and cardiovascular risk," *The Lancet*, vol. 382, no. 9889, pp.339–352,2013.
13. Y.-L. Chiu, M.-J. Jhou, T.-S. Lee, et al., "Health data-driven machinelearning algorithms applied to risk indicators assessment for chronickidney disease," *Risk Management and Healthcare Policy*, pp. 4401–4412, 2021.
14. S. Dutta, R. Sikder, M. R. Islam, et al., "Comparing machine learningalgorithms in early CKD detection," *Journal of Computer Science andTechnology Studies*, 2024.
15. M.S.Bhuiyan,I.K.Chowdhury,M.Haider,etal.,"Advancementsin early detection of lung cancer using machine learning algorithms,"*Journal of Computer Science and Technology Studies*, 2024.
16. E. Dritsas and M. Trigka, "Lung cancer risk prediction with machinelearning models," *Big Data and Cognitive Computing*, vol. 6, no. 4, p.139, 2022.
17. P. K. Bhowmik, M. N. Miah, M. K. Uddin, et al., "Optimizing lungcancer risk prediction with machine learning algorithms," *Journal ofMedical and Health Studies*, vol. 5, no. 4, pp. 35–48, 2024.
18. G. Gupta, V. Kumar, and P. Karuppanan, "Lung cancer risk predictionusing machine learning and deep learning," in *Proc. IEEE SCES Conf.*, pp.1–6,2024.
19. K.MohanandB.Thayyil,"Machinelearningtechniquesforlungcancerprediction," *International Journal of Data Informatics and IntelligentComputing*, 2023.
20. A. Pathan et al., "Lung cancer prediction using XGBoost and featureengineering," *International Journal of Applied Artificial Intelligence*,2024.
21. M.Islametal.,"Ensemblemethodsforlungdiseasedetection,"*ClinicalData Intelligence Journal*, 2022.
22. R. Gupta et al., "Lung cancer risk stratification using logistic regression(PLCOm2012)," *Clinical Screening Research*, 2023.
23. V. Rani et al., "Traditional and modern techniques for lung cancer riskprediction," *International Journal of Cancer Studies*, 2022.
24. S. Raoof et al., "Gaps and challenges in machine learning–based lungcancer prediction," *Medical Informatics Review*, 2020.
25. M. Thallam et al., "Explainability challenges in machine learninghealthcare models," *AI in Medicine Review*, 2020.