iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78922

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 III Mar 2026- Available at www.ijraset.com

Disease Prediction System using Machine Learning

Date Sahil Dagadu[1], Wagadare Om Tukaram[2], Mayur Santosh Lagad[3], Sarthak Shantaram Agale[4], Guide:-Nawale S.K.[5]
Department of Information Technology, Samarth Polytechin, Belhe

Abstract: ***This paper presents a medical-history-based potential disease prediction algorithm designed to assist in intelligent medical decision-making. Leveraging healthcare big data and deep learning, the proposed method aims to identify diseases that may be overlooked due to a patient's limited medical knowledge. Inspired by recommendation systems, the model applies deep learning to analyze historical medical records and predict possible diseases. It combines high-order and low-order relations between medical data using a machine learning approach, resulting in improved prediction accuracy.***
Keywords: Healthcare Big Data, Deep Learning, Recommendation System, Disease Prediction, Medical History, Factorization Machine, Attention Network, Medical Informatics.

I. INTRODUCTION

The rapid development of internet technologies and digital transformation in the healthcare industry has led to the emergence of massive electronic medical data. Electronic Medical Records (EMRs), online medical consultations, automated diagnostic tools, and cloud-based healthcare platforms have replaced traditional paper-based systems, ushering in an era of Healthcare Big Data. This data encompasses a wide spectrum, including clinical records, diagnostic imaging, health insurance claims, biomedical research, environmental influences, public health statistics, and behavioral information. Effectively analyzing and leveraging this vast dataset has the potential to transform healthcare by enhancing disease diagnosis, treatment efficiency, service quality, and the personalization of medical recommendations.

One critical application of healthcare big data analytics is potential disease prediction. It allows for identifying diseases that a patient might develop, particularly those they may overlook due to limited medical expertise. In many cases, patients fail to undergo necessary diagnostic tests simply because they are unaware of potential conditions associated with their symptoms. This lack of foresight may delay timely diagnosis and treatment, which can lead to the worsening of health conditions. Hence, a system capable of accurately predicting potential diseases can assist both patients and healthcare providers in proactively addressing health risks.

In this perspective, diseases are treated as items, and the patient's medical history acts as a record of interactions. Drawing on the collaborative filtering and hybrid recommendation models used in online systems, the proposed methodology identifies patterns in historical medical data to suggest likely future diseases.

Traditional recommendation techniques, such as Collaborative Filtering (CF) and Content- Based Filtering, have limitations when applied to healthcare data. They either assign equal weight to all past interactions (i.e., previous diseases) or fail to capture complex, latent relationships among diseases. More advanced models like hybrid systems attempt to bridge these gaps, but they still struggle with high-dimensional data, sparsity, and the need for semantic relevance between diseases.

II. MOTIVATION

Patients may miss important medical examinations because they lack professional knowledge. A system that suggests potential diseases based on historical medical data can guide targeted check-ups, ensuring better prevention and timely treatment. Traditional disease prediction models fail to capture complex relations between various diseases. There is a growing need for a model that can explore both low- and high-order relationships among diseases efficiently, handle sparse data, and provide accurate predictions.

III. LITERATURE REVIEW

Author :Nianyin Zeng, Zidong Wang, Yurong Li, Min Du and Xiaohui Liu

The main purpose of this paper is to handle the dynamic modeling problem with state constraints by combining the extended Kalman filtering and constrained op timization algorithms via the maximization probability method. More specifically, a recently developed SPSO algorithm is used to cope with the constrained optimization problem by converting it into an unconstrained optimization one through adding a penalty term to the objective function.

Author :Nianyin Zeng, Zidong Wang, Yurong Li, Min Du∗ and Xiaohui Liu
In this paper, a mathematical model for sandwich type lateral flow immunoassay is developed via short available time series. A nonlinear dynamic stochastic model is considered that consists of the biochemical reaction system equations and the observation equation. After specifying the model structure, we apply the extend Kalman filter (EKF) algorithm for identifying both the states and parameters of the nonlinear state-space model. It is shown that the EKF algorithm can accurately identify the parameters and also predict the system states in the nonlinear dynamic stochastic model through an iterative procedure by using a small number of observations.

Author :PAULINE JOHANSSON RN,MSc 1,GO ¨RAN PETERSSON MD, PhD2 and GUNILLA NILSSON
Aim Theaimofthis study was to describe one nurses experience of using a personal digital assistant (PDA) in nursing practice. Background Nurses handle large amounts of information and a PDA may contain valuable information that nurses need in their daily work. Methods In this qualitative single case study, data were collected through an open ended interview with one registered nurse and were analysed by content analysis. Results The findings show that the PDA provides immediate access to information anywhere and at anytime, with advantages for both the nurse and for her patients. The PDA increased her confidence and efficiency in practice; it was easier to keep up-to-date and spend more time with the patient. Furthermore, the PDA was perceived as improving patient safety and patient participation.

Author :Taoying Li *, LinlinJin, Zebin Wu and Yan Chen
The recommendation algorithm in e-commerce systems is faced with the problem of high sparsity of users' score data and interest's shift, which greatly affects the performance of recommendation. Hence, a combined recommendation algorithm based on improved similarity and forgetting curve is proposed. Firstly, the Pearson similarity is improved by a wide range of weighted factors to enhance the quality of Pearson similarity for high sparse data. Secondly, the Ebbinghaus forgetting curve is introduced to track a user's interest shift. User score is weighted according to the residual memory of forgetting function.

Author :Abbadi, A.E.
The Author Index contains the primary entry for each item, listed under the f irst author's name. The primary entry includes the coauthors' names, the title of the paperorotheritem,anditslocation, specified bythepublicationabbreviation, year, month, and inclusive pagination. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication abbreviation, month, and year, and inclusive pages. Note that the item title is found only under the primary entry in the AuthorIndex**.**

IV. EXISTING SYSTEM

Traditional disease prediction approaches often rely on:

1) Collaborative Filtering (CF): Focuses on past interactions but treats all historical data equally.
2) Content-Based Models: Depend heavily on explicitly defined features and fail to uncover deep relations.
3) Hybrid Models: Combine multiple techniques but may still fall short on high-order feature analysis.

These systems have limitations such as:

- Inability to capture deep, nonlinear relationships.
- Equal weighting of symptoms or diseases.
- Overfitting due to sparse and high-dimensional data.

V. PROPOSED SYSTEM

The proposed system is a machine learning-based model that predicts potential diseases using a patient's medical history. It treats disease prediction like a recommendation system. The system combines machine learning techniques to learn complex, high-level patterns to capture simpler, low-level relationships between diseases. An attention mechanism is used to give more importance to the most relevant past diseases. The outputs from naïve bayes parts are merged to predict the probability of a patient having a specific disease. This approach helps doctors and patients identify likely health risks early and take preventive action

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

VI. ARCHITECTURE

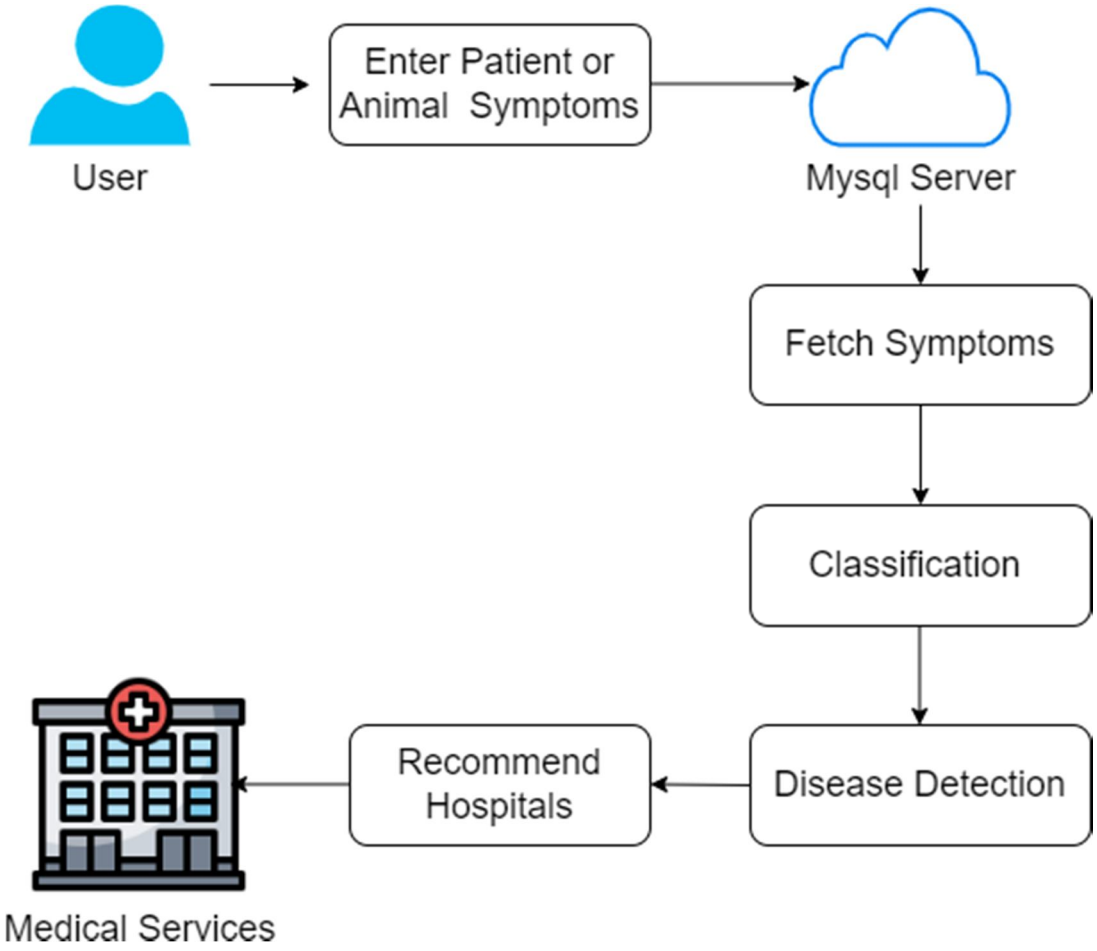


Fig: "Architecture of the Disease Prediction System"

VII. ALGORITHMS

A. *Naive Bayes*

- Given training dataset D which consists of documents belonging to different class say Class A and Class B
- Calculate the prior probability of class A=number of objects of class A/total number of objects
- Calculate the prior probability of class B=number of objects of class B/total number of objects
- Find NI, the total no of frequency of each class
- Na=the total no of frequency of class A
- Nb=the total no of frequency of class B
- Find conditional probability of keyword occurrence given a class:
- P (value 1/Class A) =count/ni (A)
- P (value 1/Class B) =count/ni (B)
- P (value 2/Class A) =count/ni (A)
- P (value 2/Class B) =count/ni (B)
- ……..
- P (value n/Class B) =count/ni (B)
- Avoid zero frequency problems by applying uniform distribution
- Classify Document C based on the probability p(C/W)
- Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A)……. P(value n /Class A)
- Find P (B/W) =P (B)*P (value 1/Class B)* P (value 2/Class B)……. P(value n /Class B)
- Assign document to class that has higher probability.

B. Random Forest

The algorithm used here is Random Forest. Random Forest is the most popular and powerful algorithm of machine learning.

- Step 1: Assume N as number of training samples and M as number of variables within the classifier.
- Step 2: The number m as input variables to decide the decision at each node of the tree; m should be much less than M.
- Step 3: Consider training set by picking n times with replacement from all N available training samples. Use the remaining of the cases to estimate the error of the tree, by forecasting their classes.
- Step 4: Randomly select m variables for each node on which to base the choice at that node. Evaluate the best split based on these m variables in the training set.
- Step 5: Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier). For forecasting, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is repeated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. i.e. classifier having most votes.

VIII. CONCLUSION

The proposed algorithm successfully integrates the strengths of machine learning mechanisms to create a robust system for predicting potential diseases based on a patient's medical history. The results show that this approach significantly outperforms traditional methods in terms of prediction accuracy and ranking quality. The model's ability to account for both high- and low-order relations enhances its real-world applicability in medical decision-making.

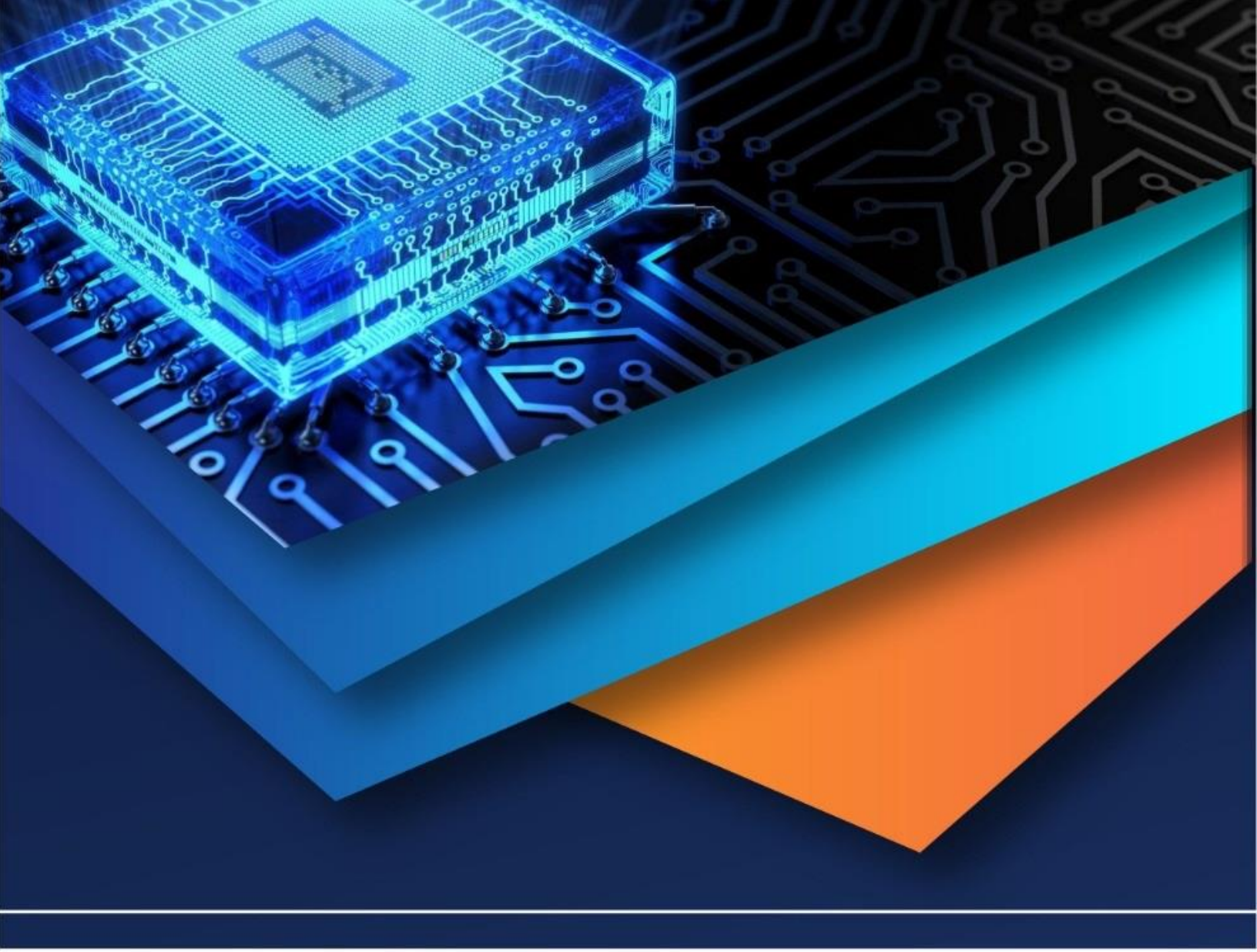
IX. FUTURE SCOPE

The authors suggest the following directions for future research:

1) Integration with Symptom Data: Using explicit symptom descriptions and side information for improved predictions.
2) Explainability: Enhancing the model with domain-specific knowledge to make predictions more interpretable for doctors and patients.
3) Real-Time Systems: Developing interactive systems for real-time use in clinics.
4) Transfer Learning: Applying the model to various healthcare datasets and patient populations for generalization.
5) Multi-modal Data Fusion: Incorporating genomic, behavioral, and environmental data for more holistic prediction models.

REFERENCES

[1] N. Zeng, Z. Wang, Y. Li, M. Du, and X. Liu, ―A hybrid ekf and switching pso algorithm for joint state and parameter estimation of lateral flow immunoassay models,‖ IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 2, pp. 321–329, 2011.
[2] N. Zeng, Z. Wang, Y. Li, M. Du, and X. Liu, ―Inference of nonlinear state space models for sandwich-type lateral flow immunoassay using extended kalman filtering,‖ IEEE Transactions on Biomedical Engineering, vol. 58, no. 7, pp. 1959–1966, 2011.
[3] P. E. Johansson, G. I. Petersson, and G. C. Nilsson, ―Personal digital assistant with a barcode reader—a medical decision support system for nurses in home care,‖ International journal of medical informatics, vol. 79, no. 4, pp. 232–242, 2010.
[4] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern information retrieval, vol. 463. ACM press New York, 1999.
[5] H.-L. Xu, X. Wu, X.-D. Li, and B.-P. Yan, ―Comparison study of internet recommendation system,‖ Journal of software, vol. 20, no. 2, pp. 350–362, 2009.
[6] F. Xue, X. He, X. Wang, J. Xu, K. Liu, and R. Hong, ―Deep item-based collaborative filtering for top-n recommendation,‖ ACM Transactions on Information Systems (TOIS), vol. 37, no. 3, p. 33, 2019.
[7] E. Christakopoulou and G. Karypis, ―Hoslim: Higher-order sparse linear method for top-n recommender systems,‖ in Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 38–49, Springer, 2014.
[8] X. He, Z. He, J. Song, Z. Liu, Y.-G. Jiang, and T.-S. Chua, ―Nais: Neural attentive item similarity model for recommendation,‖ IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 12, pp. 2354–2366, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY