Multi Disease Prediction and Hospital Recommendation

Pradeepth S Kumar[1] Seema Nagaraj[2]

[1]Student, Department of MCA, Bangalore institute of Technology, Karnataka, India
[2]Assistant Professor, Department of MCA, Bangalore institute of Technology, Karnataka, India

---***---

Abstract

Healthcare systems worldwide face challenges in providing timely and accurate diagnoses, especially in rural areas and locations with limited resources. To address this problem, this project proposes a unified framework for predicting multiple diseases using machine learning (ML), deep learning (DL), and ensemble methods. The system employs various models, including Random Forest, Support Vector Machines, Naïve Bayes, Gradient Boosting, and Neural Networks. These models are trained on clinical datasets and symptom-based inputs to predict a range of diseases, such as diabetes, heart disease, liver disorders, kidney disease, Parkinson's, breast cancer, and lung cancer.

With a modular design, the framework can manage symptom-driven predictions and disease-specific classifiers, making it flexible and scalable. Better preprocessing, feature selection, and hyperparameter tuning improve accuracy and generalization. Big data technologies efficiently manage large-scale medical records. The system also provides hospital recommendations, making it suitable for telemedicine platforms, rural healthcare centers, and personal wellness applications.

Results from several studies demonstrate high precision, recall, and sensitivity, with ensemble models like gradient boosting achieving accuracy close to 99%. This project enables proactive healthcare through early disease detection. It also reduces the burden on medical infrastructure, supports physicians in clinical decision-making, and provides individuals with access to affordable diagnostic support.

1.INTRODUCTION

Healthcare is essential to human life and has a direct effect on social and economic well-being. However, millions of people around the world still struggle to access timely and accurate diagnoses. This is due to limited medical infrastructure, a shortage of experts, and high costs, particularly in rural and underdeveloped areas. These challenges often lead to delayed treatment and serious health problems. With the rapid growth of data and improvements in artificial intelligence, machine learning (ML) and deep learning (DL) techniques have emerged as promising solutions for automated disease prediction and healthcare support [1], [2]. Many studies have applied ML models to predict specific diseases such as diabetes, liver disorders, and heart disease. For example, Naïve Bayes and KNN have been tested on the PIMA dataset for diabetes classification, where Naïve Bayes performed better [3]. Likewise, ensemble approaches like Gradient Boosting have shown strong results in predicting liver disease, with accuracy levels nearing 99% [5]. While effective, these methods are confined to single-disease contexts and cannot handle multiple conditions at once. To overcome this limitation, researchers have suggested multi-disease prediction systems that use various algorithms, including Random Forest, SVM, Decision Trees, and ANN [1], [4]. Some studies have even integrated big data frameworks like Apache Spark and web-based interfaces for better scalability and accessibility [4]. Mobile-based solutions have been created to let users input symptoms and receive predictions for up to 40 diseases [2]. However, the accuracy of these predictions depends heavily on the symptoms reported by users. Despite these advancements, existing systems still face significant challenges. Many do not offer hospital recommendations, give limited advice on precautionary measures, and are not fully integrated into telemedicine workflows [2], [6]. Moreover, there are still accessibility gaps, especially in low-resource settings. To tackle these issues, this project proposes an integrated multi-disease prediction and hospital recommendation system. This system will combine symptom-based prediction, disease-specific modules, ensemble learning, and hospital guidance. The goal is to provide accurate, scalable, and user-friendly healthcare assistance, bridging the gap between patients and healthcare providers.

II. LITERATURE SURVEY

1. Symptom-based multi-disease prediction (client/mobile focus)

Early research on predicting multiple diseases through symptoms shows that end-to-end apps can let users select symptoms and receive a list of likely conditions. A study in Procedia Computer Science developed a mobile self-assessment app that combines KNN, SVM, Decision Tree, Random Forest, Naïve Bayes, and ANN by averaging probability vectors. It used a Kaggle symptom-disease dataset and reported high accuracy in training and testing for about 40 diseases. The study also described a practical client-server design with thresholds for the top outputs. Another paper on multi-disease assessment looked at Random Forest, Naïve Bayes, and SVM across around 41 diseases. It featured an easy-to-use interface where users can add or remove symptoms and get predictions from different algorithms. The authors noted that structured symptom inputs work well, but there is still demand for multimodal data.

2. Multi-disease systems with ML/DL and big data

Recent research seeks to improve on simple apps by combining machine learning and deep learning with big data systems like Apache Spark and web deployment using Flask. This approach can handle various data sources, including electronic health records, images, and genetic data. These systems have shown strong results in breast cancer, lung cancer, and diabetes, supporting scalability through Spark and a variety of models like Random Forest, Gradient Boosting, and CNNs to assist in predictions.

3. Disease-specific models (structured clinical features)

Single-disease studies provide baseline methods and insights that can help your project:

Diabetes (PIMA): Comparative research shows that Naïve Bayes outperforms KNN with standard preprocessing and confusion-matrix evaluations. This is useful when class priors and assumptions of conditional independence are mostly correct.

Liver disease (large clinical cohort): A large study with 30,691 records evaluated bagging, boosting, and voting methods. It found that Gradient Boosting had the best overall performance, achieving about 98.8% accuracy, with precision, recall, and F1 scores around 98.5%. The study also looked at feature importance, revealing that biochemical markers were more influential than demographics and included ROC/AUC comparisons.

4. Ensemble learning patterns and takeaways

Ensemble methods generally outperform single models across various areas. In symptom-based applications, averaging probabilities helps stabilize predictions from different learners. In clinical-feature settings, tuned boosting methods such as Gradient Boosting provide the best accuracy and reliability. They also feature clear metric reporting, including confusion matrices, ROC, AUC, and feature contribution plots, which help in understanding and reviewing models. Method surveys within the liver disease study list successful ensemble methods for other conditions like cardiovascular problems, breast, skin, and thyroid diseases, and myocardial infarction, highlighting the effectiveness of these approaches.

5. Datasets, pipelines, and deployment considerations

Symptom-driven studies often depend on public symptom-disease databases, such as those from Kaggle. They use preprocessing, multi-model training, and thresholding in a client-server framework. For models focusing on specific diseases, standard clinical datasets like PIMA allow for quick comparisons and evaluations of algorithms. Research in big data stresses using Spark for distributed ETL and training, while Flask is suitable for lightweight inference APIs. This enables scalable and flexible deployment in healthcare.

6. Gaps and open challenges

Despite promising results, the literature shows recurring limitations that your project can address:

Generalization and external validity: Many studies report internal test metrics, but few validate these across multiple sites or in prospective studies.

Data modality breadth: While symptom-only inputs are useful, they miss important information in labs, imaging, and notes. Authors have specifically called for multimodal pipelines to tackle this issue.

Imbalanced data and calibration: Several reports point out sensitivity to class imbalances and insufficient focus on probability calibration, which is vital for risk assessment.

Workflow integration: Few systems link predictions to actionable next steps, such as recommendations for hospitals or doctors, precautions, or multilingual text-to-speech for accessibility. Your project's Phase-1 plan will directly target these gaps with user interface, text-to-speech, and recommendation modules.

7. Summary position for this project

The field shows the potential for multi-disease prediction from symptoms using ensemble methods, solid baselines for single diseases based on clinical features, and the ability to scale with big data tools. Building on these foundations, your project aims for an integrated, modular approach that combines symptom-based screening, disease-specific classifiers, ensemble learning, and deployment features such as hospital recommendations and voice assistance.

III. EXISTING SYSTEM

Several disease prediction systems using machine learning and deep learning have been proposed in recent years, but most are limited in scope and functionality. Many existing models focus on predicting a single disease. For example, there is diabetes detection using the PIMA dataset and liver disease prediction with ensemble learning techniques. In these cases, algorithms like Naïve Bayes and Gradient Boosting have shown good accuracy. While effective for specific conditions, these methods cannot diagnose multiple diseases at once.

Symptom-based mobile applications have also been developed to predict up to 40 diseases using classifiers like Random Forest, SVM, and Naïve Bayes. However, their accuracy heavily depends on user-reported symptoms, and they often struggle with overlapping features from different diseases.

More advanced research systems use big data and deep learning, employing tools like Apache Spark for large-scale processing and Flask for deployment. These systems have demonstrated high accuracy for diseases such as breast cancer, lung cancer, and diabetes. Still, they mainly exist in research settings and are not widely available to patients, especially in rural or under-resourced areas. Additionally, most existing solutions focus only on prediction and do not provide actionable insights such as precautionary measures or hospital recommendations. This highlights the limitations of current systems in terms of scalability, accessibility, and usability in real-world settings, showing the need for a more integrated multi-disease prediction framework.

IV. PROPOSED SYSTEM

The proposed system is a platform that predicts multiple diseases and recommends hospitals. It uses machine learning, deep learning, and ensemble learning to provide real-time healthcare support. Unlike current models that focus on a single disease or only symptoms, this system takes a modular approach. Each disease, including diabetes, heart disease, liver disorders, kidney disease, Parkinson's, breast cancer, and lung cancer, has a specialized predictive module trained on specific datasets. There is also a general symptom-based module that identifies possible illnesses based on user input and provides probability scores, precautionary measures, and hospital recommendations.

To improve accuracy and reliability, the system uses data preprocessing, feature selection, and hyperparameter tuning. It integrates supervised learning algorithms such as Random Forest, SVM, Naïve Bayes, Gradient Boosting, and Neural Networks. Additionally, big data tools help manage large-scale medical data. A user-friendly interface makes the system accessible on smartphones, laptops, and web platforms, catering to the needs of both urban and rural healthcare.

By combining prediction, precaution, and hospital recommendations in one framework, the proposed system connects patients with healthcare providers, reduces the hospital workload through early diagnosis, and allows individuals to actively manage their health. The first major benefit of the system is its ability to predict multiple diseases at once with high accuracy, which is an improvement over existing systems that focus on one disease. The second benefit is the inclusion of hospital recommendations and precautionary guidance, making the system more practical and useful for everyday healthcare..

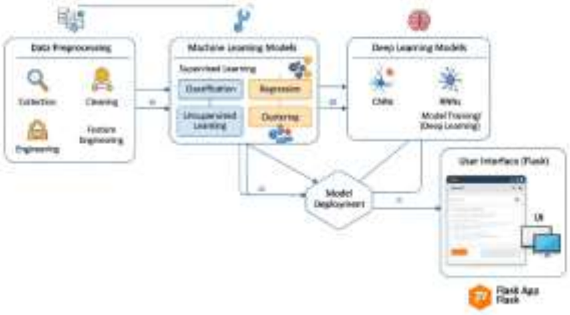

Fig 1: Proposed Model

V. IMPLEMENTATION

The proposed multi-disease prediction and hospital recommendation system is implemented in several phases. It begins with data collection and preprocessing. Publicly available datasets like the PIMA Diabetes dataset, the Indian Liver Patient dataset, and Kaggle's symptom-disease datasets are used to train the models. Preprocessing methods, such as handling missing values, normalization, and feature selection, improve data quality and ensure consistency across different disease modules.

The system has a modular design. Each disease prediction uses a specific machine learning or deep learning model. Algorithms like Random Forest, Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting, and Artificial Neural Networks are implemented and fine-tuned for accuracy. In some cases, ensemble methods are used to increase robustness and reduce overfitting. Each model is evaluated with performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix to ensure reliable predictions.

A general symptom-based prediction module is also included. It takes symptoms provided by the user as input and outputs the top possible diseases along with their probability scores. The system offers precautionary measures and suggests nearby hospitals through integrated APIs. The hospital recommendation module uses location-based services, like Google Maps API, to find the closest healthcare centers for immediate consultations.

The user interface is designed to be accessible on both web and mobile platforms. Flask handles backend integration, while HTML, CSS, and JavaScript comprise the front-end interface. The system is lightweight enough to run on smartphones and laptops, making it suitable for rural areas with limited resources.

By combining disease prediction, precautionary guidance, and hospital recommendations, the implementation ensures the system is both technically sound and user-friendly. It aims to be practically useful in real-world healthcare situations.

VI. CONCLUSIONS

The proposed multi-disease prediction and hospital recommendation system demonstrates how machine learning and deep learning can address real healthcare problems. By combining specific models for various diseases with a general symptom-based module, the system offers timely predictions for multiple conditions, including diabetes, heart disease, liver disorders, kidney disease, lung cancer, breast cancer, and Parkinson’s. Unlike current single-disease solutions, the modular design allows for growth and flexibility, enabling the framework to incorporate additional disease models as needed.

Adding precautionary measures and hospital recommendation features makes the system more than just a diagnostic tool; it serves as a practical healthcare assistant that guides users to the right medical care. The use of ensemble learning techniques, large data handling, and a user-friendly interface ensures both accuracy and ease of use. This makes the system suitable for both urban and rural areas.

This project helps bridge the gap between medical knowledge and patient access by providing an affordable, effective, and scalable healthcare solution. With further enhancements like integration with wearable devices, electronic health records, and real-time monitoring, the system could evolve into a smart healthcare platform that empowers individuals, supports medical professionals, and reduces pressure on healthcare systems.

VII. FUTURE ENHANCEMENTS

Although the proposed multi-disease prediction and hospital recommendation system provides accurate predictions and practical healthcare advice, several areas need improvement to enhance its effectiveness and usability. One important addition is the integration of Internet of Things (IoT) devices and wearable sensors, such as smartwatches, fitness trackers, and medical monitoring devices. This integration would allow the system to gather real-time health data, including heart rate, oxygen levels, blood pressure, and glucose levels, enabling continuous monitoring and early alerts.

Another improvement is the use of cloud-based deployment. This change would allow the system to manage larger datasets, scale easily for multiple users, and offer access across devices without requiring heavy local resources. It could also integrate with electronic health records (EHRs), ensuring that predictions are personalized based on a patient’s medical history.

The system could also use natural language processing (NLP) and chatbot-based interaction. This feature would allow users to describe their symptoms in everyday language rather than selecting from predefined options. Adding multilingual support and text-to-speech (TTS) features would improve accessibility for users in rural areas or those who do not speak English.

To encourage clinical adoption, future versions could employ explainable AI (XAI) techniques to clarify the reasoning behind predictions. This method can help build trust among doctors and patients. Finally, connecting with telemedicine platforms could create a complete digital healthcare system. Users would receive predictions and recommendations plus the option to consult with doctors online directly through the application.

By implementing these improvements, the system can evolve into a smart healthcare platform that offers real-time monitoring, personalized healthcare, and smooth communication between doctors and patients. This progress can result in more proactive and accessible healthcare services.

VIII REFERENCES

[1] H. Karwa, P. Gupta, R. Agrawal, G. S. Virdi, A. Kumar, and S. Jain, “Multi-disease prediction with machine learning,” *International Journal of Health Sciences*, vol. 6, no. S2, pp. 9477–9483, 2022.

[2] A. Patil and A. Jadon, “Auto-labelling of bug report using natural language processing,” in *Proc. 2023 IEEE 8th Int. Conf. Convergence in Technology (I2CT)*, IEEE, 2023.

[3] A. Aburakhia and M. Alshayeb, “A Machine Learning Approach for Classifying the Default Bug Severity Level,” *Arabian Journal for Science and Engineering*, pp. 1–18, 2024.

[4] M. Q. Shatnawi and B. Alazzam, “An Assessment of Eclipse Bugs’ Priority and Severity Prediction Using Machine Learning,” *Int. J. Commun. Netw. Inf. Secur.*, vol. 14, no. 1, pp. 62–69, 2022.

[5] H. Bani-Salameh, M. Sallam, and B. Alshboul, “A deep-learning-based bug priority prediction using RNN-LSTM neural networks,” *e-Informatica Software Engineering Journal*, vol. 15, no. 1, 2021.

[6] H. A. Ahmed, N. Z. Bawany, and J. A. Shamsi, “Capbug—a framework for automatic bug categorization and prioritization using NLP and machine learning algorithms,” *IEEE Access*, vol. 9, pp. 50496–50512, 2021.

[7] S. D. Immaculate, M. F. Begam, and M. Floramary, “Software bug prediction using supervised machine learning algorithms,” in *Proc. 2019 Int. Conf. Data Science and Communication (IconDSC)*, IEEE, 2019.

[8] A. Yadav and S. S. Rathore, “A Hierarchical Attention Networks based Model for Bug Report Prioritization,” in *Proc. 17th Innovations in Software Engineering Conf.*, 2024.

[9] W. Zheng, et al., “Duplicate Bug Report detection using Named Entity Recognition,” *Knowledge-Based Systems*, vol. 284, p. 111258, 2024.

[10] J. S. H. Al-Bayati, M. AlShamma, and F. N. Tawfeeq, “Enhancement of Recommendation Engine Technique for Bug System Fixes,” *J. Adv. Inf. Technol.*, vol. 15, no. 4, 2024.

[11] A. Lamkanfi, J. Pérez, and S. Demeyer, “The eclipse and mozilla defect tracking dataset: a genuine dataset for mining bug information,” in *Proc. 2013 10th Working Conf. Mining Software Repositories (MSR)*, IEEE, 2013.

[12] A. Kukkar and R. Mohana, “A supervised bug report classification with incorporate and textual field knowledge,” *Procedia Computer Science*, vol. 132, pp. 352–361, 2018.

[13] T. Hirsch and B. Hofer, “Using textual bug reports to predict the fault category of software bugs,” *Array*, vol. 15, p. 100189, 2022.

[14] A. A. Ahmad, et al., “Deep Bug Reports Processing (DBRP): A Systematic Literature Review,” 2023.

[15] T. Thomas and B. Hofer, “Root cause prediction based on bug reports,” in *Proc. IEEE Conf.*, 2020.