Power System Technology
ISSN:1000-3673

Received: 06-11-2024 *Revised: 15-12-2024* *Accepted: 05-01-2025*

Advanced Predictive System for Diagnosing Patient Disease through Machine Learning Techniques

Vivek Raj Singh[1], Dr. Shwet Ketu[2]

[1] CSE Student, Computer Science and Engineering, Galgotias University Noida, India

[2] Assistant Professor, Computer Science and Engineering, Galgotias University Noida, India

***Corresponding Author:** Vivek Raj Singh, CSE Student, Department of Computer Science Galgotias University, Uttar Pradesh, India

Email: vivekrajsingh97@gmail.com

ABSTRACT

The aim of this research is to develop a web application that predicts multiple diseases, including Diabetes, Breast Cancer, Heart Disease, Alzheimer disease, brain tumor, Covid-19, and Pneumonia, using machine learning and deep learning models. The models were trained on large datasets sourced from Kaggle. The study includes data collection, preprocessing, model selection, and deployment of the trained models. The application uses Python Flask as the web framework, and models such as CNN (for Covid-19, and Pneumonia) are integrated to ensure accurate predictions. This project achieved prediction accuracies of up to 97% for certain diseases, showing the viability of machine learning in healthcare applications. The findings suggest that the developed system could support early detection of disease and improve healthcare outcomes.

Keywords: Disease Prediction, Machine Learning, CNN, Deep Learning, Flask, Healthcare, Kaggle

I. INTRODUCTION

The field of healthcare has evolved dramatically over the past few decades ,with artificial intelligence(AI) playing an increasingly important role in diagnosis ,treatment planning and patient monitoring. Predictive models powered by machine learning (ML)and deep learning (DL) offer an unprecedented opportunity to diagnose disease early ,thus enabling more effective treatment plans and reducing the burdens on healthcare systems. These technologies help analyze vast amount of patient data, identifying patterns and correlation that may not be readily apparent to human practitioners.

The development of predictive models for disease diagnosis is not new; however, the focus has primarily been on individual diseases. For example, various studies have successfully applied machine learning to predict diabetes, heart disease, or cancer. However, this research seeks to unify the prediction of multiple diseases into a single application, offering a

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

comprehensive diagnostic tool. By predicting multiple diseases in a single system, healthcare professionals can use the platform for broader diagnostic purposes, thus improving efficiency in medical evaluations.

The web-based application introduced in this study uses both ML and DL models to predict seven diseases:

- **Diabetes**: A chronic condition that affects how the body processes glucose.
- **Breast Cancer**: One of the most common cancers, especially in women, which early detection greatly increases the chances of survival.
- **Heart Disease**: A range of cardiovascular conditions that can lead to life-threatening events such as heart attacks.
- **Alzheimer disease**: A condition that leads to loss of memory which mainly affects cortex region of brain.
- **Brain tumor** : Diseases that affect brain functions and lead to fatal death
- **Covid-19**: A disease that affects our lungs badly leading oxygen deficiency and sudden death
- **Pneumonia**: An infection that inflames air sacs in the lungs, which can be caused by viruses, bacteria, or fungi.

The aim of this project is to develop a user-friendly web application that integrates multiple predictive models, providing users with a straightforward interface to input data and receive diagnostic predictions. The platform is built on Python's Flask framework, offering seamless integration with machine learning models. The system accepts both numerical data (such as medical test results) and image data (such as X-rays for Pneumonia detection), making it a versatile tool for disease prediction.

In this paper, we will explore the datasets used for training the models, the methods for preprocessing and feature selection, and the algorithms applied for each disease prediction. The results demonstrate high accuracy for the majority of the models, especially for diseases like brain tumor and breast cancer, where accuracy exceeds 96%. We also explore the challenges and limitations, including the relatively lower accuracy of liver disease prediction, and suggest future improvements.

II. METHODOLOGY

The methodology for this research involves a step-by- step approach to developing a web-based application that predicts multiple diseases using machine learning (ML) and deep learning (DL) models. The primary focus is on creating a system that is user-friendly, accurate, and scalable, with the ability to diagnose seven diseases: Diabetes, Breast Cancer, Heart

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Disease, Covid-19, Alzheimer disease, breast cancer, and Pneumonia. This section explains the dataset collection process, preprocessing techniques, model selection, and the overall architecture of the web application.

A. Datasets

Datasets were sourced from Kaggle, a well-known platform that provides open-access datasets for various machine learning tasks. The datasets used in this research cover diverse medical conditions, with both structured (tabular) and unstructured (image-based) data. Each dataset was carefully selected based on quality, size, and relevance for training ML and DL models. The following datasets were used:

- **Diabetes -Dataset**-The Indians Diabetes Database, consisting of diagnostic measurements like glucose, insulin levels, and body mass index (BMI).
- **Breast Cancer Dataset**- The Wisconsin Diagnostic Breast Cancer Dataset, which contains 30 features including radius, texture, and area of cell nuclei.
- **Heart Disease Data set** - The Cleveland Heart Disease Dataset, which includes 14 features such as cholesterol- levels, resting blood- pressure and age.
- **Covid-19 disease**- A data set of X ray image set of infected lungs and normal lungs.
- **Alzheimer disease**: A data set of brain mainly cortex part of brain with x ray image.
- **Brain tumor** - A dataset of microscopic images of brain with tumor x ray analyses.
- **Pneumonia Dataset**: A chest X-ray image dataset, with binary classification between pneumonia-infected and normal lungs.

Each dataset was divided into training, validation, and test sets using an 80-10-10 split to ensure robust model training and evaluation.

B. Data Processing

Data preprocessing is a critical step in any machine learning or deep learning task. The preprocessing methods used in this research vary depending on the type of data (structured or image-based).

C. Structured Data Preprocessing (For ML Models

For diseases- such as- diabetes, breast cancer, heart- disease, the datasets are tabular, containing numerical and categorical features. The following preprocessing steps were applied:

- **Missing Data Handling**: Missing values were either filled using median imputation or removed if the missing data exceeded a threshold (e.g., 20%).

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

- **Feature Scaling**: All numerical features were standardized using Z-score normalization, ensuring that each feature has a mean of 0 and a standard deviation of 1.
- **One-Hot Encoding**: Categorical features (e.g., gender) were converted into binary vectors using one-hot encoding to make them compatible with machine learning algorithms
- **Feature Selection**: Correlation analysis and feature importance methods were used to eliminate redundant features, reducing the risk of overfitting.

D. Image Data Processing(For DL Models)

For diseases such as Covid 19 and pneumonia, the datasets consist of medical images. Preprocessing steps for image data included:

- **Image Resizing**: All images were resized to a uniform size (128x128 pixels) to ensure consistency inmodel training.
- **Normalization**: Pixel values were normalized to a range between 0 and 1 to improve the convergence of deep learning models.
- **Data Augmentation**: To combat overfitting, data augmentation techniques such as rotation, flipping, and zooming were applied to artificially increase the size of the dataset.

E. Model Selection

The choice of model depends on the type of data being processed (tabular or image-based). Different models were employed for each disease based on the complexity of the dataset and the nature of the disease being predicted.

F. Machine Learning Model for Tabular Data

For diseases like diabetes, breast cancer, heart disease, Alzheimer disease, and covid-19 disease, traditional machine learning algorithms were employed. The following models were tested:

- **Logistic Regression**: Used for binary classification tasks (e.g., diabetes and breast cancer). Logistic regression was chosen for its simplicity and interpretability.
- **Random Forest**: A decision tree-based ensemble method was used for heart disease and kidney disease prediction due to its ability to handle high-dimensional data and its robustness against overfitting.
- **Support- Vector Machine (SVM)**-SVM was used to classify liver disease, given its ability to find optimal decision boundaries in high dimensional spaces.
- **K--Nearest Neighbors (KNN)**: Applied to breast cancer and kidney disease, as it is known to perform well on small datasets.

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

G. *Deep learning Model for Image data*

For image-based diseases such as covid-19 and pneumonia, Convolutional Neural Networks (CNNs) were selected due to their exceptional performance in image classification tasks. The following architecture was used:

- **CNN Architecture**: A deep CNN with multiple convolutional layers, ReLU activation functions, and max- pooling layers was designed to capture spatial hierarchies in the images. Dropout layers were added to prevent overfitting. The final layer uses a softmax activation for binary classification (infected or uninfected for malaria, pneumonia or healthy for pneumonia).

H. *Model Training and Evaluation*

The models were trained using the following setup:

- **Training Phase**: Each model was trained on the training set using 10 folds cross validation to optimize the hyperparameters and avoid overfitting.
- **Loss Function**: For ML models, binary cross- entropy loss was used. For CNN models, categorical cross- entropy was employed.
- **Optimizer**: The Adam optimizer was used for both machine learning and deep learning models, due to its adaptive learning rate capabilities.
- **Evaluation Metrics**: The performance of each model was evaluated using accuracy, precision, recall, F1- score, and confusion matrices. Specific focus was placed on minimizing false negatives, especially for critical diseases such as cancer and heart disease.
- **Validation and Testing**: After tuning the models using the validation set, final testing was performed on the test set to ensure generalization.

I. *Web Application Development*

The web application was developed using the **Flask** web framework. Flask was chosen for its simplicity and scalability. The key components of the web app include:

- **Frontend**: Built using HTML, CSS, and JavaScript, with Bootstrap for responsive design. Users can input their data for ML models or upload images for DL models.
- **Backend**: Python Flask routes handle requests from the frontend and call the trained models to generate predictions.
- **Model Integration**: The models were serialized using Python's Pickle library, allowing them to be loaded and used for real-time predictions within the web app.
- **Deployment**: The app was tested locally and deployed on cloud-based platforms to ensure its scalability and accessibility.

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

III. PROBLEM STATEMENT

In today's healthcare landscape, early detection and diagnosis of diseases are critical for effective treatment and patient outcomes. However, many healthcare systems, particularly in resource-limited settings, face challenges in providing timely and accurate diagnoses for multiple diseases due to a lack of specialized equipment, skilled personnel, and financial constraints. Furthermore, diseases such as -diabetes, breast cancer, Heart Disease, Covid-19, brain tumor, Alzheimer disease, and pneumonia require distinct diagnostic tests, which are often expensive, time-consuming, and inaccessible to underprivileged populations.

The problem is further compounded by the increasing number of patients who require medical attention, creating a significant burden on healthcare infrastructure. This scenario often leads to delays in diagnosis, misdiagnosis, and ultimately, poor health outcomes, especially for diseases that require immediate intervention.

1) Key issues addressed by this research include:

1. **Lack of Accessible and Affordable Diagnostics**: In many areas, advanced diagnostic tests are expensive and not easily accessible to a large segment of the population.

2. **Time-Consuming Diagnostic Procedures**: Traditional diagnostic methods for diseases like breast cancer, heart disease, and pneumonia require various tests, resulting in delays in receiving crucialtreatment.

3. **Overburdened Healthcare Systems**: With limited healthcare professionals and growing patient numbers, healthcare systems struggle to provide timely and effective diagnoses.

4. **Disease-Specific Diagnostic Tools**: Most diagnostic systems focus on specific diseases, requiring separate tools for each condition, which is inefficient and costly.

IV. EXISTING SYSTEM

Current disease diagnosis relies on specialized clinical tests and expert evaluations, each tailored to specific conditions. For instance, diabetes is diagnosed through blood glucose tests, breast cancer through mammography or biopsy, and heart disease through ECGs or echocardiograms. While accurate, these methods are costly, time-consuming, and often inaccessible, particularly in low-resource settings.

The fragmented approach to diagnostics—requiring separate tools for each disease—creates inefficiencies in healthcare systems. Additionally, the growing demand for medical care puts significant strain on healthcare infrastructure, leading to delays, misdiagnoses, and increased costs.

Although machine learning (ML) and deep learning (DL) are being explored for disease

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

prediction and medical imaging, existing solutions remain focused on single diseases and have not been widely adopted. There is a critical need for an integrated, automated system capable of predicting multiple diseases using ML and DL models, offering a unified, cost-effective solution to improve diagnostic efficiency and accessibility.

V. PROPOSED SYSTEM

This proposed system is a multiple disease prediction platform using Machine Learning (ML) and Deep Learning (DL) models, deployed as a web-based application. It is designed to predict multiple diseases, including Diabetes, breast Cancer, heart Disease, kidney disease, Alzheimer Disease, covid-19, and pneumonia based on user-input health data and medical images.

1) Key Components:

1. **Data Input**:

- **Structured Data Input**: Patient health parameters (e.g., glucose, cholesterol) for diseases like Diabetes, Heart Disease, etc.
- **Image Input**: Medical images (e.g., X- rays for Pneumonia,).

2. Prediction Models:

- **Machine Learning Models**: Used for diseases with structured data (e.g., Diabetes, Heart Disease).

3. **Deep Learning (CNN)**: Used for image- based diseases (e.g., covid-19, Pneumonia).

4. Preprocessing:

- Normalization, standardization, and image preprocessing (e.g., resizing and augmentation).

5. Web Application:

- Developed using **Flask**, it offers a user-friendly interface for data entry and real-time diseaseprediction.

Key Features of the Proposed System

1. Multi-Disease Prediction Capability:

The system is capable of predicting seven critical diseases, providing a single platform for multiplehealth conditions.

2. Integration of Both ML and DL Models:

Machine learning models are used for diseases based on structured health data (e.g., Diabetes, Heart Disease), while deep learning models are applied to image-based diseases (e.g., Pneumonia, Covid-19).

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

3. Real-Time Disease Prediction:

The web app offers real-time predictions once users input data, providing instant feedback based on the trained models.

4. **User-Friendly Interface***:*

The system is designed with a simple and intuitive web interface, ensuring that healthcare professionals and non-experts can easily use the application.

5. Scalability:

The system is scalable and can accommodate additional disease models as needed. It can also be expanded to include new data types or integrate with electronic health record (EHR) systems in the future.

VI. INPUT AND OUTPUT DESIGN

Input Design:

- **Structured Data Input:**
  - User-provided health parameters for diseases like Diabetes, Heart Disease, etc.
  - Example inputs: Age, Blood Pressure, Glucose Level, BMI.
  - **Input Method:** Form fields on the web app.
- **Medical Image Input*:* Output Design:**
- **Prediction Results:**
  - **For Structured Data:** Disease likelihood (e.g., "85% likelihood ofvDiabetes").
  - **For Image Data:** Disease detection with confidence (e.g., "Pneumonia detected with 92% confidence").
  - Recommendations (e.g., "Consult a healthcare provider").
- **User Interface:**
  - Simple forms for input.
  - Clear, color-coded prediction results (e.g., high risk in red, low risk in green).
  - Optional visual output like charts for prediction probabilities.

This design ensures quick, easy data input and clear, actionable prediction results for users.

VII. SYSTEM DESIGN

System Architecture- The architecture diagram for the different diseases prediction web application

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

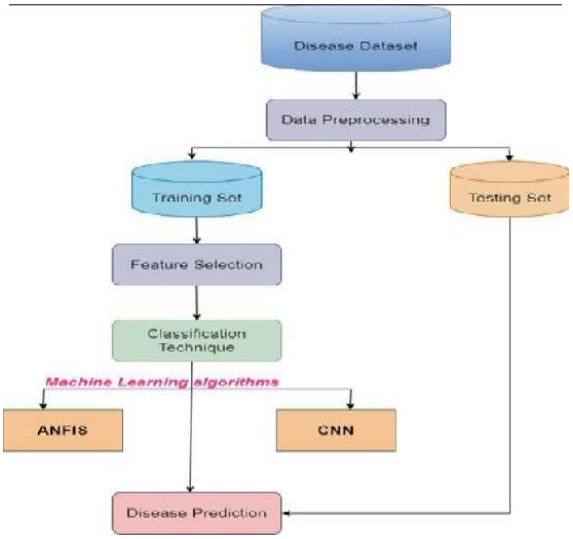

Fig 1: System Design

VIII. RESULTS

The results for all the ML(machine learning) models and of final completed project are shown in the below following figures and tables:

The deep learning models for Malaria and Pneumonia, based on CNN architectures, achieved high accuracy levels due to the large dataset sizes and image-based features. Machine learning models also showed high accuracy, particularly for kidney disease and diabetes prediction.

SN.	Disease Name	Algorithm Name	Proposed system accuracy
1	Diabetes	Random Forest	78%
2	Heart disease	XGBoost	86.96%
3	Pneumonia	CNN	83.17%
4	Brain Tumor	CNN	95%
5	Breast cancer	Random Forest	96%
6	Covid 19	CNN	93%
7	Alzheimer disease	CNN	73%

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

IX. FUTURE SCOPE

Future research should focus on improving the accuracy and robustness of models, particularly for diseases with lower prediction accuracy, such as liver disease. Further enhancements could include:

- **Ensemble Learning**: Combining different machine learning and deep learning models in an ensemble could improve prediction performance, especially for complex diseases.
- **Real-Time Data Integration**: Integrating real-time data from medical devices or patient wearables (such as continuous glucose monitors for diabetes patients) could enhance the app's utility for ongoing health monitoring.
- **Expansion of Disease Coverage**: Future versions of the web app could incorporate more diseases especially those of significant public health concern, such as COVID-19 or mental health conditions, using natural language processing (NLP) for symptom analysis.
- **Cloud Deployment**: Deploying the app on cloud platforms like AWS or Google Cloud could increase its scalability and accessibility, allowing the app to handle large-scale usage across the globe.

Future Scope of the Multiple Disease Prediction System

1. Expansion to More Diseases:

The system can be extended to predict additional diseases by integrating more datasets and

developing new models for conditions such as

Alzheimer’s, Parkinson’s, and various cancers.

2. SS

The system could integrate with wearable health devices (e.g., smartwatches, fitness trackers) to collect real-time data like heart rate, oxygen levels, and physical activity, enabling.

[2] Kaggle: continuous health monitoring and timely disease prediction.

3. Telemedicine Integration:

Incorporating the platform with telemedicine services would allow users to instantly consult with healthcare professionals based on the prediction results, enabling faster diagnosis and treatment.

4. Personalized Health Insights:

Future versions of the system could use advanced analytics and personalized recommendations

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

based on users' historical health data, lifestyle habits, and family medical history to provide more tailored disease risk assessments.

5. **Mobile App Development:**

Developing a mobile app version of the platform would improve accessibility, allowing users to monitor their health and predict diseases on the go, providing an even more user-friendly experience.

6. **Improvement in Prediction Models:**

Ongoing research into advanced ML/DL algorithms could lead to even more accurate and faster disease predictions. Techniques like ensemble learning, reinforcement learning, and better image processing methods can further improve model performance.

7. **Multilingual Support:**

Adding support for multiple languages would increase the system's usability across different regions and demographics, making it accessible to non- English speaking populations.

8. **Integration with Electronic Health Records (EHR):**

The system can be integrated with hospital EHR systems, enabling automated prediction based on patient records, thus aiding doctors in decision-making and enhancing clinical workflow.

9. **Predictive Preventive Healthcare:**

The system can evolve to not only predict diseases but also offer preventive measures based on real-time data analytics, promoting a preventive healthcare model that encourages healthier lifestyles.

REFERENCES

1. Harleen Kaur, Siri Krishan Wasan, and Amit Goel. "An Empirical Study on Applications of Data Mining Techniques in Healthcare." Journal of Computer Science, 2006.
2. Yan, Ke, et al. "Deep Learning for Malaria Detection in Thin Blood Smear Images." IEEE InternationalConference on Image Processing (ICIP), 2018.
3. Ramesh, A. N., et al. "Artificial intelligence in medicine." Annals of the Royal College of Surgeons of England, 86.5 (2004): 334-338.
4. Ravi, D., et al. "Deep Learning for Health Informatics." IEEE Journal of Biomedical and HealthInformatics, 2017.
5. Patel, Shyamal, et al. "A review of wearable sensors and systems with application in rehabilitation." Journal of NeuroEngineering and Rehabilitation, 2012.
6. Keesara, Sirina, Andrea Jonas, and Kevin Schulman. "COVID-19 and health care's digital

Power System Technology
ISSN:1000-3673
Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

revolution."The New England Journal of Medicine, 2020.

7. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
8. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
9. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.