Health Prediction System

Mr. Pradip Bhakare[1], Aditya Arabt[2], Ayush Ingel[3], Ankush Kharate[4], Nikhli Chitode [5]

[1]*Asst.Professor, Department of ENTC, MGICOET Shegaon, Maharashtra.*
[2,3,4,5] *Students, Department of ENTC, MGICOET Shegaon, Maharashtra.*

***

Abstract - Chronic diseases like diabetes, heart disease, breast cancer, and Parkinson's remain major global health challenges. Early detection is crucial to reducing long-term health risks. This research introduces a Health Prediction System that combines machine learning (ML) with web technologies to deliver real-time disease risk assessments. Users can either manually input clinical parameters or upload medical reports, which are processed using Optical Character Recognition (OCR) techniques. The system features a modular design with four dedicated ML models: Support Vector Machine (SVM) for diabetes and Parkinson's, and Logistic Regression for heart disease and breast cancer. Models were trained on trusted datasets, including the PIMA Indian Diabetes dataset, UCI Cleveland Heart dataset, Wisconsin Breast Cancer dataset, and Parkinson's voice dataset. A Flask-powered backend handles input routing and prediction, achieving an average response time between 200–300 milliseconds. A dual-input mechanism enhances flexibility—users can type data manually or extract it automatically from scanned reports using Tesseract OCR and OpenCV preprocessing. The intuitive web interface, built with HTML, CSS, and JavaScript, offers immediate feedback with built-in validation to ensure data integrity. Performance evaluation, based on accuracy, F1-score, and ROC-AUC, showed that the breast cancer model achieved the highest accuracy (97.1%), while Parkinson's and heart disease models also delivered strong results. Designed for offline use, the system prioritizes user privacy and accessibility. This paper details the system's design, implementation, and broader applications in home healthcare, mobile clinics, and underserved communities, demonstrating the potential of machine learning and OCR technologies to enhance early diagnosis and health awareness.

Key Words: Health Prediction, Machine Learning, OCR, Flask, SVM, Diabetes, Breast Cancer, Parkinson's

I. INTRODUCTION

The rising prevalence of chronic diseases such as diabetes, heart disease, breast cancer, and Parkinson's disease underscores the urgent need for early risk assessment. Traditional diagnostic methods, while highly accurate, often require extensive clinical resources, specialized labs, and significant time—factors that limit accessibility, especially in under-resourced areas. This has fueled the demand for faster, technology-driven alternatives that can deliver reliable health insights in a more accessible and affordable manner.

Recent progress in machine learning (ML) has opened new possibilities for health prediction by analyzing structured and unstructured medical data. Coupled with tools like Optical Character Recognition (OCR), modern systems can now extract, process, and interpret health parameters from scanned documents or manual entries with minimal user effort. Such integration offers a practical route to decentralized healthcare, empowering users with instant health assessments outside traditional clinical settings. This research introduces a Health Prediction System designed to bridge machine learning intelligence with personalized healthcare. The platform enables users to assess their risk for four major diseases Type 2 Diabetes, Heart Disease, Breast Cancer, and Parkinson's via a user-friendly web interface. Inputs can either be entered manually or extracted from medical reports using OCR and computer vision techniques.

Dedicated ML models are trained for each disease, leveraging benchmark datasets like the PIMA Indian Diabetes dataset, UCI Heart Disease dataset, Wisconsin Breast Cancer dataset, and the Parkinson's voice dataset. Models were selected based on proven classification performance, with Support Vector Machine (SVM) and Logistic Regression used for their efficiency, interpretability, and reliability in binary prediction tasks.

Built as a lightweight, offline-capable system, the platform suits both personal use and broader applications such as field deployment or telemedicine services. It also maintains transparency through detailed evaluation metrics, including confusion matrices, ROC curves, and feature correlation heatmaps, ensuring trust and usability in real-world healthcare scenarios.

This paper outlines the complete lifecycle of the system from design to deployment while emphasizing its role in advancing early diagnosis, preventive care, and health empowerment.

II. Motivation

Chronic diseases such as diabetes, heart disease, breast cancer, and Parkinson's disease continue to impose a significant global health burden, especially in low-resource regions where early diagnosis and continuous medical monitoring are often limited or inaccessible. While medical science has made great strides in treating these conditions, the challenge of early detection remains critical. Delayed diagnosis not only worsens patient outcomes but also increases healthcare costs, limits treatment effectiveness, and diminishes quality of life.

With the increasing availability of structured healthcare data and advancements in machine learning, there lies an opportunity to develop systems that can offer timely, personalized health assessments. Several studies have shown that predictive models, when properly trained and validated, can perform on par with traditional diagnostic methods in detecting early signs of disease. However, existing machine learning applications in healthcare often require high-end infrastructure, assume technical expertise, or lack accessibility for general users.

This project was born from the need to bridge this gap to create a lightweight, accurate, and user-friendly prediction system that empowers individuals to screen for multiple health conditions in real-time. The motivation was to democratize health monitoring by making it accessible even to non-technical users and those without immediate access to clinical facilities.

A unique aspect of this system is its dual input mechanism, allowing users to either manually enter data or upload lab reports that are processed using Optical Character Recognition (OCR). This flexibility reduces barriers for users who may not be familiar with digital forms or may only have physical medical documents. Additionally, choosing interpretable models like Logistic Regression and SVM ensures that predictions are not only accurate but also explainable a critical factor in healthcare decision-making.

III. METHODOLOGY

The methodology behind this project follows a systematic approach, combining well-established data science practices with practical software engineering. The goal was to develop a multi-disease prediction system that is both accurate and accessible, while supporting dual input modes manual entry and medical image upload. This section describes the steps involved, from dataset preparation to machine learning model integration and evaluation.

System Flow for Prediction of Heart Disease

The disease prediction system follows a streamlined and intelligent workflow designed to offer quick, accurate, and user-friendly health assessments. It begins with users either entering their medical data manually through an online form or uploading an image of their health report. In the case of image input, the system uses OCR technology (Tesseract) to automatically extract relevant clinical values from the image. Once the data is collected, it goes through preprocessing steps—such as normalization, cleaning, and formatting—to ensure compatibility with machine learning models.

The user then selects the disease they wish to screen for, including options like Diabetes, Heart Disease, Breast Cancer, or Parkinson's. Based on this selection, the appropriate pre-trained model (e.g., SVM or Logistic Regression) is loaded from local storage. The input data is then transformed into a numerical vector and passed through the model for prediction. The outcome, which could indicate either low or high risk, is interpreted into a human-readable message and displayed instantly on the web interface. This entire process is handled in real time, offering a smooth and responsive experience, even on low-resource systems.

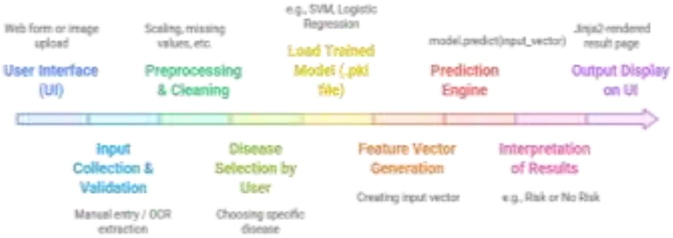
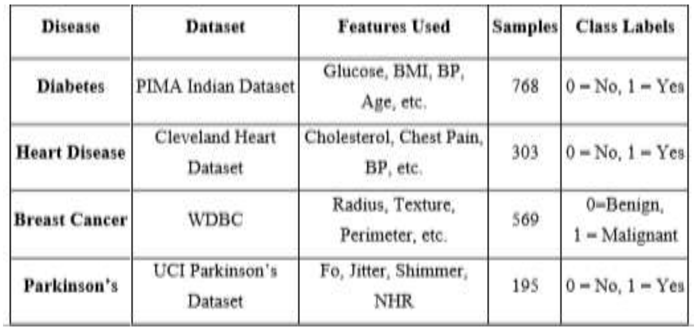

Web form or image upload
User Interface (UI)
Scaling, missing values, etc.
Preprocessing & Cleaning
e.g., SVM, Logistic Regression
Load Trained Model (.pkl file)
model.predict(input_vector)
Prediction Engine
Jinja2-rendered result page
Output Display on UI
Input Collection & Validation
Manual entry / OCR extraction
Disease Selection by User
Choosing specific disease
Feature Vector Generation
Creating input vector
Interpretation of Results
e.g., Risk or No Risk

Fig 3.1: System Flow for Prediction of Heart Disease

Dataset Preparation

This project utilized four well-established datasets to train and evaluate the health prediction models. For diabetes, the PIMA Indian Diabetes dataset was chosen due to its inclusion of key features like glucose, BMI, and age. The Cleveland Heart Disease dataset, commonly used for cardiac prediction tasks, provided clinical attributes such as chest pain type and cholesterol. Breast cancer prediction relied on the Wisconsin Diagnostic dataset, which includes 30 numerical features describing cellular structure. Lastly, the UCI Parkinson's Disease dataset offered vocal signal-based features, including jitter, shimmer, and fundamental frequency.

Each dataset was pre-analyzed for feature distribution, class balance, and formatting. The selection of these datasets ensured diversity in data types and high relevance for early disease detection. Their widespread use in published studies further validated their reliability for building accurate and generalizable machine learning models.



Disease	Dataset	Features Used	Samples	Class Labels
Diabetes	PIMA Indian Dataset	Glucose, BMI, BP, Age, etc.	768	0 – No, 1 – Yes
Heart Disease	Cleveland Heart Dataset	Cholesterol, Chest Pain, BP, etc.	303	0 – No, 1 – Yes
Breast Cancer	WDBC	Radius, Texture, Perimeter, etc.	569	0=Benign, 1 – Malignant
Parkinson's	UCI Parkinson's Dataset	Fo, Jitter, Shimmer, NHR	195	0 – No, 1 – Yes

Table 3.1: Classification of Dataset

Data Processing

Effective data preprocessing was essential to ensure the reliability and accuracy of the predictive models used in this system. Each dataset underwent a series of standardized preprocessing steps to improve data quality, handle inconsistencies, and prepare the features for model training.

To begin with, missing or biologically unrealistic values were addressed. In datasets like the PIMA Indian Diabetes dataset, features such as insulin levels and skin thickness occasionally recorded zeroes—values that are physiologically implausible. These were treated as missing data and replaced using the mean of the respective feature columns to avoid skewing the model.

For datasets containing categorical outputs, such as the breast cancer dataset where the diagnosis is labeled as “M” (Malignant) or “B” (Benign), label encoding was applied to convert these classes into binary numerical form (1 and 0, respectively).

Following this, feature scaling was performed using the Standard Scaler class from the scikit-learn library. This step normalized all feature values to have a mean of zero and a standard deviation of one. Standardization is particularly important for algorithms like Support Vector Machines (SVM), which are sensitive to the scale of input features.

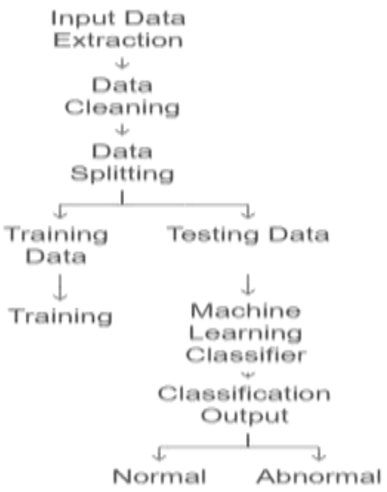


Fig.3.2: Flow Diagram Preprocessing

Lastly, all datasets were split into training and testing subsets using an 80:20 ratio. A stratified splitting approach was employed to ensure that the distribution of class labels remained consistent across both subsets. This helped prevent class imbalance and ensured that the test data fairly represented the original dataset.

By maintaining consistency in preprocessing across all datasets, the methodology supported the development of robust, fair, and accurate machine learning models tailored for each disease type.

Feature Selection

Feature selection played a major role in enhancing model performance by reducing noise and improving interpretability. For each dataset, only the most relevant features—those with strong correlation to the target variable—were retained for training. In the diabetes dataset, glucose, BMI, age, and insulin as key indicators of diabetic risk. Similarly, in the heart disease dataset, features such as chest pain type, maximum heart rate, cholesterol, and ST depression (oldpeak) were prioritized based on prior research and statistical relevance. For

To identify the most informative features, correlation heatmaps and domain-specific knowledge were combined with basic statistical analysis. Irrelevant or redundant attributes were excluded to prevent overfitting and to streamline the input data. This selective approach not only improved model accuracy but also reduced computational complexity, ensuring faster and more reliable predictions.

Model Selection And Training

To ensure optimal performance and interpretability, different machine learning models were chosen based on the characteristics of each disease-specific dataset. For diabetes and Parkinson’s disease, the Support Vector Machine (SVM) algorithm was selected due to its strong performance with small datasets and its ability to handle non-linear patterns and high-dimensional features effectively. In contrast, Logistic Regression was used for heart disease and breast cancer prediction, given its simplicity, transparency, and proven accuracy with structured clinical data. These models were implemented using the scikit-learn library in Python. Once trained and validated, each model was serialized using the joblib module and stored as a .pkl file. This made them easy to load into the Flask backend for real-time inference without the need for retraining.

OCR and Image Input Workflow

A key innovation in this system is its support for image-based input, allowing users to upload scanned medical reports or capture them using a webcam. These images are processed using OpenCV to enhance

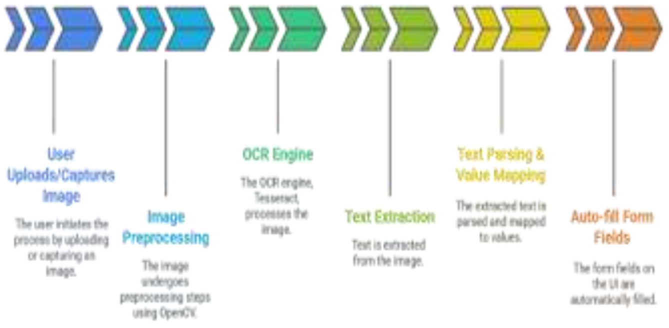

User Uploads/Captures Image
The user initiates the process by uploading or capturing an image.
Image Preprocessing
The image undergoes preprocessing steps using OpenCV.
OCR Engine
The OCR engine, Tesseract, processes the image.
Text Extraction
Text is extracted from the image.
Text Parsing & Value Mapping
The extracted text is parsed and mapped to values.
Auto-fill Form Fields
The form fields on the UI are automatically filled.

Fig.3.3: OCR Workflow for Image-Based Input

readability: they are first converted to grayscale, then smoothed using Gaussian blur to reduce noise, and finally thresholder to improve contrast. Resizing is applied to standardize dimensions for consistent OCR performance. The pre-processed image is then passed to Tesseract OCR, which extracts relevant text. To improve accuracy, configuration settings are optimized for numeric and tabular data commonly found in medical documents. Once the text is extracted, regular expressions are used to identify and map values such as glucose levels, cholesterol, or heart rate to their corresponding form fields. To ensure user control and minimize errors, the system allows users to manually review and edit the auto-filled values before proceeding with prediction. This dual input capability makes the platform both flexible and user-friendly.

Model Integration and Prediction

Once the machine learning models were trained and validated, they were integrated into the system using a Flask-based backend. Each model was serialized into a .pkl file using the joblib library to facilitate easy loading during runtime without the need for retraining. Based on the disease selected by the user, the backend dynamically loads the corresponding model and applies the same preprocessing steps—such as scaling and formatting—that were used during training.

The prediction workflow begins when a user submits data, either manually or via OCR. The input is routed through Flask, where it is first validated and then transformed into a suitable format for model inference. After preprocessing, the model generates a binary output indicating whether the disease is likely present (1) or not (0). This result is then returned to the user interface and displayed in real-time, typically within a second of form submission.

This modular approach ensures that each model functions independently while maintaining consistency in the overall workflow. The design also allows for easy expansion — additional models for other diseases can be integrated with minimal adjustments to the existing architecture.

Model Evaluation

To assess the predictive performance of each disease-specific model, a combination of standard evaluation metrics was employed, including Accuracy, Precision, Recall, F1 Score, and ROC-AUC. These metrics offer both overall and class-specific insight into the model's ability to generalize to unseen data. Precision reflects the percentage of true positive predictions among all positive predictions, while Recall captures how effectively the model identifies actual positive cases. The F1 Score, being the harmonic mean of Precision and Recall, provides a balanced measure for imbalanced data scenarios. ROC-AUC scores further supported binary classification quality by quantifying the model's ability to distinguish between classes across thresholds. The following table summarizes the performance of each model based on these metrics:

Table 3.2: Model Overview

Disease	Model Used	Accuracy (%)	F1 Score	Precision	Recall
Diabetes	Support Vector Machine (SVM)	77.9	0.78	0.75	0.81
Heart Disease	Logistic Regression	85.2	0.85	0.83	0.86
Breast Cancer	Logistic Regression	97.1	0.97	0.96	0.98
Parkinson's	SVM	90.5	0.91	0.90	0.92

These results indicate that all models performed well within their respective domains, with the breast cancer model demonstrating the highest accuracy and F1 Score. The Parkinson's prediction model also showed robust generalization, while the diabetes and heart disease models maintained solid performance with room for further optimization.

IV. RESULT ANALYSIS

Fig 4.1: VS_Code Interface Showing Python program of Health Prediction System

Confusion Matrices (for Each Disease Model)

Confusion matrices were used to evaluate each machine learning model's ability to distinguish between disease-positive and disease-negative cases. These matrices provide crucial insights into true positives, false positives, true negatives, and false negatives, which are vital in healthcare settings where misclassification can delay treatment or cause undue stress.

Diabetes(SVM):

The confusion matrix for diabetes prediction shows a moderate balance between correct and incorrect predictions. While the Support Vector Machine accurately identified most diabetic patients, some cases were misclassified either way. This suggests good generalization but highlights the difficulty in distinguishing borderline cases.

Heart Disease (Logistic Regression):

For heart disease, the confusion matrix indicates strong performance. The model correctly classified most patients with few errors, minimizing both false positives and false negatives. This balance strengthens its reliability for clinical screening purposes.

Breast Cancer (Logistic Regression):

The breast cancer model demonstrated near-perfect classification. Few cases were misclassified, reflecting high precision and recall, making it highly suitable for early tumor detection where accuracy is critical.

Parkinson's Disease(SVM):

The Parkinson's model showed excellent sensitivity, with minimal false negatives. Although slightly more false positives were observed, overall reliability remained high critical for detecting early-stage neurological symptoms.

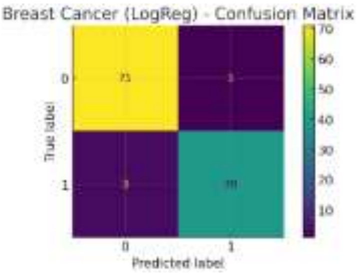
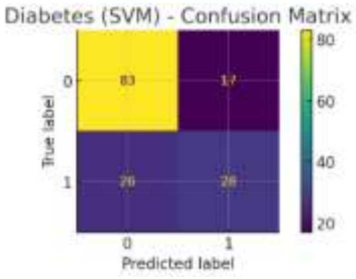


Fig 4.2: Confusion Matrix of Breast Cancer



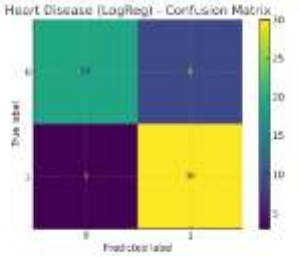


Fig 4.3: Confusion Matrix of Diabetes

Fig 4.4: Confusion Matrix of Heart Disease

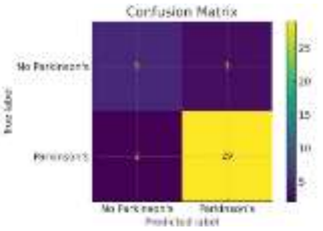


Fig 4.5: Confusion Matrix of Parkinson's Disease

ROC Curves and AUC Comparison

The Receiver Operating Characteristic (ROC) curve is a vital tool for evaluating a model's ability to distinguish between classes. A curve closer to the top-left corner indicates better performance, while the Area Under the Curve (AUC) provides a single value summarizing overall classification ability. In healthcare prediction, ROC analysis is critical, as both false positives and false negatives can have significant impacts.

Diabetes (SVM):

The ROC curve for diabetes prediction yielded an AUC of approximately 0.84, indicating good separation between diabetic and non-diabetic cases. While not perfect, the model demonstrated balanced and reasonable diagnostic capability, especially given overlapping feature patterns.

Heart Disease (Logistic Regression):

The heart disease model achieved an AUC of around 0.89. The steep ROC curve reflects strong discriminative power, ensuring reliable identification of high-risk individuals with minimal misclassification.

Breast Cancer (Logistic Regression):

This model achieved an outstanding AUC of **0.98**, with the ROC curve closely hugging the top-left corner. It reflects the model's near-perfect ability to differentiate between malignant and benign tumors.

Parkinson's Disease (SVM):

The Parkinson's model produced an AUC of approximately **0.95**, with a sharp ROC curve suggesting excellent sensitivity to early-stage neurological symptoms based on subtle vocal feature variations.

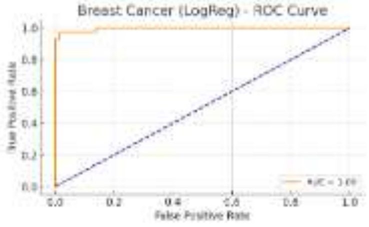


Fig 4.6: ROC Curve of Breast Cancer

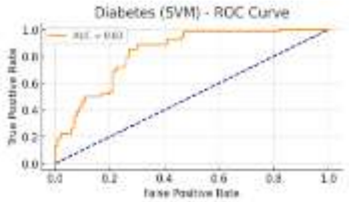


Fig 4.7: ROC Curve of Diabetes

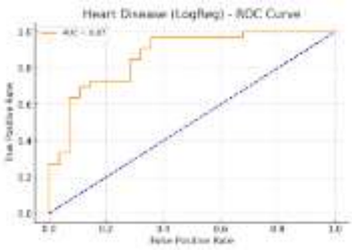


Fig 4.8. ROC Curve of Heart Disease

Feature Correlation Heatmaps

Feature Feature correlation heatmaps were used to visualize how input variables relate to each other and the target outcome. They helped identify key clinical features, minimize redundancy, and guide effective model training.

Diabetes Dataset:

The heatmap showed glucose had the strongest correlation with diabetes, followed by BMI, age, and insulin levels. This aligned with clinical expectations, confirming these features as critical indicators. Weaker features like skin thickness and blood pressure were included but had less influence.

Heart Disease Dataset:

In heart disease prediction, chest pain type, maximum heart rate (thalach), and ST depression (oldpeak) showed moderate-to-strong correlations with disease presence. Surprisingly, cholesterol had a lower correlation, highlighting the complexity of cardiac risk factors.

Breast Cancer Dataset:

Strong internal correlations were observed among radius mean, perimeter mean, and area mean, reflecting tumor size and geometry. The heatmap also helped eliminate redundant metrics, streamlining model input without sacrificing predictive power.

Parkinson's Dataset:

In Parkinson's data, jitter, shimmer, and HNR emerged as highly correlated features. These vocal measures captured subtle frequency and amplitude variations linked to motor control decline, justifying their key role in disease prediction.

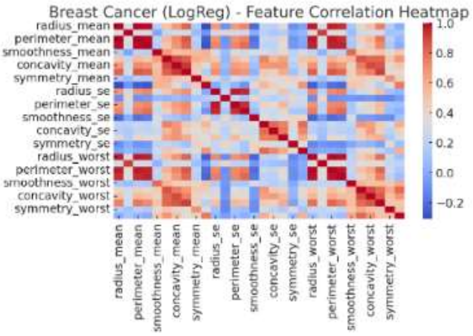


Fig 4.9: Feature Correlation Heatmap of Breast Cancer

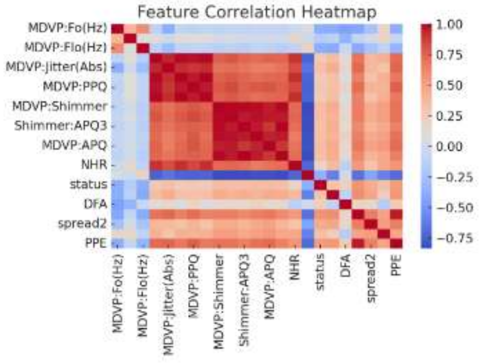


Fig 4.10: Feature Correlation Heatmap of Parkinson's

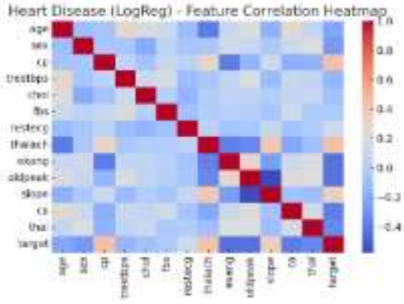


Fig 4.11: Feature Correlation Heatmap of Heart Disease

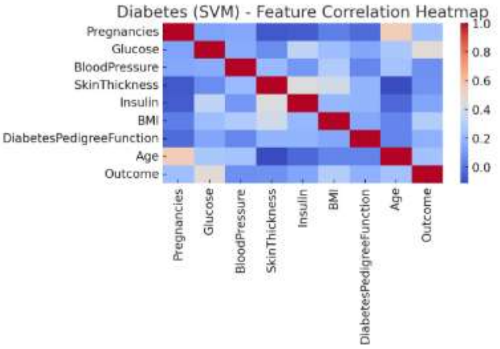


Fig 4.12: Feature Correlation Heatmap of Diabetes

V. CONCLUSION

The increasing prevalence of chronic diseases like diabetes, heart disease, breast cancer, and Parkinson's highlights the urgent need for early, accessible, and reliable health screening tools. However, real-world challenges—such as delayed diagnosis in rural areas, dependence on manual data entry, and fragmented, single-disease systems—limit the effectiveness of traditional methods. This project addresses those challenges by developing a unified, real-time Health Prediction System that integrates machine learning models with optical character recognition (OCR), allowing users to input data manually or extract it directly from medical reports. Built using a lightweight Flask backend and interpretable ML algorithms like SVM and Logistic Regression, the system delivers instant predictions for multiple diseases through a clean, browser-based interface. By balancing accuracy with usability and automation with flexibility, this solution brings predictive healthcare a step closer to real-world application. It empowers individuals, supports frontline health workers, and opens the door to wider adoption of AI in health without compromising simplicity or reliability

REFERENCES

[1] Majhi, B., Kashyap, A., Mohanty, S. S., Dash, S., Mallik, S., Li, A., & Zhao, Z. (2024). An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm. *BMC Medical Imaging*, *24*(156). https://doi.org/10.1186/s12880-024-01335-z

[2] Govindu, A., Kumar, R., & Sharma, R. (2023). A hybrid approach using PCA and logistic regression for Parkinson's disease detection. *Procedia Computer Science*, *218*, 249–261 https://doi.org/10.1016/j.procs.2023.04.031

[3] Journal of Engineering Sciences. (2023). Heart disease prediction using machine learning. *Journal of Engineering Sciences*, *14*(4), 447. https://jespublication.com/issues/2023041212.pdf

[4] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012072. https://doi.org/10.1088/1757-899X/1022/1/012072

[5] Moumita, P., Pradhan, R., Dey, N., & Gupta, P. (2021). Biomarkers for detection of Parkinson's disease using machine learning—a short review. In *Soft Computing Techniques and Applications* (pp. 461–475). Springer. https://doi.org/10.1007/978-981-15-7394-1_43

[6] Nilashi, M., et al. (2020). Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. *Expert Systems with Applications*, *159*, 113562. https://doi.org/10.1016/j.eswa.2020.113562

[7] Singh, S., & Xu, W. (2020). Robust detection of Parkinson's disease using harvested smartphone voice data: A telemedicine approach. *Telemedicine and e-Health*, *26*, 327–334. https://doi.org/10.1089/tmj.2018.0271

[8]] X. Yang, Q. Ye, G. Cai, Y. Wang and G. Cai, (2022), "PD-ResNet for Classification of Parkinson's Disease from Gait," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-11, 2022, Art no. 2200111, 10.1109/JTEHM.2022.3180933

[9] A. U. Haq et al., "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings," in IEEE Access, vol. 7, pp. 37718-37734, 2019, doi: 10.1109/ACCESS.2019.2906350.

[10] Mei Jie, Desrosiers Christian, Frasnelli Johannes, (2021), "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature", in Frontiers in Aging Neuroscience, vol. 13, doi: 10.3389/fnagi.2021.633752.

[11] Amalia Luque, Alejandro Carrasco, Alejandro Martín, Ana de las Heras, (2019), "the impact of class imbalance in classification performance metrics based on the binary confusion matrix", Pattern Recognition, Volume 91, Pages 216-231, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.02.023.

[12] J. R. Barr, M. Sobel and T. Thatcher (2022), "Upsampling, a comparative study with new ideas," 2022 IEEE 16th International Conference on Semantic Computing (ICSC), pp. 318-321, doi: 10.1109/ICSC52841.2022.00059.

[13] Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS. Interob server variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Springer. 2013;22(2):153–60.

[14] Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020;. https ://doi.org/10.1080/03772 063.2020.17139 16.

[15] Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci. 2020;1(4):206.

[16] Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229