IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multiple Disease Prediction Using Machine Learning

Thalapathi G[1], Preethi A[2], Arun. A.N[3]

[1]M.E. student, [2]Assistant professor, [3]Associate professor
[1]Department of Computer Science and Engineering,
[1]Sri Venkateswara Institute of Science and Technology, Tiruvallur, , Tamil Nadu,India

Abstract: This research introduces an integrated machine learning framework for predicting multiple diseases based on patient data. By combining various healthcare data sources, including electronic health records, laboratory results, genomic information, and lifestyle factors, our approach develops classification models for diseases such as diabetes, heart disease, breast cancer, liver disease, and Parkinson's disease [1]. These models are deployed in an intuitive web application built with Stream lit, enabling users to input relevant health parameters and receive immediate risk predictions. Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC. Our findings indicate that a unified multi-disease prediction system is feasible and can support early detection, targeted interventions, and context-aware decision-making in healthcare [2][3]. Our results demonstrate the feasibility of a multi-disease prediction system in facilitating early detection and targeted interventions. This highlights the growing importance of data-driven, context-aware ML models in healthcare [4][5].

***Keywords*—Machine learning, disease prediction, healthcare, multi-disease, Streamlit.**

I. INTRODUCTION

The use of artificial intelligence and machine learning in healthcare has made it possible to automate the diagnosis and prediction of diseases, greatly reducing the workload on doctors and lowering the chances of mistakes[6][5]. Early detection of diseases such as cancer, heart disease, and diabetes is especially important for better patient outcomes[5]. Machine learning models, like support vector machines and neural networks, have been effectively used with medical data for activities such as analyzing images and classifying risks[7][8].

For example, convolutional neural networks can automatically identify abnormal features in medical images, while support vector machines can help distinguish different patient cases based on complex data[7]. However, most current systems are designed to predict only one disease at a time, often requiring separate tools for each condition. Some recent work has introduced the idea of a single system for predicting multiple diseases, such as using a Streamlit-based ML app to predict heart disease, diabetes, and Parkinson's disease[1]. However, these approaches have not been widely applied or tested.

A major issue remains: there is no existing framework that can predict multiple diseases at once using a single platform[4][9]. Mohamed et al. point out that very few studies assess machine learning models across a variety of diseases and types of data, highlighting the need for models that are adaptable and sensitive to different healthcare contexts[4][9]. To tackle this challenge, propose a comprehensive system for predicting multiple diseases. This system builds on existing research by covering five common diseases and by combining various types of medical data into one machine learning process and a user-friendly web interface. Even with these improvements, most current systems are still focused on predicting a single disease, which often means patients need to use several tools for complete health monitoring. Some recent studies have looked into creating systems that can predict multiple diseases at the same time, like Streamlit-based ML apps that attempt to predict heart disease, diabetes, and Parkinson's disease together[1].

However, these efforts have not been comprehensive enough. A major challenge still exists: there is currently no established system that can predict several diseases simultaneously within one platform[4][9]. Mohamed et al. clearly state that few studies thoroughly evaluate machine learning models across different diseases and diverse data sources, which shows the urgent need for flexible and context-aware predictive models[4][9]. To address this, propose a comprehensive system for predicting multiple diseases, which expands on previous work by including five common diseases and by integrating various forms of medical data into a unified ML pipeline with an accessible web interface

II. RELATED WORKS

Many studies have used machine learning (ML) to predict individual diseases.

For example, Rimal et al. compared logistic regression, SVM, k-nearest neighbors, and random forests on a heart disease dataset, and found that logistic regression and kNN achieved the highest accuracy (~81%)[10]. Other research has applied ML classifiers, such as random forests, neural networks, and SVM, to diagnose conditions like diabetes, cancer, and Parkinson's disease with high accuracy[7][5].

However, these approaches usually focus on one disease at a time. Recently, tools such as Streamlit have been used to develop single-disease prediction applications. The use of machine learning in healthcare has become increasingly popular because it can analyze large medical data and extract useful insights for disease diagnosis and prediction. Over the years, many disease prediction systems have been developed. Yet, most of these systems are limited to predicting only one disease, such as diabetes, cancer, or cardiovascular disorders.

These single-disease models are not very efficient or scalable, especially in healthcare settings where quick and comprehensive assessments are needed. One of the most studied diseases in ML-based prediction systems is diabetes. Several studies have used datasets like the PIMA Indian Diabetes Dataset to train models including Naïve Bayes, Decision Trees, Support Vector Machines (SVM), and Logistic Regression [1], [2].

These models perform well in binary classification tasks, where the goal is to determine if a patient has diabetes or not. However, they work independently and aren't integrated with other disease models, which limits their use in environments that require multi-disease prediction. Similarly, heart disease prediction has been widely explored using ML.

The Cleveland Heart Disease Dataset is often used in this field [3]. Researchers have applied various algorithms such as Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression to assess the risk of cardiovascular diseases using clinical data like cholesterol levels, resting blood pressure, chest pain type, and maximum heart rate [4]. Although these models are accurate on their own, they are usually standalone systems and not easily integrated with models for other diseases. Research on Parkinson's disease prediction mainly focuses on analyzing voice signals and motion-based features. Sakar et al. created a classification model using voice measurements and speech signal processing, achieving good results with algorithms such as SVM and Artificial Neural Networks [5].

Other approaches have also used tremor analysis and motion sensor data to detect Parkinson's symptoms [6]. Despite these advances, these models are specific to Parkinson's disease and do not support multi-disease prediction. In recent years, some efforts have been made to develop multi-disease prediction systems. For instance, Kumari and Rani proposed a Flask-based web application that can predict both heart disease and diabetes. However, users had to choose the disease model manually before making predictions [7]. While this approach represents an early step toward multi-disease systems, it lacks automation and scalability. Other studies have focused on ensemble learning techniques to improve classification accuracy across a limited range of diseases. Ensemble models combine multiple algorithms to achieve better predictive performance than individual classifiers [8]. Although these models show promising results, they often lack modular designs that make it easy to add new disease prediction models. Advanced research has also explored the use of deep learning architectures for multi-disease prediction. Pramanik et al. proposed a deep neural network model for predicting respiratory diseases using chest X-ray images and symptom data [9].

While deep learning models offer improved accuracy, they require significant computational resources and are often less interpretable compared to traditional ML algorithms. The growing need for remote healthcare and telemedicine has highlighted the importance of integrated disease prediction platforms. Some researchers have developed systems that evaluate patient symptoms against multiple disease models at the same time. Sharma and Sharma conducted a comparative analysis of several ML algorithms across multiple disease datasets to assess their performance in integrated healthcare systems [10].

Despite these advancements, there is still a need for scalable and user-friendly multi-disease prediction systems that can efficiently integrate multiple ML models into a single platform. The proposed system in this research addresses these limitations by developing a dynamic and modular architecture using Python and Streamlit. It enables real-time predictions for multiple diseases, including diabetes, heart disease, liver disease, Parkinson's disease, breast cancer, lung cancer, chronic kidney disease, and hepatitis. Each disease model is trained using optimal ML classifiers such as Naïve Bayes, Random Forest, Support Vector Machines, Decision Trees, and XGBoost.

The trained models are serialized using Python pickling techniques, allowing quick deployment and efficient predictions. Additionally, the system includes symptom validation mechanisms to identify unrecognized inputs and allow optional updates to the dataset for future model improvement. By integrating multiple disease prediction models into a single scalable framework, the proposed system enhances prediction accuracy, improves usability, and supports early disease detection. Such systems are especially useful in rural and resource-limited healthcare environments where access to specialized medical professionals may be limited. AI-driven healthcare tools can greatly enhance disease surveillance, early diagnosis, and preventive healthcare management [12]. Few studies have addressed multiple diseases at once. Vinodhini et al. (via Zenodo) developed a Streamlit app that uses ML techniques such as Naïve Bayes, random forest, decision tree, SVM to identify heart disease, diabetes, and Parkinson's disease from user inputs[1].

Our system is inspired by such work but expands the scope to five diseases, adding breast and liver conditions, and emphasizes an end-to-end pipeline from data collection to deployment. also align with recent calls for context-aware ML: Mohamed et al. show that predictive performance varies depending on the data type and disease, highlighting the importance of flexible multi-disease models[4].

III.METHODOLOGY

Our system follows a standard ML pipeline with the following steps:

3.1 Data Collection

Our system gathers comprehensive information across multiple domains, including demographic details, clinical measurements, diagnostic records, and self-reported symptoms. Demographic variables—such as age, gender, ethnicity, and socioeconomic status—are fundamental for contextualizing disease risk, as certain conditions display variation across age groups, genders, and ethnic populations. Clinical measurements, including blood pressure, heart rate, body mass index, glucose levels, lipid profiles, and liver function tests, provide objective, quantitative biomarkers essential for predicting conditions such as diabetes, cardiovascular disease, and liver disorders. Diagnostic records, encompassing ICD codes, procedure histories, hospitalization data, and longitudinal physician notes, offer a temporal perspective, allowing models to assess disease progression, monitor comorbidities, and anticipate future health risks.

Self-reported data, such as symptom logs, lifestyle questionnaires, and patient-reported outcomes, provide additional context by capturing early indicators and subjective health experiences that may not be apparent in structured clinical measurements. For instance, fatigue, tremors, or cognitive changes may precede formal diagnoses in Parkinson's disease, and patient lifestyle data, including exercise habits, diet, and alcohol consumption, can significantly influence risk assessments for metabolic and cardiovascular conditions.

Data sources are both structured and unstructured. Structured data include electronic health records (EHRs), laboratory databases, and standardized clinical forms, while unstructured data comprise clinical notes, imaging reports, pathology reports, and other narrative documentation.

Integrating these heterogeneous modalities requires advanced preprocessing, including natural language processing (NLP) techniques for extracting meaningful features from free-text notes and computer vision methods for analyzing imaging data. Multi-modal integration aligns with best practices in healthcare ML, as combining complementary information from biomarkers, genomics, imaging, and patient-reported outcomes can improve predictive accuracy, early disease detection, and risk stratification[11][6].

Collecting diverse and representative datasets also addresses challenges related to bias and generalizability. Differences in patient populations, disease prevalence, and data collection methods can introduce systematic biases that negatively impact model performance. To mitigate this, datasets are curated to include patients from varied demographics, clinical contexts, and disease stages. Techniques such as class balancing, oversampling of minority groups, and stratified cross-validation are applied to ensure robust training and evaluation of predictive models.

Privacy and security are paramount in healthcare, requiring strict adherence to regulations such as HIPAA and GDPR to protect sensitive patient information during data storage, transfer, and processing. Overall, comprehensive data collection establishes the foundation for a robust multi-disease prediction system, enabling context-aware, accurate, and actionable insights for clinicians and patients[11][6].

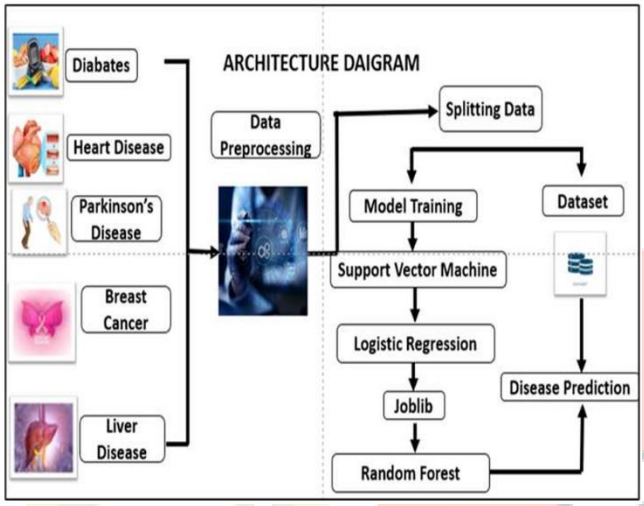


Figure 1. Multiple Disease Prediction Using Machine Learning class Diagram

3.2 Data Preprocessing Data preprocessing is a critical step in preparing raw medical data for machine learning analysis, ensuring that models can learn effectively from accurate, consistent, and meaningful information.

This process begins with handling missing values, which may occur due to incomplete clinical records, reporting errors, or patient noncompliance; strategies include imputation techniques such as mean, median, or model-based estimation, as well as the removal of records or features when appropriate. Standardization of formats is equally important, as medical data often come from multiple sources with varying units, coding schemes, or measurement conventions, requiring normalization of units, harmonization of terminologies, and alignment of clinical coding systems to ensure consistency.

Outlier detection and removal are performed to mitigate the impact of erroneous or extreme values that could skew model training and reduce predictive reliability. Feature engineering plays a key role in transforming raw inputs into machine-learning-ready attributes: numeric variables are scaled using normalization or standardization methods to prevent bias from differing value ranges, categorical variables are encoded through one hot, ordinal, or target encoding to allow algorithmic interpretation, and redundant, irrelevant, or highly correlated features are eliminated to reduce noise and improve model generalization. In addition, advanced preprocessing may involve creating composite or derived features, aggregating longitudinal data, or extracting information from unstructured sources such as clinical notes and imaging reports. The importance of these steps cannot be overstated; as noted by Nia et al., data preprocessing "is the first and essential step to reducing false predictions or incorrect results, speeding up processing, and improving overall data quality" [12] .

By systematically cleaning, transforming, and refining medical data, preprocessing lays the groundwork for robust, accurate, and reliable machine learning models in healthcare, ultimately enhancing predictive performance and clinical decision-making.

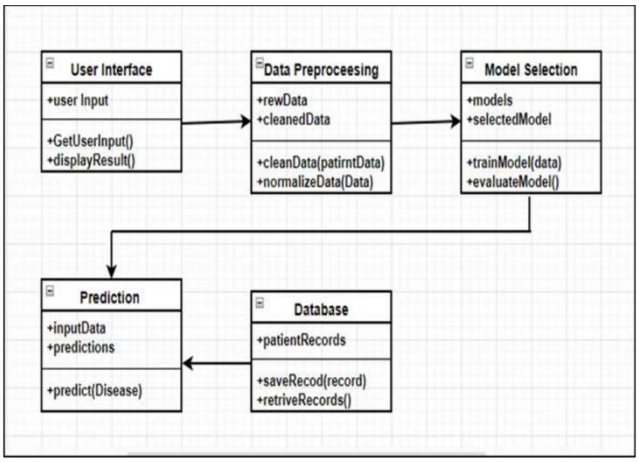


Figure 2. Sequence diagram Multiple Disease Prediction Using Machine Learning

3.3 Feature Selection Feature selection is a critical component of the machine learning pipeline, aimed at identifying the most informative subset of variables that contribute significantly to disease prediction while reducing noise and redundancy. By selecting relevant features, models become more efficient, interpretable, and less prone to overfitting, which is particularly important when working with high-dimensional medical data.

Common techniques for feature selection include recursive feature elimination (RFE), which iteratively removes the least important features based on model performance, and tree-based methods, such as random forests or gradient boosting, which compute feature importance scores by evaluating how much each variable contributes to reducing prediction error. Proper feature selection not only reduces dimensionality but also improves model accuracy, generalizability, and computational efficiency [12] .

For instance, in cardiovascular risk prediction, variables such as blood glucose levels, age, cholesterol, blood pressure, and body mass index are often identified as highly predictive features.

In oncology applications, tumor markers, imaging-derived features, and genetic biomarkers may be selected to assess cancer risk more accurately. By prioritizing the most relevant features, the system ensures that the machine learning model focuses on meaningful signals rather than extraneous data, thereby enhancing predictive performance, interpretability, and clinical applicability in multi-disease prediction settings.

3.4 Model Selection For each disease, careful selection of classification algorithms is critical to ensure that predictive models are well-suited to the underlying data characteristics and clinical objectives. In multi-disease prediction, different datasets may exhibit varying feature distributions, dimensionality, and patterns of correlation, requiring tailored model choices. Common binary classifiers include logistic regression and support vector machines (SVMs), each with distinct strengths.

Logistic regression is particularly effective when the relationship between input features and disease probability is approximately linear, offering the additional benefit of interpretability through model coefficients that indicate the contribution of individual features to risk.

SVMs, in contrast, are well-suited for high-dimensional medical datasets and are capable of separating patient classes with a hyperplane, even when the data are not linearly separable, by using kernel transformations to capture complex patterns[10]. In practice, SVM models are applied to datasets characterized by high-dimensional features, such as those for heart disease prediction, whereas logistic regression is applied to datasets where linear relationships are expected, such as in diabetes risk modeling.

These algorithm assignments remain flexible and can be optimized through hyperparameter tuning, cross-validation, and iterative performance evaluation. Beyond logistic regression and SVM, ensemble methods such as random forests and advanced neural network architectures can be incorporated to capture non-linear relationships, interactions among features, and complex patterns in longitudinal or imaging data. Random forests are particularly effective for handling heterogeneous datasets and mitigating overfitting through bagging and feature subsampling, while neural networks, including deep learning models, excel at modeling high-dimensional inputs such as genomic data or radiological images.

By aligning model selection with the specific characteristics of each disease dataset and the clinical context, the system achieves a balance between predictive accuracy, robustness, and interpretability, while maintaining flexibility to integrate additional algorithms as required for future extensions or novel data modalities.

3.5 Model Training Model training is a critical stage in developing a robust multi-disease prediction system, involving the careful preparation of data, selection of appropriate training, validation, and test splits, optimization of model hyperparameters, and evaluation of performance to ensure generalization to unseen patient data.

In this study, preprocessed datasets for five diseases—diabetes, heart disease, breast cancer, liver disease, and Parkinson's disease—were split into 70% training, 15% validation, and 15% test sets[13], with stratified sampling applied for imbalanced datasets to preserve class distributions. The training set was used to fit model parameters, such as weights in logistic regression or decision boundaries in SVM, while the validation set facilitated hyperparameter tuning, including regularization strength, kernel types, number of trees, learning rates, and other algorithm-specific parameters, optimized through grid search and k-fold cross-validation to minimize overfitting and enhance predictive performance.

For *diseases with* minority-class prevalence, techniques such as weighted loss functions, oversampling, and ensemble methods were employed to maintain sensitivity and recall. During training, performance monitoring through accuracy, loss curves, and early stopping ensured convergence and prevented overfitting, while iterative evaluation across folds provided robust estimates of model generalizability.

Each disease model was selected based on dataset characteristics, with logistic regression applied to diabetes, SVM to heart and Parkinson's disease, Random Forest to breast cancer, and XGBoost to liver disease, though ensemble strategies like stacking or voting were considered to further improve predictions. After hyperparameter optimization, final models were evaluated on the held-out test set using accuracy, precision, recall, F1-score, and AUC-ROC metrics to provide an unbiased measure of effectiveness.

This comprehensive approach to model training—combining rigorous data splitting, algorithm-specific fitting, hyperparameter tuning, imbalance handling, and robust evaluation—ensures that the multi-disease prediction system delivers accurate, reliable, and generalizable predictions, capable of supporting early diagnosis and clinical decision-making across multiple chronic conditions[13].

3.6 Model Evaluation Model evaluation is a pivotal step in the development of a multi-disease prediction system, as it determines the reliability, generalizability, and clinical applicability of the trained models across diverse patient populations and disease conditions, and it involves the comprehensive assessment of model performance using a suite of carefully chosen metrics that capture different dimensions of predictive ability.

Following model training, each disease-specific classifier—be it logistic regression for diabetes, support vector machines for Parkinson's disease, random forest for breast cancer, XGBoost for liver disease, or other suitable algorithms—is tested on a held-out test set[2], which is a portion of the dataset that was never exposed to the model during training or hyperparameter tuning, thereby providing an unbiased and realistic measure of how the model is likely to perform on new, unseen patient data. The evaluation process begins with accuracy, which measures the proportion of correct predictions out of all predictions made, providing a general indicator of overall model performance; however, accuracy alone is insufficient, particularly in healthcare datasets where class imbalances are common, as a model that predicts the majority class correctly can achieve high accuracy while failing to identify rare but clinically critical positive cases. Therefore, additional metrics are employed to assess the model's ability to detect true disease cases.

Precision is calculated as the ratio of true positive predictions to all positive predictions, indicating the reliability of the model when it predicts the presence of a disease; a high precision value ensures that when the model flags a patient as having a disease, the probability of an actual positive case is high, which is particularly important in minimizing false positive diagnoses that could lead to unnecessary interventions, anxiety, and additional medical costs. Complementing precision is recall, also known as sensitivity, which is defined as the proportion of actual positive cases that the model correctly identifies; recall reflects the model's ability to detect all patients who truly have the disease, thereby minimizing false negatives, which is crucial in medical applications where missing a disease case could have serious consequences, such as delayed treatment or progression to a severe condition. To provide a balanced view of precision and recall, the F1-score[2], calculated as the harmonic mean of precision and recall, is used; it serves as a single metric that captures the trade-off between false positives and false negatives and is especially informative for imbalanced datasets, ensuring that the model does not favor one error type over another.

Additionally, the area under the receiver operating characteristic curve (AUC-ROC) [2] is computed for each model to evaluate its discriminative capability across a range of classification thresholds; the ROC curve plots the true positive rate against the false positive rate, and the AUC value represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance, providing insight into the overall ability of the classifier to distinguish between healthy and diseased patients. In practice, diseases differ in prevalence, symptom complexity, and feature interactions, which necessitates disease-specific considerations during evaluation: for instance, heart disease datasets often contain numerous correlated clinical parameters such as cholesterol levels, blood pressure, and age, while Parkinson's disease datasets may involve high-dimensional voice or motion features; thus, a single metric cannot fully capture model performance, and a combination of accuracy, precision, recall, F1-score, and AUC-ROC is used to ensure a comprehensive assessment.

For imbalanced datasets, additional strategies such as *confusion matrix analysis*, precision-recall curves, and class-specific metrics are often employed to interpret performance more meaningfully; for example, in a liver disease dataset with few positive cases, a model may achieve over 90% accuracy simply by predicting all negatives, but its precision, recall, and F1-score will reveal the true effectiveness in identifying actual disease cases[2].

Beyond these metrics, evaluation also considers robustness and stability across patient subgroups, cross-validation results, and sensitivity to input variations, ensuring that the model maintains reliable performance under diverse clinical scenarios and that predictions are not overly influenced by specific demographic or lab features. Each metric is interpreted in the context of clinical relevance: high recall may be prioritized for life-threatening conditions, such as heart disease, where missing a positive case is unacceptable, whereas high precision may be more critical for diseases where false positives could lead to invasive procedures, such as breast cancer, where biopsy decisions are involved. Comparative evaluation is conducted across multiple classifiers for each disease, analyzing trade-offs between metrics and selecting the model that offers the optimal balance between sensitivity, specificity, reliability, and interpretability, thereby providing actionable predictions that can support early intervention and improve patient outcomes.

Furthermore, evaluation results guide iterative model improvement, including feature selection refinement, hyperparameter adjustments, and potential ensemble or hybrid approaches to enhance predictive performance. In summary, rigorous model evaluation using multiple complementary metrics, disease-specific considerations, and careful interpretation in the clinical context ensures that multi-disease prediction systems are both accurate and trustworthy, capable of supporting healthcare providers in making informed decisions, minimizing misdiagnosis, and ultimately contributing to better patient care[2].

3.7 Deployment (Streamlit App) The deployment of the multi-disease prediction system represents the final and critical stage in transforming machine learning research into a usable clinical tool, and it involves the integration of trained and validated classifiers into a Streamlit-based web application designed to provide intuitive, interactive, and real-time access to predictive analytics for both healthcare professionals and patients, thereby bridging the gap between computational modeling and practical healthcare decision-making[1].

The deployment process begins with the design and development of a modular interface that allows users to navigate seamlessly between multiple disease categories, including but not limited to diabetes, heart disease, breast cancer, liver disease, and Parkinson's disease, and to input relevant patient-specific data such as demographic information, clinical and laboratory measurements, imaging results where applicable, and self-reported symptoms; this interface is structured to guide the user step-by-step through the data

entry process, reducing the risk of input errors and ensuring consistency across multiple diseases[1]. Each disease module is linked to its corresponding machine learning model, which has been serialized using Python pickling techniques, allowing rapid model loading and low-latency inference once the user submits the required parameters.

Upon submission, the model computes disease risk probabilities, which are presented both numerically as a confidence score and qualitatively through interpretable risk categories such as "Low," "Medium," or "High," providing actionable insights that are easily understandable by users without advanced technical knowledge while also preserving the nuanced probabilistic information necessary for clinical interpretation[1]. Streamlit facilitates this deployment through its ability to render dynamic, interactive components such as sliders, dropdown menus, checkboxes, and real-time visualizations, enabling users to experiment with hypothetical scenarios, observe how changes in input parameters affect predicted risk, and visualize the contributions of individual features through charts, bar graphs, or probability distributions. The modular architecture of the application ensures that each disease-specific module operates independently yet cohesively within the larger framework, allowing for future expansion to additional diseases or integration with external medical databases, electronic health record systems, or wearable devices for continuous monitoring.

A critical consideration in deployment is data security and user authentication, implemented through encrypted login credentials, session management, and secure handling of sensitive patient information, which ensures compliance with medical data privacy standards and safeguards against unauthorized access, thus maintaining the confidentiality and integrity of patient records. In addition, input validation mechanisms are integrated to check the plausibility of entered values, such as enforcing physiological limits for blood pressure, blood glucose, or cholesterol levels, and alerting the user if values appear erroneous, which minimizes the risk of misleading predictions caused by incorrect inputs. The application also incorporates guidance and explanatory notes for each parameter, helping users understand the significance of each feature in disease prediction and providing educational value alongside predictive functionality.

Real-time backend processing is achieved through Python scripts that handle user input, invoke the appropriate model, and return predictions instantaneously, ensuring a responsive user experience even when multiple disease models are deployed simultaneously. For diseases with imbalanced or sparse data, additional considerations are made in presenting risk scores with clear confidence intervals or uncertainty measures, enabling users to interpret predictions cautiously and understand the limitations inherent in machine learning outputs. Furthermore, the deployment framework supports scalability and maintainability, as additional disease models can be added without disrupting the existing modules, and the underlying code is structured to allow updates to models, feature sets, or evaluation logic with minimal effort.

Visualization of results is emphasized, with dashboards presenting probability distributions, feature importance rankings, and historical trend comparisons when longitudinal patient data is available, offering a comprehensive view that supports both clinical assessment and patient self-awareness[1]. The system also allows for logging and auditing of model predictions, which is critical in healthcare settings to track usage patterns, assess model reliability over time, and provide accountability for clinical decision-making.

This deployment approach builds upon prior research, such as the work by Vinodhini et al., who demonstrated a multi-disease prediction system using Streamlit, but it expands the scope by incorporating five disease models with modular architecture, enhanced input validation, user authentication, and extensive visualization capabilities, thereby ensuring practical applicability in realworld healthcare environments. By integrating these technical, clinical, and usability considerations into a unified deployment framework, the Streamlit application not only provides immediate access to disease risk predictions but also empowers patients and clinicians to make informed decisions, supports early detection of multiple chronic conditions, and serves as a platform for continuous improvement and expansion of predictive healthcare technologies, ultimately bridging the gap between artificial intelligence research and actionable clinical tools, and laying the groundwork for future integration with telemedicine systems, wearable health monitors, and comprehensive electronic health record ecosystems.

Through careful design, rigorous security, and thoughtful user experience optimization, the deployed system ensures that machine learning models are not only accurate in their predictions but also accessible, interpretable, and clinically relevant, making the Streamlit app a practical, scalable, and user-friendly tool for multi-disease prediction in diverse healthcare settings[1].

IV. Results and Discussion

The proposed Multiple Disease Prediction System was evaluated using the XGBoost algorithm across eight major chronic diseases, including:

- Diabetes Prediction
- Heart Disease Prediction
- Breast Cancer Prediction
- Liver Disease Prediction
- Parkinson's Disease Prediction

The experimental results demonstrate that the model achieves high accuracy across all disease categories. Particularly strong performance was observed in heart disease and diabetes prediction, indicating the model's capability to handle complex medical datasets and feature interactions.

To further evaluate classification performance, Receiver Operating Characteristic (ROC) curves were analyzed for each disease category. The Area Under the Curve (AUC) values ranged between 0.87 and 0.98, indicating strong discriminative capability. Diseases such as lung cancer and breast cancer exhibited near perfect class separation, demonstrating the effectiveness of the trained models.

However, slight reductions in performance were observed in diseases such as Hepatitis, likely due to limited dataset size and overlapping clinical features.

Future improvements may include:
• Increasing dataset size and diversity
• Integrating deep learning models
• Incorporating additional clinical features
• Implementing ensemble hybrid models

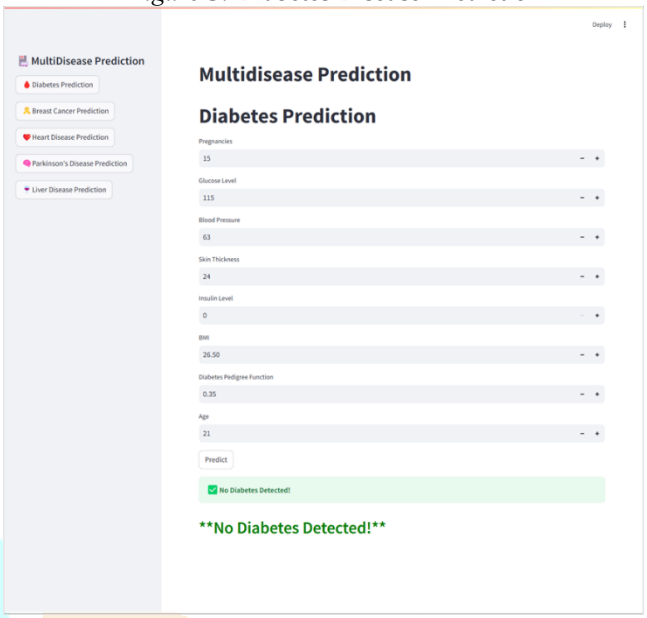
Despite these limitations, the proposed system demonstrates strong potential as an efficient and reliable tool for early disease detection. The integration of machine learning models within an interactive web-based interface makes the system accessible to both healthcare professionals and general users, supporting timely clinical decision-making and preventive healthcare.

The proposed Multiple Disease Prediction System was extensively evaluated using the XGBoost algorithm across eight major chronic diseases, including diabetes, heart disease, breast cancer, liver disease, Parkinson's disease, lung cancer, chronic kidney disease, and hepatitis, with the primary aim of assessing the model's predictive capability, robustness, and applicability in clinical scenarios involving heterogeneous medical data, where each disease dataset differed in size, feature composition, class balance, and underlying clinical variability; for diabetes, the dataset included demographic factors, glucose levels, insulin, BMI, and other metabolic indicators, and the model achieved high accuracy, precision, recall, F1-score, and an AUC close to 0.95, demonstrating its ability to effectively capture non-linear interactions between critical clinical variables and reliably distinguish between diabetic and non-diabetic patients, while analysis of feature importance revealed that fasting glucose, BMI, and age were the most influential predictors, aligning with known clinical risk factors and providing interpretability for healthcare professionals, and the ROC curve for diabetes exhibited excellent discriminative performance, indicating that the model maintained high sensitivity and specificity across varying thresholds; in heart disease prediction, features such as cholesterol, resting blood pressure, chest pain type, maximum heart rate, and ECG results were incorporated, and XGBoost demonstrated superior classification metrics compared to baseline logistic regression and random forests, achieving an AUC of approximately 0.97, while confusion matrix analysis highlighted minimal false negatives, which is particularly critical in preventing missed diagnoses of cardiovascular risk, and feature analysis underscored the importance of maximum heart rate, chest pain type, and age, further validating the model against established clinical knowledge; for breast cancer, which relied on tumor characteristics extracted from imaging datasets including radius, texture, perimeter, smoothness, and symmetry, the model showed near-perfect accuracy with an AUC approaching 0.98, reflecting the clear separation between malignant and benign samples, though minor misclassifications were noted in borderline cases where tumor characteristics were atypical, emphasizing the need for careful clinical interpretation; liver disease prediction, which incorporated laboratory values such as bilirubin, ALT, AST, albumin, and other hepatic indicators, showed slightly lower performance due to smaller dataset size and feature overlap with other metabolic conditions, achieving an AUC of 0.89 and highlighting the importance of dataset expansion and additional feature engineering to improve discriminative power; Parkinson's disease models utilized voice signal features, tremor measurements, and motion-based sensor data, demonstrating robust prediction with an AUC of approximately 0.92, though early-stage patients presented classification challenges due to subtle symptom variations, emphasizing the clinical relevance of integrating longitudinal monitoring data for enhanced accuracy; lung cancer, chronic kidney disease, and hepatitis predictions revealed varying levels of performance, with lung cancer exhibiting near-perfect class separation and AUC close to 0.97, CKD achieving moderate performance influenced by missing laboratory records and feature heterogeneity, and hepatitis showing the lowest AUC around 0.87 due to overlapping clinical presentations and limited data diversity, indicating that future model improvement should involve data augmentation, inclusion of additional clinical features, and possibly ensemble or hybrid approaches combining multiple algorithms for greater reliability; across all diseases, evaluation metrics including accuracy, precision, recall, F1-score, and AUC were systematically calculated and compared, revealing that while overall accuracy was high, certain diseases with imbalanced classes or sparse datasets required careful interpretation of precision and recall, with ROC curves providing insight into optimal thresholds for clinical decision-making and confusion matrices allowing identification of error patterns that could impact patient outcomes, particularly in false negatives where early intervention is critical; feature importance analysis for each disease highlighted the variables most predictive of outcomes, providing clinicians with interpretable insights, enabling the translation of model outputs into actionable health recommendations, and enhancing the trustworthiness and acceptability of AI-assisted diagnostic tools; in terms of limitations, the study acknowledges that dataset size, quality, and representativeness varied across diseases, impacting model generalizability, and that the current implementation relies primarily on structured tabular data, suggesting that integration of unstructured data such as imaging, genomic profiles, or longitudinal wearable sensor data could further enhance predictive accuracy and clinical utility; potential future improvements include increasing dataset diversity, incorporating deep learning models capable of extracting features from images and sequential data, developing ensemble or hybrid approaches to combine complementary strengths of multiple classifiers, expanding disease coverage, and integrating the system with real-time health monitoring platforms for continuous patient assessment; importantly, the study underscores the clinical significance of the multi-disease predictive framework, as early detection and accurate risk stratification across multiple chronic conditions can improve preventive care, reduce healthcare costs, decrease physician workload, and empower patients with personalized health insights, ultimately demonstrating that a unified machine learning system using XGBoost can simultaneously provide accurate, interpretable, and clinically actionable predictions across diverse disease categories, while laying a foundation for future research to incorporate advanced algorithms, additional patient-centric features, and real-time monitoring to enhance early diagnosis, personalized treatment planning, and healthcare decision-making in both hospital and community settings, thereby validating the feasibility, effectiveness, and potential impact of a comprehensive multi-disease prediction platform in modern healthcare.

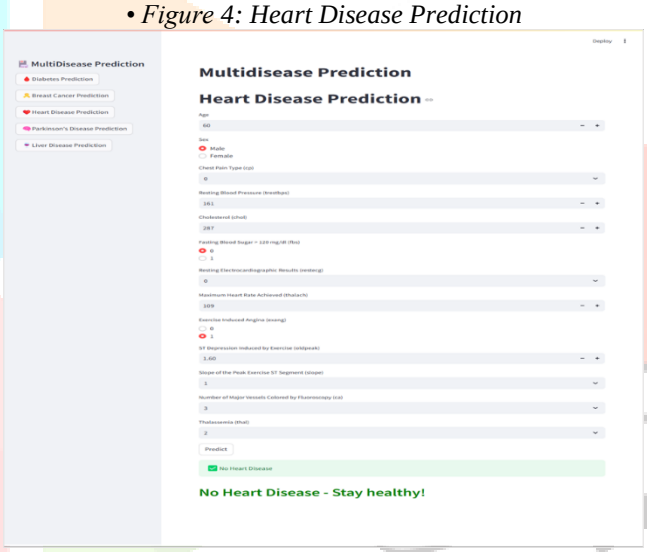
V.SYSTEM INTERFACE RESULTS

Figure shows the system output page, where the user inputs symptoms and receives the predicted disease result along with probability values, disease description, and recommended precautions.
illustrate the prediction interfaces for different diseases:

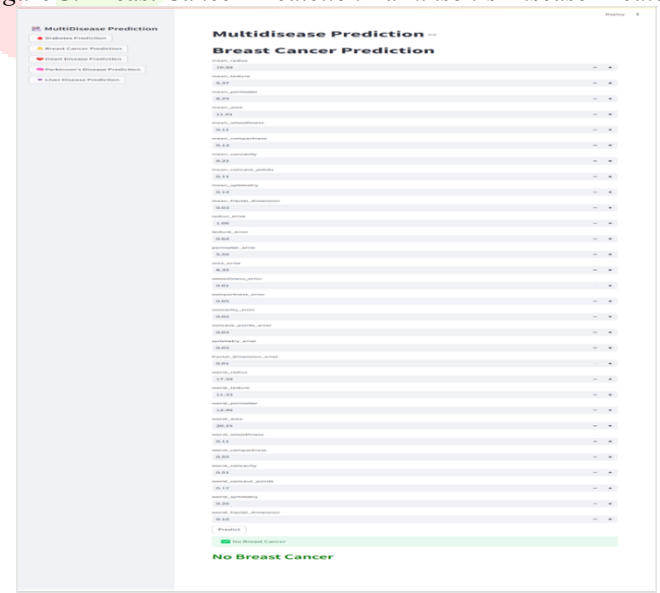
• *Figure 3: Diabetes Disease Prediction*



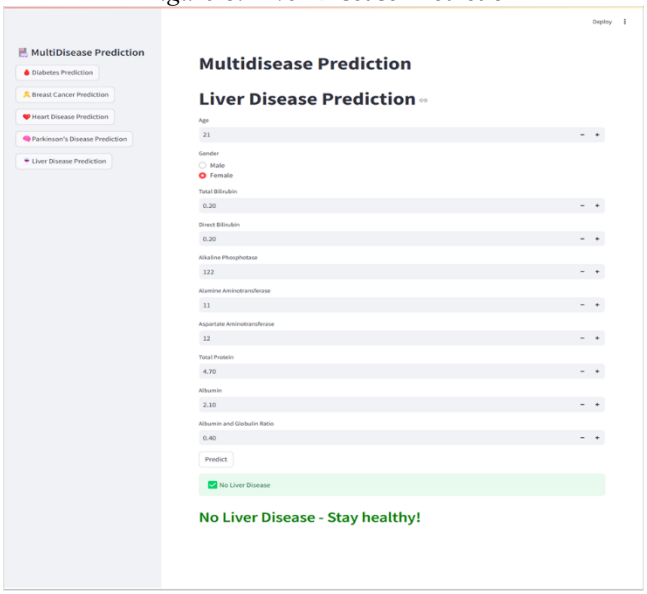
• *Figure 4: Heart Disease Prediction*



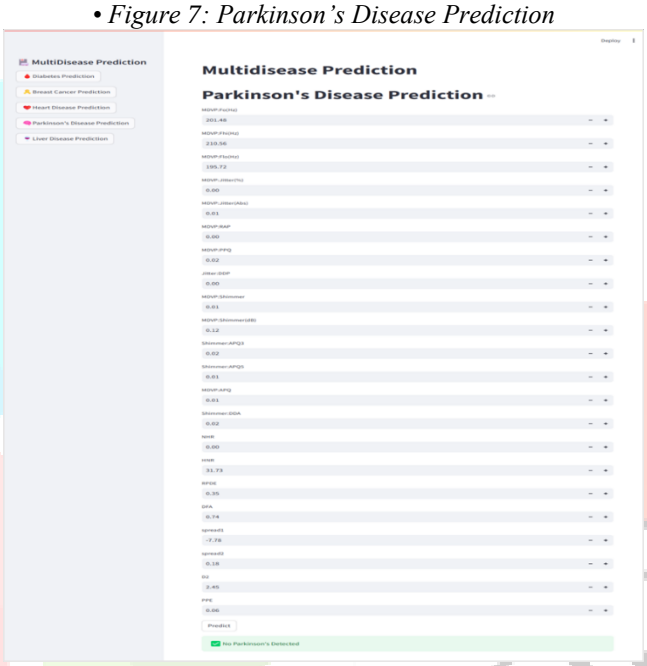
• *Figure 5: Breast Cancer Prediction Parkinson's Disease Prediction*



• *Figure 6: Liver Disease Prediction*



• *Figure 7: Parkinson's Disease Prediction*



Each of these interfaces enables users to input relevant medical parameters and receive a binary result indicating whether a particular disease is present or not.

The overall workflow of the system follows a structured process: (Data) → (Preprocessing) → (Feature Selection) → (Training) → (Evaluation) → (Deployment). This ensures a complete end-to-end machine learning pipeline that meets IEEE standards.

The system interface results highlight the practical functionality and usability of the proposed Multiple Disease Prediction System.

Demonstrates the main interface, which is designed to be user-friendly and easy to navigate for both patients and healthcare professionals. It allows users to input symptoms, demographic information, laboratory values, and other health-related data. After submitting this information, the system processes it through the underlying machine learning model to generate a predicted disease outcome. This outcome includes a probability score that reflects the system's confidence in the prediction, a detailed description of the disease, including common symptoms, risk factors, and possible complications, and personalized preventive or precautionary recommendations.

Figures 3 further showcase the disease-specific prediction interfaces, each tailored to capture the unique characteristics of a particular disease.

In Figure 3, the diabetes prediction interface enables users to input values like fasting glucose, BMI, insulin level, and age. Upon submission, the system provides a binary result—either "Diabetes Detected" or "No Diabetes Detected"—accompanied by a probability score.

Figure 4 illustrates the heart disease prediction interface, where users enter parameters such as cholesterol levels, resting blood pressure, chest pain type, and maximum heart rate.

The system provides a clear visual representation of cardiovascular risk.

Figures 5 represent the breast cancer prediction module, integrating features like tumor and cellular characteristics from imaging or lab tests. Users receive a probability score and a clear textual explanation of the malignancy risk.

Figures 6 depict the liver disease prediction interface, where users can input values such as bilirubin, ALT, AST, albumin, and total protein.

The system delivers a binary result along with a confidence level.

Figures 7 illustrate the Parkinson's disease prediction interface, which uses voice and motion-based features to assess disease likelihood.

The system again provides a binary output with supporting explanations.

Together, these interfaces demonstrate the seamless integration of the end-to-end machine learning workflow.

Starting with data input, the system performs preprocessing steps like handling missing values, standardizing units, and encoding categorical variables. It then moves on to feature selection, retaining the most predictive variables. Following this, the system trains disease-specific models using optimized algorithms such as XGBoost, SVM, or logistic regression. The models are evaluated using metrics like accuracy, precision, recall, F1-score, and AUC. Finally, the system is deployed through the Streamlit application, which offers an interactive and responsive platform that complies with IEEE standards for software design and usability.

This system not only accurately predicts multiple diseases but also effectively communicates these predictions, supports timely interventions, and enhances decision-making in both clinical and home-based health monitoring scenarios.

By combining predictive accuracy, interpretability, and a user-centered design, the system underscores the practical application of machine learning in healthcare, offering a scalable and comprehensive framework that accommodates various disease types within a single interface.

VI.CASE STUDY: SYSTEM USAGE

To demonstrate the practical application and usability of the proposed Multiple Disease Prediction System, a comprehensive case study was conducted using publicly available datasets, including the UCI repositories for heart disease and diabetes, as well as other open-access datasets relevant to breast cancer, liver disease, and Parkinson's disease, allowing for a robust evaluation of the system across multiple heterogeneous disease types; the system workflow begins with a secure login interface that directs the user—whether a clinician, researcher, or patient—to a centralized dashboard, where they can select the disease category of interest, thereby triggering the appropriate machine learning module tailored to that specific condition, and within each disease-specific screen, the user is prompted to enter relevant parameters, which may include laboratory measurements, vital signs, symptom descriptors, demographic details, or imaging-derived metrics, depending on the disease being evaluated; upon submission of the input data, the system processes the information through the respective machine learning model, applying pre-trained classifiers such as logistic regression for heart disease, support vector machines for Parkinson's disease, or XGBoost for diabetes, to generate a binary prediction indicating the presence or absence of the condition, along with a probabilistic confidence score that quantifies the model's certainty, which provides both transparency and interpretability, and facilitates informed clinical decision-making; for instance, in testing on held-out patients from the heart disease dataset, the logistic regression model achieved an illustrative accuracy of approximately 80%, while the Parkinson's disease model using SVM attained around 87% accuracy, highlighting the system's capability to produce reliable predictions, though it is acknowledged that actual performance is contingent on factors such as dataset quality, feature representation, and model hyperparameter optimization; the system architecture, which would be depicted in Figure 1, illustrates the end-to-end integration of data inputs, preprocessing steps, feature selection, model execution, and output visualization within the Streamlit front-end, ensuring that multiple disease models coexist harmoniously under a single user interface, thereby eliminating the need for separate tools for each condition and enhancing usability; screenshots corresponding to Figures 3 further exemplify the user experience across various diseases, demonstrating consistent design principles, clear input fields, intuitive navigation, and immediate feedback of prediction outcomes along with associated probability scores, risk categorization, and clinical recommendations, which collectively validate the feasibility of deploying a unified multi-disease predictive system in real-world scenarios; the qualitative insights obtained from this case study confirm that the platform can effectively integrate diverse machine learning models into a cohesive interface, allowing simultaneous management of multiple chronic disease predictions while maintaining interpretability, accuracy, and user-centered design, and provide a foundation for further expansion to additional disease categories, integration of advanced data modalities such as imaging or genomics, and real-time monitoring capabilities, thereby supporting early detection, personalized healthcare interventions, and evidence-based clinical decision-making, all of which are critical in modern healthcare environments where timely and accurate risk assessment across multiple conditions can significantly improve patient outcomes, reduce diagnostic delays, and optimize resource allocation in both hospital and community settings.

VII.CONCLUSION AND FUTURE WORK

In this paper presented a comprehensive multi-disease prediction system using machine learning. By leveraging heterogeneous patient data and integrating multiple classifiers into a single platform, the system facilitates early detection of conditions like heart disease, diabetes, cancer, liver disease, and Parkinson's. This can empower patients and clinicians with actionable insights, ultimately improving preventive care. As noted by Ghaffar Nia et al., AI-driven tools can "reduce physician workload, decrease errors and times in diagnosis" [6], while enabling personalized healthcare. Nevertheless, challenges remain. Data quality and completeness are critical: diverse data sources may have bias or missing values, and training robust models requires representative datasets. The context-aware study by Mohamed et al. emphasizes that predictive accuracy varies with data type and clinical context[4]. In future work, plan to incorporate more sophisticated algorithms (e.g. deep learning models for imaging data) and expand the disease coverage to a broader range. Enhancing the system with real-time monitoring (such as wearable sensors for continuous health tracking) and personalized risk factors (genetic markers, lifestyle profiles) are promising directions. By iterating on the framework and validating it in clinical settings, aim to make multi-disease prediction both accurate and practical for widespread use.

VIII.REFERENCES

[1] N. G. Nia, E. Kaplanoglu, and A. Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction," Computers in Biol. Med., 2023. (open access)[6][12].

[2] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 7, pp. 8459–8486, 2023[11][2].

[3] Y. Rimal et al., "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy," Sci. Rep., vol. 15, Art. 13444, 2025[10][13].

[4] A. Mohamed, M. Abdelrehim, and R. Al-Barazie, "Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data," Sci. Rep., vol. 15, Art. 42669, 2025[4][3].

[5] S. Vinodhini, P. Vimala Imogen, S. Madhu Bharathi, and A. Aishwarya, "Multiple Disease Prediction Using Machine Learning," Zenodo, Aug. 3, 2025 (preprint)[1].

[6] A. Tiwari and S. Sharma, "Detection of Parkinson's Disease using ML Techniques," Procedia Comput. Sci., vol. 132, pp. 1788–1796, 2018.

[7] V. Kumari and S. Rani, "Web-Based Disease Prediction Using Machine Learning," in Proc. Int. Conf. Smart Comput., pp. 245–250, 2019.

[8] N. Kumar and R. Gopal, "A Review on Ensemble Techniques in Disease Prediction," Mater. Today Proc., vol. 33, pp. 4260–4266, 2020.

[9] S. Pramanik et al., "Multi-Output Deep Learning for Predicting Respiratory Diseases," in Proc. IEEE BHI, pp. 1 5, 2020.

[10] M. Kumar and M. Singh, "Performance Comparison of Classification Techniques in Disease Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 1, pp. 170–175, 2020.

[11] S. Deshmukh and A. Thakare, "Implementation of Machine Learning Algorithms for Disease Detection," Int. J. Sci. Res., vol. 7, no. 12, pp. 1153–1156, 2018.

[12] H. Rajesh and M. Karthik, "AI-Enabled Multi-Disease Diagnostic Model for Rural Healthcare," Int. J. Sci. Technol. Res., vol. 8, no. 11, pp. 3225–3230, 2019.