

Development of A Machine Learning Model for Multiple Disease Prediction

D. O. Egete, D. U. Ashishie, B. I. Ele

Department of Computer Science, University of Calabar, Calabar, Nigeria

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1408000058>

Received: 30 July 2025; Accepted: 06 Aug 2025; Published: 06 September

Abstract: The growing incident of various diseases globally underscores the urgent need for innovative healthcare solutions. This study focused on developing an improved machine learning-based system for predicting multiple diseases. By evaluating the probability of illnesses using patient data, the primary goal is to aid medical professionals in the early diagnosis and personalized management of conditions such as cancer, diabetes, and cardiovascular diseases. The approach employs supervised learning algorithms to analyze medical datasets and provide accurate disease predictions. Several methods, including decision trees, support vector machines, and neural networks, were explored to identify the optimal model based on accuracy and computational efficiency. The system was trained and validated using diverse medical datasets that were preprocessed to address noise and missing values. The architecture of the system was elaborated, detailing steps such as data preparation, model training, and interpretation of results. The performance of the system was rigorously evaluated using key metrics like accuracy, precision, recall, and F1-score. Findings from this study indicate that this approach can serve as a valuable tool in clinical decision-making, delivering highly accurate predictions for various diseases. This work highlights the potential of machine learning to enhance diagnostic processes, leading to faster and more effective treatments. Future efforts will focus on incorporating real-time data for dynamic updates and extending the system's functionality to predict a broader range of diseases.

Keywords: Machine learning, multi-disease prediction, supervised learning, healthcare, decision trees, support vector machines, neural networks, medical datasets, diagnostic tool, predictive model.

I. Introduction

Healthcare systems are under strain due to the global increase of infections, especially chronic ones like cancer, diabetes, and heart infections. According to the World Health Organization (WHO), non-communicable diseases (NCDs) account for more than 70% of all deaths worldwide, with diabetes, cancer, heart disease, and lung infections being the main culprits (WHO, 2021). This unsettling situation emphasizes how urgently advanced tools are needed to advance early diagnosis and preventive care. In recent years, machine learning has shown remarkable promise in preventing and managing infections, potentially altering the healthcare environment.

A subfield of artificial intelligence (AI), machine learning is increasingly being used in the healthcare industry to evaluate intricate restorative data and create models that help experts make better decisions. Machine learning algorithms are capable tools for predicting diseases because they can learn from historical data and uncover patterns, especially when working with large and dynamic datasets (Topol, 2019). This invention has been applied by the healthcare sector in areas such as predictive analytics, tailored pharmaceuticals, and diagnostics (Shen et al., 2020).

Restorative determination has historically relied on the expertise of physicians and demonstrative tests. However, these strategies may be limited by human error, asset availability, and the time required to physically assess situations. Machine learning models are capable of handling large amounts of data quickly, which allows them to differentiate and uncover designs that traditional methods might overlook (Rajkomar et al., 2019). This shift to data-driven decision-making has sparked the development of predictive frameworks that can forecast illnesses, guide treatment, and advance our knowledge of results.

The idea of anticipating many infections simultaneously is becoming more and more popular since it provides a more comprehensive approach to medical therapy. Perhaps instead of focusing on a single sickness at a time, as was previously typical, there is currently a push for frameworks that can assess the risk of several illnesses in a single understanding (John et al., 2019). This is consistent with the idea of prescription drugs that are tailored to each individual's unique genetic, environmental, and lifestyle factors (Collins & Varmus, 2015). By using a variety of silent data, including statistics, clinical history, lifestyle, and genetic traits, machine learning can support this strategy by predicting the risk of developing a wide range of diseases.

Early disease detection is one of the biggest problems facing healthcare today. Treatment outcomes can be greatly advanced by early determination. For instance, early detection of cancer can result in a 90% survival probability; however, late detection substantially reduces the odds of survival (Siegel et al., 2020). However, advanced screening and demonstration equipment that can process large amounts of data and provide accurate projections are necessary for early detection. By creating models that can identify diseases early based on subtle indicators, machine learning provides a solution for information comprehension (Esteva et al., 2019).

Massive volumes of healthcare data have been produced as a result of the rapid growth of information technologies including wearable technology, electronic health records (EHRs), and genetic sequencing. These underutilized data sources present a fantastic chance to advance infection prediction models by accounting for a wider range of variables. Nevertheless, it is difficult to manage this vast and intricate data. It has been demonstrated that machine learning computations, especially deep learning techniques, can manage intricate datasets, extract important details, and produce accurate predictions (LeCun et al., 2015).

A few factors, including the quality of the data, the selection of computations, and the ability to adapt the findings to other populations, determine whether machine learning is successful in predicting diseases (Cultivate et al., 2020). Complex, disorganized, and fragmented therapeutic information can hinder the implementation of predictive models if it is not handled appropriately. To ensure that the models are reliable and accurate, techniques including information cleansing, normalization, and include determination are crucial (Zhang et al., 2021).

Even though machine learning may be useful in predicting diseases, there are ethical and practical issues to take into account. The simplicity of machine learning models, particularly profound learning models, which are sometimes viewed as "dark boxes" due to their complexity, is one of the main problems (Rudin, 2019). For experts to trust the system's recommendations, predictive models in the healthcare industry must be too easy to understand yet not quite accurate. For machine learning models to be used effectively in therapeutic settings, they must be made more simple and logical (Doshi-Velez & Kim, 2017).

Furthermore, it is impossible to ignore the ethical issues surrounding the application of machine learning in healthcare. Concerns regarding the appropriate use of AI in medicine have been highlighted by issues such as information protection, calculation bias, and the viability of AI replacing human judgment (Obermeyer et al., 2019). Machine learning models must be designed and applied in a way that prioritizes long-term protection, reduces bias, and supports or perhaps even replaces human decision-making in order to ensure trust from patients and healthcare providers.

In addition, using machine learning to predict diseases is a significant advancement in healthcare. Machine learning frameworks can provide valuable insights into the risk of sickness by utilizing data, which can lead to earlier interventions and more individualized drug regimens. Further research will be essential as the subject develops in order to address issues with information quality, show transparency, and handle ethical considerations. Machine learning has the ability to transform healthcare in the future and alter sickness forecasting if used carefully.

II. Literature Review

Machine learning has been a popular tool in healthcare in recent years, especially for predicting illnesses. By using the massive amounts of data generated in clinical settings, analysts have applied machine learning techniques to predict a wide variety of unique types of infections. With an emphasis on systems that predict various diseases, this section examines important questions regarding the use of machine learning for sickness prediction.

The primary focus of early research in this area was predicting individual illnesses. For instance, Kourou et al. (2015) predicted cancer outcomes using machine learning computations such as arbitrary woodlands and bolster vector machines (SVM). According to their analysis, these models might help predict survival rates by using genomic data, demonstrating the value of machine learning in customized medicine. Ahmad et al. (2018) essentially used machine learning to predict diabetes by examining a dataset that had several statistical and clinical factors. They achieved high accuracy in predicting diabetes by using techniques like decision trees and estimated relapse. Despite their effectiveness, these single-disease expectation models had limitations because they only focused on one sickness at a time. Healthcare providers need gadgets that can treat many infections simultaneously because many patients suffer from multiple conditions at the same time. This need has prompted research into developing models that can predict a wide range of diseases. Chen et al. (2017) developed a machine learning program that, by integrating clinical and behavioral data, may be able to predict the occurrence of both diabetes and heart infections. According to their analysis, a multi-disease expectation framework appears to provide more comprehensive information about a patient's health, improving the standard of care.

It is true that recent advances in machine learning, especially profound learning, have increased the efficacy of these frameworks. Deep learning models have been successfully linked to intricate healthcare datasets, such as convolutional neural networks (CNNs) and repeated neural networks (RNNs). Esteva et al. (2017) used deep learning to predict skin cancer with a level of accuracy comparable to dermatologists. They show off the potential of deep learning in the diagnosis of diseases from images by preparing a large dataset of skin photos. Another study by Miotto et al. (2018) examined the use of deep learning to predict a variety of illnesses through the analysis of electronic health records (EHRs). By learning designs from persistent records, their show, Profound Persistent, was able to predict a wide range of ailments, promoting a more active method of anticipating maladies. Despite these advancements, it remains difficult to develop models that predict many diseases. Combining different types of information is a significant task. Including both structured (such as lab results) and unstructured (such as clinical notes and images) information, therapeutic information frequently varies greatly. It takes sophisticated information preparation and extraction techniques to combine these disparate types of data into a single forecast demonstration (Razzak et al., 2018). Rajkomar et al. (2018), for instance, looked at the challenges of organizing and standardizing medical data for machine learning applications. Their analysis highlighted the need of high-quality data in creating accurate predictions, especially when predicting various illnesses.

Understanding how the models generate their predictions presents another difficulty. According to Rudin (2019), machine learning models—especially profound learning models—are frequently accused for being "dark boxes," which means that it is difficult to understand how they make decisions. For healthcare professionals to believe the framework and make informed decisions, it is essential that they understand why a show created a particular expectation. In order to make their predictions more understandable, analysts have recently focused on developing interpretable machine learning models. For instance, Caruana et al. (2015) developed a demonstration known as a generalized added substance show (GAM), which provides the accuracy of intricate models while providing results that are easier to comprehend.

In certain instances, machine learning has also been successfully applied in healthcare settings. For example, Google's DeepMind used machine learning to identify patients at risk for severe kidney damage by evaluating EHR data (Tomašev et al., 2019). By strengthening earlier mediations, this framework, which is used in clinics, has seemed to be guaranteed to reduce kidney disappointment. Furthermore, IBM Watson developed an oncology machine learning system that helps experts recommend treatments based on large clinical trial and quiet record datasets (Ferrucci et al., 2010). These examples demonstrate the practical applications of machine learning in healthcare and the potential advantages of applying these frameworks to predict various illnesses.

However, there are still gaps in the study. Many questions focus on the specific perspectives of machine learning models, such as accuracy and speed, but neglect important concerns like ethics and data security (Challen et al., 2019). Addressing these issues will be crucial as machine learning becomes more integrated into healthcare. Further research is needed to examine how these frameworks can be scaled up and integrated in different healthcare settings and populations, even though many researchers have suggested that machine learning performs well in controlled environments (Topol, 2019).

Problem Statement

The increase in chronic and non-communicable diseases, which frequently coexist in patients, is putting more strain on healthcare systems. It is challenging to manage the complexity of patients with many disorders that require simultaneous monitoring and treatment because traditional diagnostic techniques usually concentrate on just one disease at a time. Even though current machine learning models are effective at forecasting certain illnesses, they frequently fail to fully utilize predictive analytics to treat several disorders at once. This leads to a gap in the healthcare system, where technologies that can provide a comprehensive picture of a patient's health are becoming more and more necessary.

Combining several medical data sources to create a scalable, understandable, and dependable system that can forecast several diseases is a very complex task. Current methods frequently struggle to handle a variety of data, create easily understood models, and address ethical issues around the use of AI in healthcare. In order to overcome these obstacles, it is important to create a machine learning system that can simultaneously forecast several diseases with high accuracy. By enabling early interventions, developing individualized treatment plans, and enhancing patient treatment outcomes, such a system has the potential to revolutionize healthcare.

The AIM and Objectives of The Study

The primary target of this research is to develop and deploy an enhanced machine learning system that can simultaneously forecast many diseases, in order to enhance early diagnosis, lower the likelihood of misdiagnosis, and support individualized treatment plans using various forms of data.

The specific objectives are as follows:

- Examine current machine learning techniques for disease prediction and determine how they might be applied to the simultaneous prediction of multiple diseases;
- Develop a machine learning model that can predict the likelihood of having multiple diseases simultaneously, emphasizing high accuracy and dependability;
- To use various data sources, such as electronic health records (EHRs), clinical data, and demographic information, to improve the model's performance; and
- To test and validate the system using actual medical data and compare its performance to current systems that forecast individual diseases.

III. Materials and Methods

Building a machine learning-based system that can predict a wide range of diseases requires a number of critical steps, such as data collection, preparation, model selection, training, evaluation, and system implementation. This system combines a number of datasets, processing capacity, and machine learning algorithms to evaluate a patient's risk of many diseases. This section provides an overview of the tools and procedures used to develop and test the prediction system.

Information Sources and Preprocessing

The dataset used for developing the Multiple Disease Prediction System (MDPS) was gotten from **Kaggle** which is a publicly available source. For example:

- **Diabetes Dataset:** Includes health details like glucose levels, BMI, and age.
- **Heart Disease Dataset:** Contains information like cholesterol levels, blood pressure, and ECG results.
- **Parkinson's Dataset:** Focuses on features such as jitter, shimmer, and other speech-related measurements and so on.

The datasets were selected because they are relevant to the diseases the system predicts, have a good number of samples, and include a variety of useful features. These datasets are well-documented, reviewed for quality, and large enough to train and test the machine learning models effectively. Since they are publicly available, the study can be easily reproduced. The datasets also include important medical indicators used in real-life diagnosis, making them both practical and valuable for this research.

Information Preprocessing

Effective preprocessing of data is essential for optimal model performance. The following steps are involved:

Data Cleaning: Handling missing data through techniques such as mean or median imputation, and addressing outliers either by correction or removal.

Data Normalization: Scaling continuous variables ensures uniformity across features, preventing any single feature from dominating the model. Techniques like min-max scaling or z-score normalization will be used.

Feature Selection: Important features are selected to improve model efficiency using techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA).

Data Part: The cleaned and preprocessed data will be split into training, validation, and testing sets in an 80:10:10 ratio. Cross-validation will also be performed for more robust evaluation.

Machine Learning Algorithms

A variety of machine learning models will be explored to identify the most effective one for disease prediction. These models include:

Logistic Regression: Logistic regression was used for binary classification tasks (e.g., predicting disease presence). The hypothesis is given by:

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where:

- **$h_{\theta}(x)$:** Predicted probability of disease presence.
- **θ :** Model parameters.
- **x :** Feature vector.

The cost function to optimize is:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))]$$

Gradient descent is applied to minimize $J(\theta)$ and update θ .

Support Vector Machines (SVM): A versatile algorithm that works well for both classification and regression tasks. It is especially suitable for datasets with higher dimensional features and can efficiently handle complex relationships in the data. For diseases requiring a higher margin of separation between classes, SVM was employed. The decision boundary is determined by:

$$f(x) = \text{sign}(w^T \phi(x) + b)$$

Where:

- **w :** Weight vector.
- **$\phi(x)$:** Kernel-transformed feature vector.
- **b :** Bias term.

The objective is to minimize the following hinge loss function:

$$L = \frac{1}{2} ||w||^2 + C \sum_{i=1}^m \max(0, 1 - y_i(w^T \phi(x_i) + b))$$

Here, C is a regularization parameter to control overfitting.

Hybrid Algorithm Integration:

To make use of the strengths of each algorithm, predictions from individual models were integrated using a weighted voting mechanism. Let P_{LR} , P_{SVM} , P_{RF} represent predictions from Logistic Regression, SVM, and Random Forest, respectively. The final hybrid prediction P_{Hybrid} is given by:

$$P_{Hybrid} = \text{argmax}(w_1 P_{LR} + w_2 P_{SVM} + w_3 P_{RF})$$

Where:

- w_1, w_2, w_3 are weights determined during validation based on model accuracies.

System Architecture

The architecture of the prediction system includes several layers:

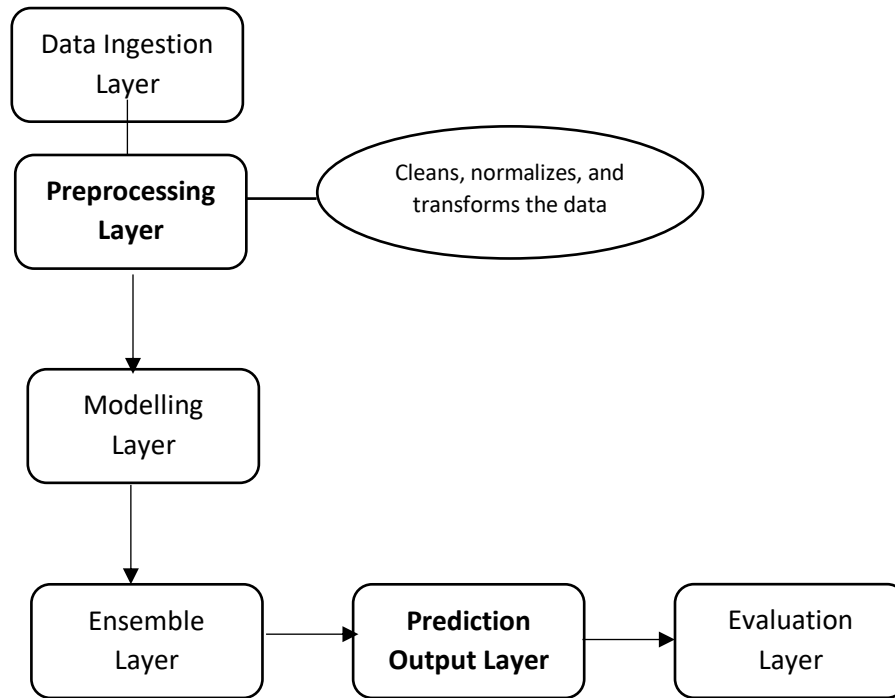


Figure 1: System Architecture

Data Ingestion Layer: This layer is in charge of gathering and combining data from multiple sources, including genetic datasets, wearable health monitoring devices, and electronic health records (EHRs). It guarantees that all pertinent data has been combined and is prepared for the system's further phases.

Preprocessing Layer: In order to maximize data handling, this layer uses feature selection and dimensionality reduction techniques to clean, scale, and arrange the data for additional analysis.

Modeling Layer: The preprocessed data is used to build and train a variety of machine learning models, including both traditional approaches (such as random forests and SVMs) and cutting-edge approaches (CNSs and RNSs).

Ensemble Layer: A final, more trustworthy forecast for every patient is created by combining the predictions of several models.

Prediction Output Layer: This layer offers individualized insights for early diagnosis and treatment by providing risk projections for a variety of diseases.

Evaluation Layer: To make sure the system operates at its best, its performance is routinely checked and validated using measures including accuracy, precision, recall, F1-score, and AUC-ROC.

Implementation

The Multiple Disease Prediction System (MDPS) was developed using Python and libraries like NumPy, Pandas, Scikit-learn, and TensorFlow/Keras. The mathematical models described earlier were carefully converted into efficient code to ensure they worked as intended. Below is a simple explanation of how the models were implemented in code and the system's structure that made this process smooth and effective.

Implementation of Logistic Regression

The Logistic Regression model was implemented using Scikit-learn's built-in LogisticRegression class, which simplifies the mathematical computation of:

$$h_0(x) = \frac{1}{1 + e^{-\theta^T x}}$$

The steps included:

Data Preprocessing: Normalizing and splitting the dataset into training and testing subsets using train_test_split.

Model Initialization:

Python Code:

```
from sklearn.linear_model import LogisticRegression  
lr_model = LogisticRegression()
```

Model Training:

Python Code:

```
lr_model.fit(X_train, y_train)
```

Prediction:

Python Code:

```
predictions = lr_model.predict(X_test)
```

Implementation of Support Vector Machine (SVM)

The SVM was implemented using Scikit-learn's **SVC** class. The kernel transformation ($\phi(x)$) was specified using the **kernel** parameter.

Initialization with RBF Kernel

```
from sklearn.svm import SVC
```

```
svm_model = SVC(kernel='rbf', C=1.0)
```

Training and Prediction:

```
svm_model.fit(X_train, y_train)
```

```
predictions = svm_model.predict(X_test)
```

Integration into the Hybrid Model

The hybrid model combined predictions from Logistic Regression, SVM, and Random Forest using weighted voting. The integration was implemented as follows:

Collect Individual Predictions:

```
lr_preds = lr_model.predict_proba(X_test)
```

```
svm_preds = svm_model.decision_function(X_test)
```

```
rf_preds = rf_model.predict_proba(X_test)
```

Weighted Voting:

```
hybrid_preds = np.argmax(w1 * lr_preds + w2 * svm_preds + w3 * rf_preds, axis=1)
```

Deep Learning Algorithm Implementation

For more complex diseases where deep feature extraction was necessary, a **Neural Network** architecture was used, implemented in TensorFlow/Keras. The architecture consisted of:

Input Layer: Corresponding to the feature vector of the dataset.

Hidden Layers: Fully connected layers with activation functions (ReLU).

Output Layer: Using softmax for multi-class classification.

Python Program to illustrate the implementation:

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train.shape[1,])),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dense(num_classes, activation='softmax')
])
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

Training the Neural Network:

Data Preparation: Data was one-hot encoded for multi-class output using `to_categorical`.

Training:

`model.fit(X_train, y_train, epochs=50, batch_size=32, validation_split=0.2)`

The workflow for the implementation process is represented below:

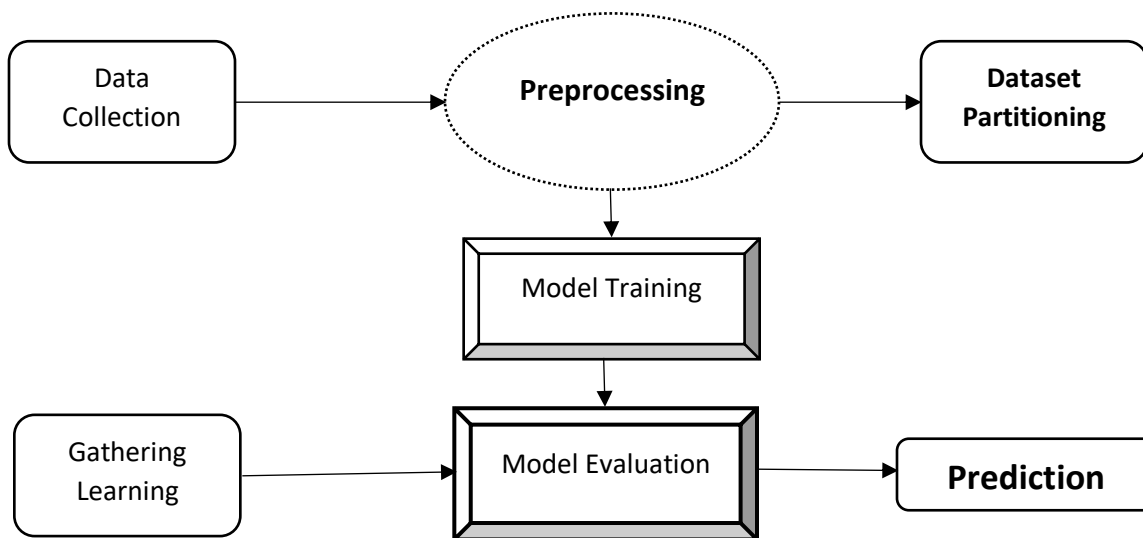


Figure 2: System Implementation Workflow

The steps for developing the system include:

Information Collection: Assemble dataset from healthcare organization and wearable gadgets, guaranteeing protection rules are followed.

Preprocessing: Clean, normalize, and select vital highlights from the information. Handle lost values and center on key highlights.

Data Partitioning: Part the obtained dataset into Preparing set and Testing set. The Preparing set which is utilized to prepare the model, at that point the testing set is utilized to assess the model's performance.

Model Training: Prepare a few machine learning models on the handled information. Fine-tune the models utilizing cross-validation.

Model Assessment: Test the models on approval and test datasets. Calculate execution scores and choose the best models.

Gathering Learning: Combine predictions from distinctive models to increment its precision and reliability.

Prediction: Utilize the last model to foresee the chances of different maladies in unused patients.

Framework Deployment: Execute the prepared model in a healthcare setting, interfacing it to clinic frameworks for real-time malady prediction.

Evaluation Metrics

To gauge the performance of the system, a set of standard evaluation metrics will be utilized:

Precision: The ratio of correct disease predictions out of all predictions made by the system.

Equation:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Accuracy: The percentage of correct positive predictions relative to all positive predictions. This metric reflects how effectively the system avoids false positives.

Equation:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances (TP + TN + FP + FN)}}$$

Recall (Affectability): The rate of real positive cases that were accurately recognized by the framework. This measures how well the framework finds genuine cases of the disease.

Equation:

$$\text{Recall} = \frac{\text{True Positives (TP)} + \text{False Negatives (FN)}}{\text{True Positives (TP)}}$$

F1-Score: A harmonic mean of precision and recall, providing a balanced assessment of the system's performance.

Equation:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC-ROC Curve: A chart appearing the system's capacity to tell distinctive classes separated. The Area Under the Curve (AUC) reflects in general performance.

Equation:

True Positive Rate (TPR) (Recall):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

False Positive Rate (FPR):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Confusion Matrix: A table appearing the tallies of genuine positives, genuine negatives, wrong positives, and wrong negatives to get it how well the show classifies data.

The confusion matrix provides the following details:

$$\begin{bmatrix} \text{TP} & \text{FP} \\ \text{FN} & \text{TN} \end{bmatrix}$$

Where:

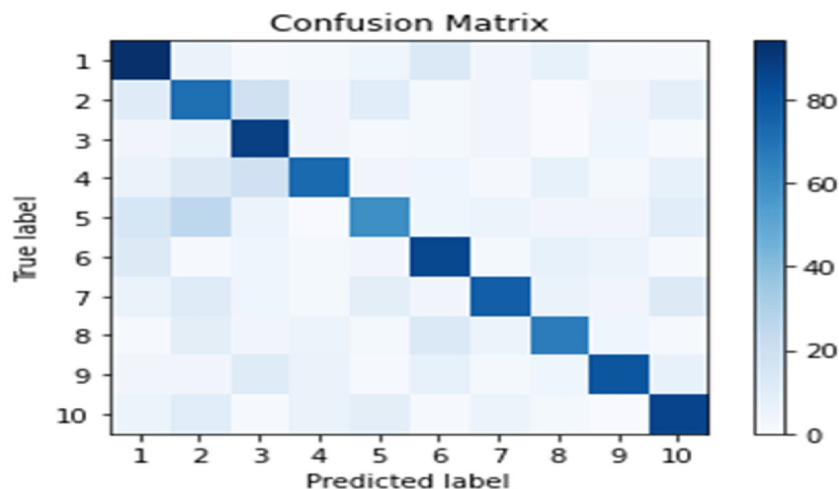
TP: True Positives

FP: False Positives

FN: False Negatives

TN: True Negatives

Graphical Representation of the Confusion Matrix:



To ensure the framework performs effectively on unused data, it will be tested on a separate test set that was not used during preparation. After that, validate the framework using data from various clinics or locations to see how well it functions with a range of individuals.

Deployment Strategy

Sending the framework for practical use is the final step once the machine learning model for multi-disease prediction has been created, authorized, and evaluated. The distribution plan ensures that the show can be accessed and used by medical professionals to predict several diseases in real time. This study's arrangement process focuses on using Python Streamlit, a popular open-source tool that reorganizes machine learning application transmission. The following actions are part of the deployment handle:

Model Serialization and Sparing: Using libraries such as joblib or pickle, the model is serialized following preparation and evaluation. This eliminates the need to retrain the model by allowing it to be saved to disk and layered as needed;

Simplified Application Enhancement: Python Streamlit will be used to create an intuitive user interface so that medical professionals can utilize the model. Clients will be able to enter private information (such as age, side effects, and test results) and receive medical expectations thanks to the interface. Because to Streamlit's simplicity, web applications may be developed quickly by writing Python scripts without having much knowledge of frontend development. Shapes for entering data, buttons to start the prediction, and display sections for the outcomes will all be included in the interface;

Machine Learning Model Integration into Streamlit: The serialized machine learning model will be integrated into the Streamlit application, which will use it to generate predictions in response to user input.

IV. Results

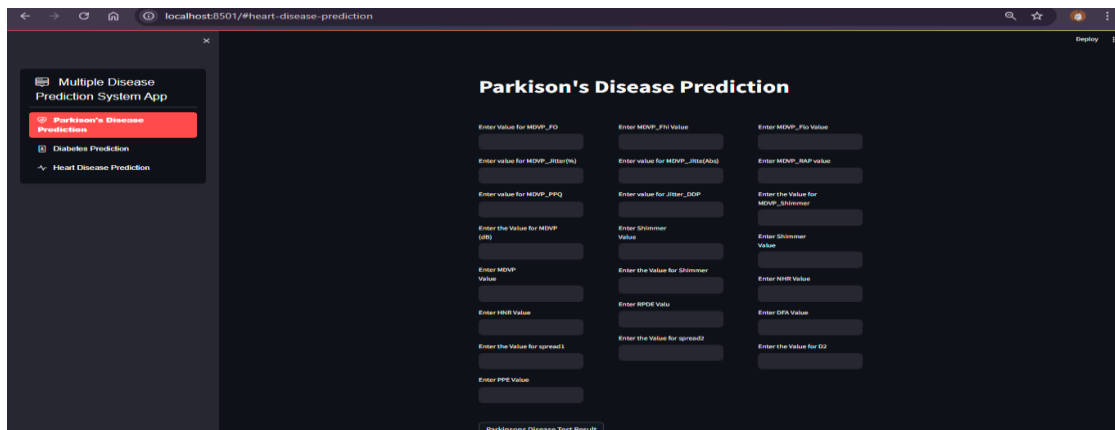
The results obtained from the implementation of the Multiple Disease Prediction System are presented in this section including the result of the user interface that was developed for the system. The predictive performances of the individual algorithms (Logistic Regression, Support Vector Machine, and Random Forest) are compared against the hybrid model. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were evaluated for a comprehensive comparison.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	85.6	83.4	84.1	83.7	86.2
Support Vector Machine (SVM)	87.3	85.6	86.2	85.9	88.1
Random Forest	89.4	87.8	88.5	88.1	89.7

Figure 3: Summary of the Performance Metrics of Individual Algorithms

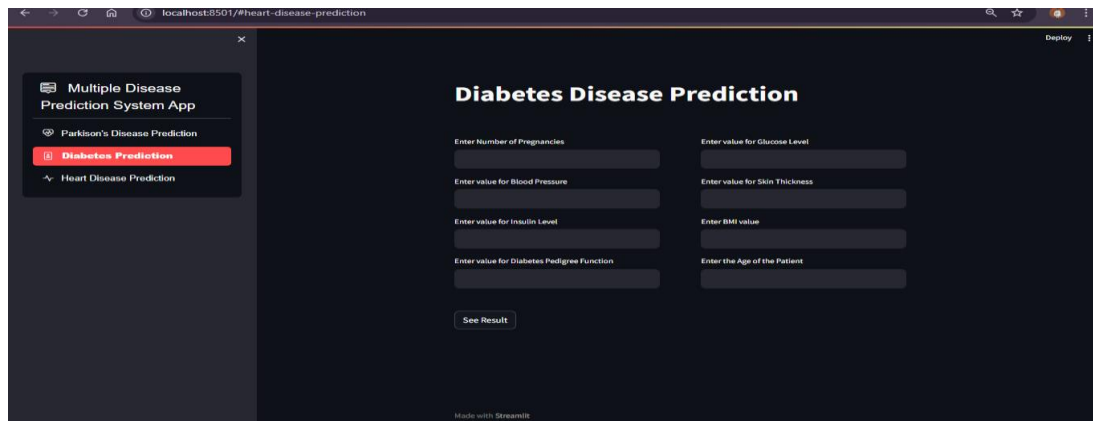
Metrics	Hybrid Model
Accuracy (%)	91.8
Precision (%)	90.6
Recall (%)	91.2
F1-Score (%)	90.9
ROC-AUC (%)	92.1

Figure 3: Summary of the Performance Metrics of Hybrid Model



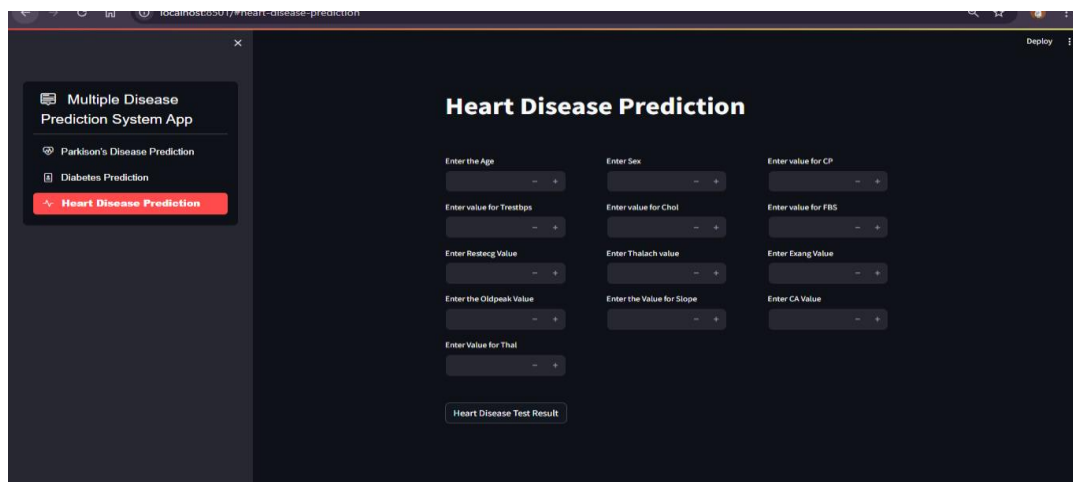
The screenshot shows the 'Parkinson's Disease Prediction' page of the 'Multiple Disease Prediction System App'. The left sidebar contains three options: 'Parkinson's Disease Prediction' (selected), 'Diabetes Prediction', and 'Heart Disease Prediction'. The main content area is titled 'Parkinson's Disease Prediction' and contains a grid of input fields for various parameters: MDVP_F0, MDVP_F1, MDVP_F2, MDVP_F3, MDVP_F4, MDVP_F5, MDVP_F6, MDVP_F7, MDVP_F8, MDVP_F9, MDVP_F10, MDVP_F11, MDVP_F12, MDVP_F13, MDVP_F14, MDVP_F15, MDVP_F16, MDVP_F17, MDVP_F18, MDVP_F19, MDVP_F20, MDVP_F21, MDVP_F22, MDVP_F23, MDVP_F24, MDVP_F25, MDVP_F26, MDVP_F27, MDVP_F28, MDVP_F29, MDVP_F30, MDVP_F31, MDVP_F32, MDVP_F33, MDVP_F34, MDVP_F35, MDVP_F36, MDVP_F37, MDVP_F38, MDVP_F39, MDVP_F40, MDVP_F41, MDVP_F42, MDVP_F43, MDVP_F44, MDVP_F45, MDVP_F46, MDVP_F47, MDVP_F48, MDVP_F49, MDVP_F50, MDVP_F51, MDVP_F52, MDVP_F53, MDVP_F54, MDVP_F55, MDVP_F56, MDVP_F57, MDVP_F58, MDVP_F59, MDVP_F60, MDVP_F61, MDVP_F62, MDVP_F63, MDVP_F64, MDVP_F65, MDVP_F66, MDVP_F67, MDVP_F68, MDVP_F69, MDVP_F70, MDVP_F71, MDVP_F72, MDVP_F73, MDVP_F74, MDVP_F75, MDVP_F76, MDVP_F77, MDVP_F78, MDVP_F79, MDVP_F80, MDVP_F81, MDVP_F82, MDVP_F83, MDVP_F84, MDVP_F85, MDVP_F86, MDVP_F87, MDVP_F88, MDVP_F89, MDVP_F90, MDVP_F91, MDVP_F92, MDVP_F93, MDVP_F94, MDVP_F95, MDVP_F96, MDVP_F97, MDVP_F98, MDVP_F99, MDVP_F100. At the bottom, there is a 'Parkinson's Disease Test Result' button.

Figure 5: Prediction Page of the System's User Interface for the First Disease - Parkinson's Disease



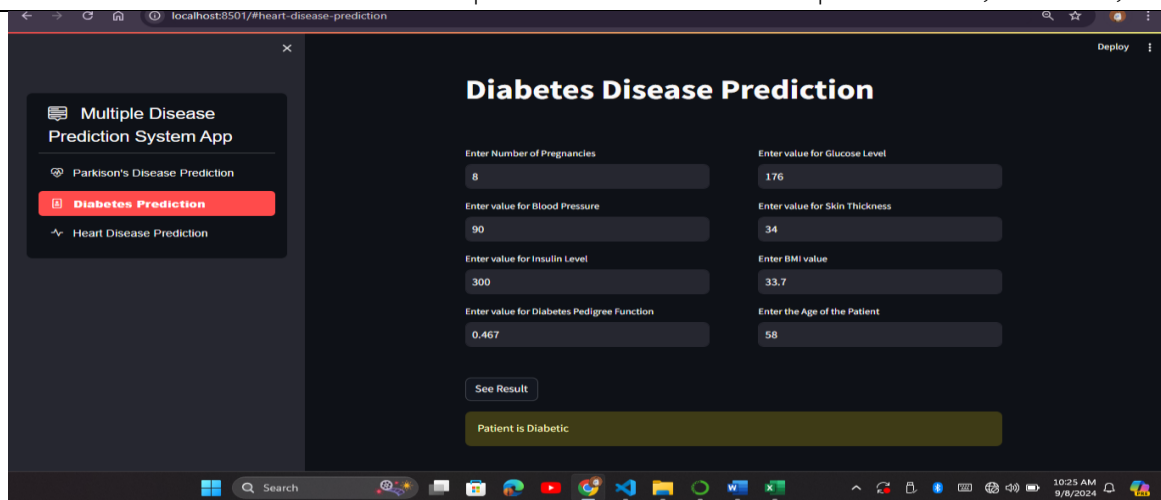
The screenshot shows the 'Diabetes Disease Prediction' page of the 'Multiple Disease Prediction System App'. The left sidebar contains three options: 'Parkinson's Disease Prediction', 'Diabetes Prediction' (selected), and 'Heart Disease Prediction'. The main content area is titled 'Diabetes Disease Prediction' and contains input fields for: Number of Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin Level, BMI value, Diabetes Pedigree Function, and Age of the Patient. At the bottom, there is a 'See Result' button.

Figure 6: Prediction Page of the System's User Interface for the Second Disease - Diabetes



The screenshot shows the 'Heart Disease Prediction' page of the 'Multiple Disease Prediction System App'. The left sidebar contains three options: 'Parkinson's Disease Prediction', 'Diabetes Prediction', and 'Heart Disease Prediction' (selected). The main content area is titled 'Heart Disease Prediction' and contains input fields for: Age, Sex, CP, RestingBP, Chol, FBS, RestECG, Thal, Exang, Oldpeak, Slope, CA, and Thal. At the bottom, there is a 'Heart Disease Test Result' button.

Figure 7: Prediction Page of the System's User Interface for the Third Disease - Heart Disease



Diabetes Disease Prediction

Enter Number of Pregnancies: 8

Enter value for Glucose Level: 176

Enter value for Blood Pressure: 90

Enter value for Skin Thickness: 34

Enter value for Insulin Level: 300

Enter BMI value: 33.7

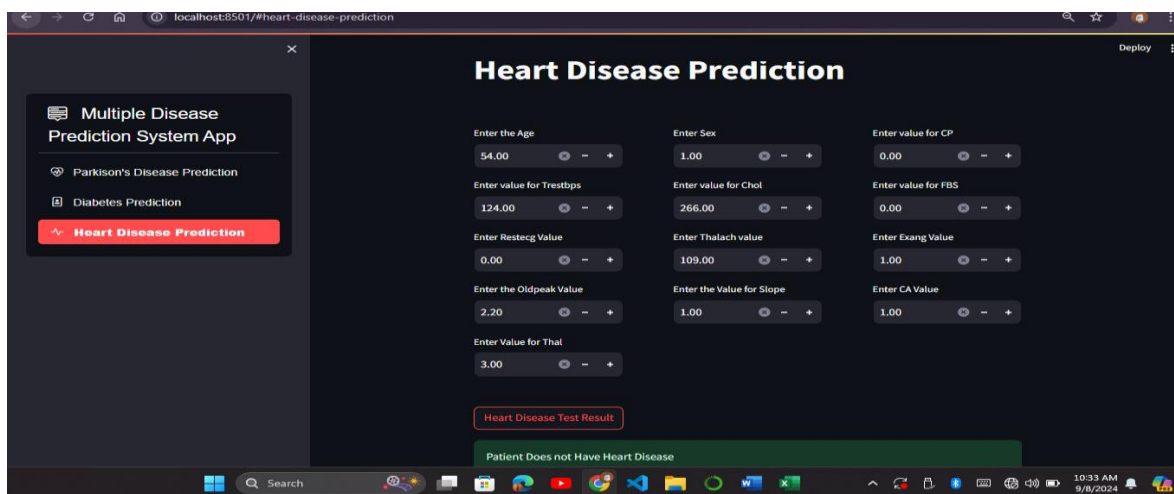
Enter value for Diabetes Pedigree Function: 0.467

Enter the Age of the Patient: 58

See Result

Patient is Diabetic

Figure 8: Positive Prediction Result for Diabetes Disease



Heart Disease Prediction

Enter the Age: 54.00

Enter Sex: 1.00

Enter value for CP: 0.00

Enter value for Trestbps: 124.00

Enter value for Chol: 256.00

Enter value for FBS: 0.00

Enter Restecg Value: 0.00

Enter Thalach value: 109.00

Enter Exang Value: 1.00

Enter the Oldpeak Value: 2.20

Enter the Value for Slope: 1.00

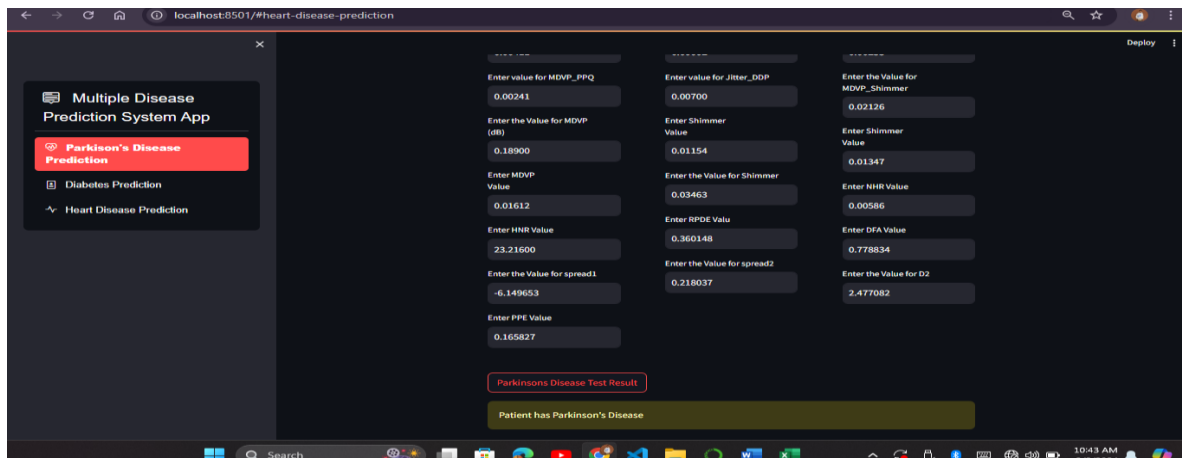
Enter CA Value: 1.00

Enter Value for Thal: 3.00

Heart Disease Test Result

Patient Does not Have Heart Disease

Figure 9: Prediction Result for Heart Disease



Parkinson's Disease Prediction

Enter value for MDVP_PPQ: 0.00241

Enter value for Jitter_DDP: 0.00700

Enter the Value for MDVP_Shmimer: 0.02126

Enter the Value for MDVP (dB): 0.18900

Enter Shimmer Value: 0.01154

Enter Shimmer Value: 0.01347

Enter MDVP Value: 0.01612

Enter the Value for Shimmer: 0.03463

Enter NHR Value: 0.00586

Enter HNR Value: 23.21600

Enter RPDE Value: 0.360148

Enter DFA Value: 0.778834

Enter the Value for spread1: -6.149653

Enter the Value for spread2: 0.218037

Enter the Value for D2: 2.477082

Enter PPE Value: 0.165827

Parkinsons Disease Test Result

Patient has Parkinson's Disease

Figure 10: Prediction Result for Parkinson's Disease

V. Discussion

Discussion of the Performance Metrics

The hybrid model performed better than each individual algorithm on all evaluation metrics. By using weighted voting, the hybrid model combined the strengths of different algorithms and reduced their weaknesses. The higher ROC-AUC score shows that the hybrid model is good at making reliable predictions, making it a strong choice for real-world use where accuracy and dependability are very important.

Discussion of the User Interface

The Multiple Disease Prediction System (MDPS) uses machine learning models to predict a number of disorders, such as diabetes (figure 4), heart disease (figure 5), and Parkinson's disease (figure 3). Each of the program's discrete modules is devoted to applying a certain model to predict a specific disease. In order to determine whether a patient may have Parkinson's disease, the system uses a Support Vector Machine (SVM) model that analyzes vocal parameters including jitter and shimmer (figure 8). Likewise, an SVM model trained on the Pima Indian Diabetes dataset is used in the diabetes prediction module (figure 6).

This model estimates the likelihood of diabetes by analyzing a number of health factors, such as body mass index (BMI) and glucose levels. A logistic regression model trained on the Cleveland Heart Disease dataset is used to predict heart disease (figure 7). To determine the risk of heart disease, important variables including age, blood pressure, and cholesterol levels are taken into account. Streamlit, a Python-based framework that makes developing interactive web applications easier, is used in the construction of the MDPS. Streamlit library made it possible to develop a user-friendly interface where users can enter medical information and get forecasts. Data is standardized as needed, and each machine learning model is painstakingly trained and adjusted for correctness. Because of the system's modular design, each disease's data loading, preprocessing, model training, and prediction procedures are guaranteed to be independent. This modularity facilitates future scalability and improves code maintainability, making it simple to incorporate more disease prediction models.

The Streamlit uses caching strategies to maximize efficiency, preventing the need to recalculate models and preprocessing operations with every user interaction. In a nutshell, the Multiple Disease Prediction System provides a user-friendly and effective tool for machine learning-based early disease identification of a variety of illnesses.

VI. Conclusion

By assisting in early diagnosis and disease management, the Multiple Disease Prediction System (MDPS) demonstrates how artificial intelligence has the potential to revolutionize healthcare. The system examines a variety of health indicators to forecast the probability of diseases like Parkinson's, diabetes, or heart disease by utilizing a number of machine learning models, each of which is tailored for a particular ailment. In the end, this skill improves healthcare delivery by helping medical practitioners make quicker, better-informed judgments. Additionally, it offers insightful information that facilitates early intervention, which is advantageous to both physicians and patients. The system's predictions are guaranteed to be accurate and useful in real-world situations through the use of dependable machine learning models, such as Support Vector Machines and Logistic Regression.

Because of its scalable architecture, the MDPS can be modified for upcoming improvements, including adding more diseases or improving the models using more advanced methods. Even for those with little technical knowledge, the system is easy to use because to the interactive interface created using Stream-lite. To sum up, the Multiple Disease Prediction System shows how machine learning can transform medical diagnostics and improve patient care by making it more efficient, accessible, and tailored to each patient's needs.

References

1. Ahmad, L. G., Kalhori, S. R., Malaekheh, N. B. & Tavakkoli-Moghaddam, R. (2018). Machine learning techniques for predicting diabetes. *Journal of Medical Systems*, 42(12), 235. <https://doi.org/10.1007/s10916-018-1072-6>
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1721-1730)*. <https://doi.org/10.1145/2783258.2788613>
4. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, R., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias, and clinical safety. *BMJ Quality & Safety*, 28(3), 231-237. <https://doi.org/10.1136/bmjqs-2018-008370>
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794)*. <https://doi.org/10.1145/2939672.2939785>
6. Chen, Y., Li, X., & Xu, S. (2017). Predicting multiple diseases based on behavioral data using machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(2), 1231-1241. <https://doi.org/10.1109/TNNLS.2016.2597167>
7. Chicco, D., Jurman, G., & Moschitti, A. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20, 16. <https://doi.org/10.1186/s12911-020-1023-4>
8. Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine*, 372(9), 793-795. <https://doi.org/10.1056/NEJMp1500523>
9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. <https://doi.org/10.48550/arXiv.1702.08608>
10. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. <https://doi.org/10.1038/nature21056>

11. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29. <https://doi.org/10.1038/s41591-018-0316-z>
12. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3), 59-79. <https://doi.org/10.1609/aimag.v31i3.2303>
13. Foster, K. R., Koprowski, R., & Skufca, J. D. (2020). Machine learning, medical diagnosis, and biomedical engineering: The view of a clinical engineer. *IEEE Pulse*, 11(3), 45-49. <https://doi.org/10.1109/MPULS.2020.2986334>
14. Gorunescu, F. (2011). *Data mining: Concepts, models, and techniques*. Springer. <https://doi.org/10.1007/978-3-642-19721-5>
15. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y. (2017). Artificial intelligence in healthcare: Past, present, and future. *Stroke and Vascular Neurology*, 2(4), 230-243. <https://doi.org/10.1136/svn-2017-000101>
16. John, R., Aro, A. Q., Mathew, R., & Manju, G. (2019). Multidisease prediction using machine learning: A new horizon in healthcare. *Journal of Medical Systems*, 43(2), 29. <https://doi.org/10.1007/s10916-019-1168-8>
17. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17. <https://doi.org/10.1016/j.csbj.2014.11.005>
18. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
19. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A. W. M., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88. <https://doi.org/10.1016/j.media.2017.07.005>
20. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 26094. <https://doi.org/10.1038/srep26094>
21. Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501-507. <https://doi.org/10.1038/s42256-019-0114-4>
22. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. <https://doi.org/10.1126/science.aax2342>
23. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Dean, J., Maisel, M., Wing, P., & Ewig, M. (2018). Scalable and accurate deep learning for electronic health records. *NPJ Digital Medicine*, 1(1), 18. <https://doi.org/10.1038/s41746-018-0029-1>
24. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges, and the future. *Current Medical Imaging Reviews*, 14(5), 574-585. <https://doi.org/10.2174/1573405614666180907121327>
25. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(7), 1110-1114. <https://doi.org/10.1093/jamia/ocaa018>
26. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
27. Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., & Wong, S. Y. (2020). Applications of artificial intelligence in medical education and practice: The future is now. *Frontiers in Medicine*, 7, 149. <https://doi.org/10.3389/fmed.2020.00149>
28. Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1), 7-30. <https://doi.org/10.3322/caac.21590>
29. Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Jelley, A., Peterson, K., Rees, G., Laing, C., Shetty, S., Tarte, S., ... De Fauw, J. (2019). A clinically applicable approach to predicting acute kidney injury in hospitalized patients. *Nature*, 572(7767), 116-119. <https://doi.org/10.1038/s41586-019-1390-1>
30. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
31. Wang, F., Casalino, L. P., & Khullar, D. (2020). Deep learning for multi-disease prediction. *Journal of Biomedical Informatics*, 112, 103604. <https://doi.org/10.1016/j.jbi.2020.103604>
32. WHO. (2021). Non-communicable diseases. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
33. Zhang, Z., Missing, C., Nevadunsky, N., Lutomski, C., Taylor, A., Henry, A., & Anagnostopoulos, C. (2021). Missing data in machine learning: A Comprehensive Review. *IEEE Access*, 9, 95690-95709. <https://doi.org/10.1109/ACCESS.2021.309398>