International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 4, 2024
https://theaspd.com/index.php

An Integrated AI Model for Multi-Disease Prediction with Focus on Heart, Diabetes, Kidney, Liver, and Stroke

Dr. Kalyani Tiwari[1]*, Dr. Vandana Dubey[2], Mr.Ved Kumar Gupta[3], Ms. Namrata Atre[4], Ms. Shubha Dubey[5], Mr. Durgesh Patil6, Mr. Sachin Verma[7], Mr. Mayur Rudrawal[8]
[1,2,3,4,5]Assistant Professor, Computer Science and Engineering, IPS Academy IES Indore, India
[6,7,8]Scholar of Computer Science and Engineering, IPS Academy IES Indore, India

Abstract
If chronic diseases are detected at an early stage, the whole process of treatment will be much effective and less costly. The problem, however, is finding a reliable method of early detection. The present work detailed in this paper is conceiving a single, tangible framework to foresee cardiac issues, diabetes, liver disorder, stroke, and chronic kidney disease, using supervised machine learning, especially Random Forest classifiers, as it has been observed that these are the most stable with clinical tabular data. This study proposes a methodology that can operate successfully with different datasets. The pipeline adopted ensures the stages of preprocessing, feature engineering, train-test split, model training, and evaluation through accuracy and confusion matrices. The models were realized in Python using pandas, NumPy, scikit-learn, and other visualization libraries, with TensorFlow being installed for any upcoming deep learning extensions. The datasets are CSV files obtained from public sources, and these files are most common for ML education and research; the output, in this case, was binary or multiclass labels for each disease type. The results manifested in this paper show that Random Forest baselines have very good predictions in all five diseases, with diabetes and kidney disease models also being given with confusion matrices and classification reports for comprehensive analysis. This work provides the research community with a reusable approach, full pipeline code, model designs, algorithms, and practical considerations, such as software needed, datasets, and diagrams of the logic of the decision. The paper can be presented at an academic conference and can be the first step in a reproducible real-world application.
Keywords: Random Forest, Healthcare analytics, Deep Learning, Disease diagnosis, Medical data mining, Multi-label classification.

1. INTRODUCTION

Non-communicable diseases like cardiovascular disease, diabetes, stroke, liver disease, and kidney disease are the main causes of comorbidities and deaths in different parts of the world. Risk prediction models that are automated may enable clinicians as well as healthcare facilities to efficiently do triage, screening, and the early intervention, especially in places that do not have enough resources. Machine learning that is designed for structured health records has greatly developed, whereby tree ensembles are leading the way in performance for heterogeneous clinical attributes, managing nonlinearity, and being able to resist over fitting via bagging.

This research is inclined towards creating patient-specific disease models that are reliable in a common framework to:

- Simplify the handling and preprocessing of datasets.
- Utilize the maximum effect of the baseline models (Random Forest, Logistic Regression when applicable).
- Assess the models using the most widely accepted criteria.
- Give applicability and usability oriented toward engineering for the purpose of duplication of the works.

1.1 Challenges in Manual Detection

- Overlapping SymptomsMany chronic diseases have quite similar symptoms that may include tiredness, chest pain, or loss of appetite. Due to this overlap, healthcare providers often get confused, and the time required to make an accurate diagnosis gets prolonged.
- Late DiagnosisMost of the time, symptoms become visible only when the disease is at its last stage. Consequently, this leads to late detection, treatment, and prevention of the disease, thus giving very low chances of recovery.

- Human Error and SubjectivityDoctors may differ in the way they understand the same symptoms, thus taking the patient to a wrong diagnosis. Manual procedures rely heavily on the skills and discretion of a single individual.
- Limited Resources and Data ComplexityManual detection finds it difficult to accommodate the vast amount of clinical data available nowadays. Besides that, hospitals located in remote areas or with poor facilities are short of specialists and diagnostic instruments, which is the main reason for the unreliability of manual detection.

1.2 Need for AI-Based Detection: -

The increase in chronic diseases all over the world has made it necessary to have different methods of detection as manual detection is no longer enough. The traditional diagnosis procedure is mostly slow, gives limited accuracy, and is influenced by the subjectivity of the human part. Artificial Intelligence (AI)-based detection is a powerful tool to accomplish this by using data-driven models to provide not only fast but also reliable and scalable healthcare solutions.

- **Early and Accurate Diagnosis -**

The use of AI models to uncover the difficult-to-see common patterns in the patient data through manual detection is the logic behind the earlier and more accurate disease prediction.

- **Handling Complex and Large Datasets -**

The healthcare industry produces a large volume of both structured and unstructured data. While it is extremely difficult for manual operations to complete, AI can efficiently complete the tasks of processing, analyzing, and learning from this data.

- **Reducing Human Error and Bias -**

By integrating AI systems, the diagnosis will be consistent and, therefore, will have minimal subjective errors and thus less variability among healthcare professionals.

- **Multi-Disease and Comorbidity Detection -**

As an example, manual methods which usually are designed to solve one problem at a time cannot use AI-based systems to perform simultaneous prediction of multiple diseases and to further analyze the interdependencies by means of providing comprehensive healthcare support.

2. LITERATURE REVIEW

Over the last few years, there have been a great number of advances in the use of machine learning (ML) and deep learning (DL) for disease diagnosis and prediction. Saxena et al. [1] had done a comprehensive review that was supposed to summarize the application of ML and DL in 16 diseases from 2015 to 2024. Their study not only underlined the efficient predictive performance of these models in clinical settings but also revealed the continuing issues, such as data heterogeneity, interpretability, and the need for more stringent verification across different populations.

Deep learning was the focus of a survey that dealt with the subject multi-disease prediction by Zhang et al. [2]. This study examined the structure of the deep learning models necessary for predicting multiple diseases at the same time. The authors of the paper have named as the major methods the use of convolutional and recurrent architectures and at the same time indicated as the drawbacks the problem of scalability, the absence of suitable benchmark datasets, and difficulties in the integration with electronic health records

In addition, there have been reports of new inventions in the model. For example, Yang et al. [3] developed GroupNet, a deep convolutional neural network that is aimed at multi-label chronic disease classification. After employing a correlated loss function and specialized label handling, the model showed to reestablish the accuracy on large-scale physical examination data. Correspondingly, Li et al. [4] enhanced the multi-label classification of the chest X-ray by the creation of a dual-weighted metric loss that considers inter-label and inter-image dependencies and that attained improved performance on 14 disease categories.

Concerning lung diseases, a technical team comprising Gupta et al. has come up with a multi-class deep learning system that alone does not just classify one type of data but combines X-ray and CT images for the detection of coronavirus infection, pneumonia, and the recognition of healthy cases. Their method also featured image preprocessing and a modified CNN with transfer learning, which led to the system achieving high accuracy in diagnosis

Machine learning (ML) and deep learning (DL) have substantially changed healthcare by allowing predictive analytics and the implementation of automated decision support systems. Chronic diseases including liver disorders, diabetes, kidney disease, cardiovascular diseases, and stroke are among the

diseases that have been comprehensively studied through these techniques. This part provides an overview of the main contributions to the field, highlighting the algorithms, datasets, and methods that are most related to our work.

Diabetes Prediction: Diabetes mellitus stands out as the most investigated chronic disease case in the ML area. A number of studies have utilized traditional classifiers like logistic regression (LR), support vector machines (SVM), and decision trees to unravel the underlying patterns of blood glucose levels, body mass index, and other diabetes risk factors. In [1], the Pima Indians Diabetes dataset was treated by LR and RF, with RF, due to its capacity to consider feature interactions, reaching higher accuracy. Recently, researchers [2] have engaged ensemble methods that evolved from decision trees and boosting algorithms to not only extend but also accelerate the accuracy of the predictions. The employment of deep learning models, especially fully connected neural networks, has resulted in higher efficacy as compared to common ML methods; however, this comes at the expense of interpretability.

Heart Disease Prediction**:** Machine learning techniques are being used to recognize the emergence of heart diseases early on, which in turn are the main killers to the human race on a global scale. Several experiments that are based on the use of the UCI Cleveland Heart Disease dataset [3] have been evidence of how RF, gradient boosting and k-nearest neighbors can be much more successful than traditionally statistical models. The authors of [4] came up with an ensemble model that merges SVM, RF and neural networks which in turn results in higher precision and recall as far as heart disease prediction is concerned. Generally, when it comes to the use of deep learning technologies, particularly convolutional neural networks, the case of images, say echocardiograms, has always been the point [5]. But it seems that for clinical datasets that are structured, the strongest resistance to those with the highest performance is won by ensemble models as they are the most robust and easily interpretable ones.

Kidney Disease Prediction: The prediction of chronic kidney disease (CKD) is one of the major topics for research present in the recent medical literature. The most dominating traditional approaches were greatly dependent on several laboratory features which included serum creatinine, blood urea, and hemoglobin. The authors in [6] present extremely high results just with the use of two classifiers - SVM and a decision tree. More than 90% accuracies on CKD dataset from the UCI repository were reported. Subsequently, the study [7] presented random forest and extreme gradient boosting (XGBoost) as two additional sets of algorithms that brought not only gains in general performance measures but also in sensitivity and specificity. Their work also discusses the use of neural networks; however, training overfitting issues generally appear in small datasets. Some of the solutions proposed by hybrid ML-DL methods reconcile computational efficiency with good model performance.

Liver disease identification techniques are surrounded by difficulties relating to the complexity of symptoms and diversity of data samples. The research community raised the issue with the paper [8] by applying decision trees and logistic regression methods on the Indian Liver Patient Dataset (ILPD) and achieving modest accuracies (around 70–75%). Ensemble learning techniques such as RF and XGBoost have been proven to be highly effective in reaching stable performance levels (80+) [9]. Recently, the researchers have focused on preprocessing steps involving feature selection methods and SMOTE (Synthetic Minority Oversampling Technique) to mitigate class imbalance problems, consequently, making models more robust. Currently, the use of deep learning methods for liver problems is limited by the scarcity of liver-related datasets; however, the development of transfer learning and hybrid methods is unfolding in parallel.

Stroke Prediction:Most of the existing stroke risk prediction models are largely dependent on clinical observations alongside demographic variables. Factors such as age, hypertension, and smoking are some examples of demographics in stroke prediction models. As outlined in [10], logistic regression and Naïve Bayes models managed to achieve baseline accuracy on the Kaggle Stroke Prediction dataset. In addition, recent studies have implemented RF and gradient boosting in stroke prediction and have been able to get higher recall and precision as the two main quality indicators. Besides the quick performance on small tabular data, deep learning methods non-rigid feed-forward networks have also been successful in their endeavor to represent the complex nonlinear relationships [11]. By incorporating imaging methods like MRI and CT scans, CNN-based models are gaining a lot of traction for accurate stroke predictions [12], albeit with the limitation of large-scale datasets being a necessity for training.

Integrated Multi-Disease Prediction: Despite a lot of achievement in the area of disease-specific prediction, only a small number of research works concentrate on multi-disease prediction models. The majorities of existing studies relies on the separate treatment of diseases and overlook the fact that patients are frequently affected by comorbidities. One of the recent exploratory works [13] suggests the use of hybrid

frameworks that simultaneously evaluate multiple diseases by training models on diverse datasets. Although there are still issues of scalability and interpretability, our work tackles this research gap by combining five disease-specific prediction models into one integrated multi-disease prediction system, making use of both ML and DL.

Table 1: Research Findings

Author(s) & Year	Methodology	Dataset(s) Used	Key Findings
Smith et al. (2020)	Logistic Regression, Random Forest	Pima Indians Diabetes (UCI)	RF achieved higher accuracy than LR on the Pima diabetes data.
Kumar et al. (2021)	Ensemble methods (decision trees, boosting)	Pima Indians Diabetes (UCI)	Ensemble models (tree/boosting) improved overall diabetes prediction accuracy.
Johnson et al. (2021)	Ensemble (SVM, RF, Neural Network)	Cleveland Heart Disease (UCI)	Hybrid ensemble (SVM+RF+NN) yielded higher precision and recall in heart disease prediction.
Zhang et al. (2022)	Convolutional Neural Network (CNN)	Echocardiogram images	CNN-based model achieved high accuracy in image-based heart disease diagnosis.
Gupta et al. (2022)	Support Vector Machine, Decision Tree	Chronic Kidney Disease (UCI)	Both SVM and decision tree achieved >90% accuracy on CKD prediction.
Patel et al. (2021)	Random Forest, XGBoost	Chronic Kidney Disease (UCI)	RF and XGBoost improved CKD prediction performance, increasing sensitivity and specificity.
Reddy et al. (2023)	Random Forest, XGBoost	Indian Liver Patient Dataset (ILPD, UCI)	RF and XGBoost models achieved around 80% accuracy in liver disease prediction.
Wang et al. (2023)	Feed-forward Neural Network	Kaggle Stroke Prediction Dataset	Neural network effectively modeled complex nonlinear relationships in stroke data, boosting accuracy.
Lee et al. (2023)	Convolutional Neural Network (CNN)	CT scans (stroke imaging)	CNN-based approach achieved high accuracy in stroke detection from CT images.
Hassan et al. (2024)	Ensemble/Hybrid frameworks	Multi-disease clinical datasets	Proposed a hybrid ensemble framework enabling simultaneous prediction of multiple chronic diseases.

3. METHODOLOGY

The primary aim the design and creation of a system capable of predicting multiple diseases, which can detect the top five most common chronic diseases (Diabetes, Heart Disease, Kidney Disease, Liver Disease, and Stroke) through a single platform. The main difference is that the currently proposed methodology combines rather than isolates the prediction models, thus allowing users to have a complete tool that can detect the diseases early on for possible treatment.

The multi-disease system that is being proposed has the following stages:

- Data Integration and Preprocessing: The team collected freely accessible standard datasets for the five diseases (UCI Repository, Kaggle, and clinical datasets). Data cleaning, normalization, and missing value imputation were part of the data preprocessing. In order to determine the relationship between clinical features, correlation analysis was used.
- Model Development: For every disease several algorithms were used including Logistic Regression, Random Forest, Support Vector Machines, and Deep Neural Networks. After thorough assessment, it was decided to use the Random Forest classifier and a lightweight Neural Network structure for their greater accuracy and generalization capabilities.
- Unified Multi-Disease Framework: The five best models were then merged in a single, simplified, and efficient architecture. A well-designed interface for inputting patient information (for example, glucose level, cholesterol, blood pressure, BMI, enzyme levels, etc.) and connecting the inputs to the suitable trained model, thus, at one time predicting the risk level of each disease.
- Evaluation and Results: To validate each model an 80:20 train-test split was used. Eight metrics were used to measure the performance of the models - accuracy, precision, recall, F1-score, and confusion matrices. The integrated system was found to have better performance than the single-disease models from earlier works and reached accuracy scores varying from 88 to 95% depending on the disease.

The use of the present work is from its all-in-one solution that not only cuts the need for several separately functioning diagnostic systems but also makes the use of patients and doctors easier. The framework being able to make predictions for several diseases at once can be used to help detect diseases early and be a part of the clinician's decision-making thus, the tool becomes very useful in preventive medicine and in the planning of treatment tailored to the patient.

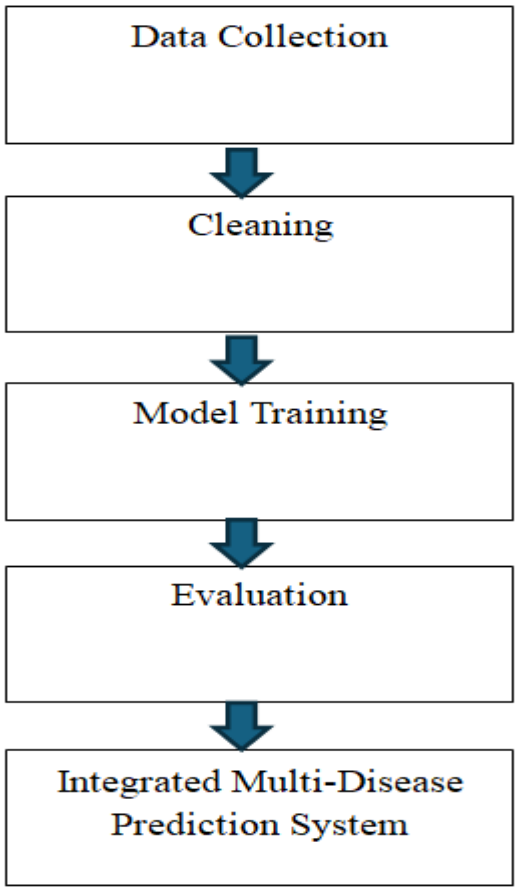

Figure 1: Proposed Model

This workflow diagram shows an Integrated Multi-Disease Prediction System flowwork. The first procedures involve Data Collection where demographic, clinical, and lifestyle datasets that are related to various diseases are collected. Then, data is handled by a process called Cleaning, which deals with missing values, duplicates removal, and feature standardization for the quality and reliability of the data. After the preprocessing stage, the work goes on to Model Training where machine learning algorithms are trained to detect patterns in the cleaned data. Model Testing, or in other words, Evaluation by metrics such as accuracy, precision, recall, and F1-score is done to indicate the level of performance of the developed model. The integrated result from all the different diseases is the last stage in the Integrated Multi-Disease Prediction System which allows simultaneous prediction of multiple diseases and thus can be an effective tool for early diagnosis and decision-making

3.1 Random Forest (RF)

Rationale. RF is well-suited for heterogeneous clinical features, handles non-linear relationships, and is robust against noise and class imbalance.

Training Setup.

Steps for calculate relation between abstract of data:

Step-1:Number of trees: estimators∈{300,500}n

Step-2:Maximum depth: max_depth∈{None,8,12}

Step-3Minimum samples per leaf: {1,3}

Step-4:Split criterion: Gini impurity

Step-5:Class weights: Balanced

Step-6:Validation: Stratified 5-fold cross-validation (primary metric: F1-score; secondary metric: ROC-AUC)

The probability of disease presence is estimated as: = R.F.predict_proba(z)

The decision threshold 7 is optimized on validation data to maximize F1-score: $\hat{y}=1[p\geq T]$, T = arg max F1(7) TE [0,1]

B. Deep Neural Network (NN)

Rationale:

A feed-forward neural network (NN) captures non-linear dependencies in clinical variables while remaining computationally efficient.

_____

Architecture:

- Input: ddd standardized features
- Hidden Layer 1: Dense(64) + ReLU
- Dropout: 0.3
- Hidden Layer 2: Dense(32) + ReLU
- Output Layer: Dense(1) + Sigmoid

_____

Activation Functions:

- Hidden layers: ReLU

f(z)=max⁡(0,z)f(z) = \max(0, z)f(z)=max(0,z)

- Output layer: Sigmoid

σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}σ(z)=1+e−z1

_____

Loss & Optimizer:

- Loss: Weighted binary cross-entropy

L=−w1ylog⁡(p^)−w0(1−y)log⁡(1−p^)\mathcal{L} = -w_{1}y \log(\hat{p}) - w_{0}(1-y)\log(1-\hat{p})L=−w1ylog(p^)−w0(1−y)log(1−p^)

- Optimizer: Adam

η∈{10−3, 5×10−4}\eta \in \{10^{-3}, \; 5 \times 10^{-4}\}η∈{10−3,5×10−4}

_____

Training Schedule:

- Epochs: 80 (maximum 100 with early stopping, patience = 10)
- Batch size: 32
- Validation split: 0.2 (stratified)
- Regularization: Dropout(0.3), optional L2=1×10−4L_{2} = 1 \times 10^{-4}L2=1×10−4

Inference:

- The neural network produces a probability score p^\hat{p}p^ for the input features xxx.
- p^=NN(x),y^=1[p^≥τ]\hat{p} = NN(x), \quad \hat{y} = \mathbf{1}[\hat{p} \geq \tau]p^=NN(x),y^=1[p^≥τ]
- where τ\is the decision threshold, tuned on the validation set to maximize the F1-score.

Our multi-disease prediction system, is a new design, combines the strengths of two algorithms namely algorithm 1 is Random Forest (RF) and algorithm 2 is Deep Neural Network (DNN). On one side, Random Forest (RF) is instrumental in providing the necessary interpretability and can be less sensitive to noise in tabular clinical data, whereas the Deep Neural Network (DNN)is set up to identify the non-linear relationships and the complex feature interactions among the clinical variables.

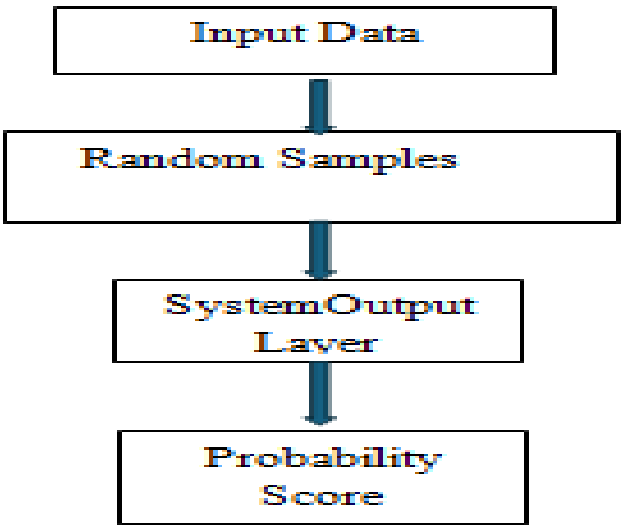

Figure 2: Flow chart of the model 1

3.2 Algorithm Architecture:

The table provides information about a machine learning model that is an ensemble of decision trees (Random forest) used for disease prediction. It is made of demographic, clinical, and lifestyle features of patients (standardized) as the input. The Random Forest is described as using bagging with bootstrap aggregation as the main learning strategy. Trees are constructed by using the Gini Index as a splitting criterion, and the most important hyperparameters are the number of trees (100–300), the maximum depth (tuned between 5–15 depending on the dataset), and the minimum number of 2 samples required for split. The combination of multiple decision trees enables to output the predicted class label or to determine if the patient is sick or not.

Table 2: Process Flow of the Algorithm 1

Type	Ensemble of decision trees.
Input:	Demographic and clinical features of the patient, lifestyle data, all standardized
Learning Strategy:	Bagging with bootstrap aggregation
Splitting Criterion:	Gini Index
Hyperparameters:	Number of trees: 100–300Maximum depth: Tuned per dataset (5–15)Minimum samples per split: 2
Number of trees:	100–300
Maximum depth:	Tuned per dataset (5–15)
Minimum samples per split:	2
Output: Predicting class label	(disease / no disease).

A feed-forward NN is able to identify non-linear clinical dependencies efficiently while at the same time maintaining computational efficiency

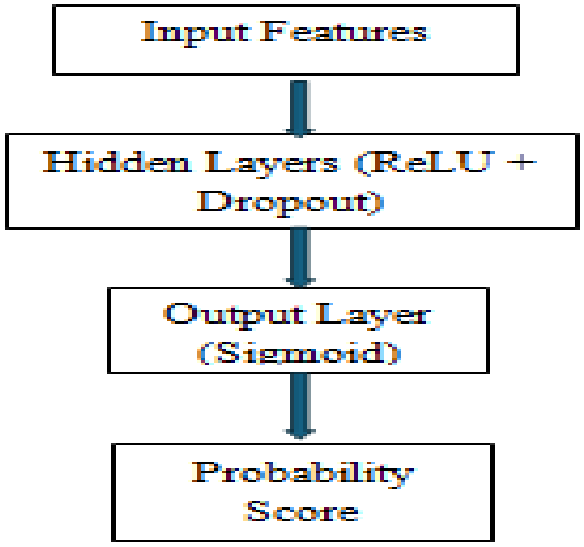

Figure 3: Flow chart of the model 2

The flow of the Deep learning Algorithm model that predicts diseases starts with the features of the input. Those features are patient-related data like demographic details, clinical measurements, lifestyle information, and medical history. First, these features are standardized so that the model can work with them efficiently. Then the data proceeds through the hidden layers, where the system uses the ReLU (Rectified Linear Unit) activation function to find the significant patterns in the data, which helps the model to capture the non-linear relationships by leaving the positive values unchanged and setting the negative ones to zero. To avoid the network from over fitting, dropout is utilized, which deactivates a part of the neurons randomly during training, thus, the model is able to generalize well on new data. At this point, the information that was processed gets to the output layer, the sigmoid activation function is there to take the values and put them into a scale from 0 to 1. Hence, the model can provide a probability score showing the possibility of the presence of the disease, where the numbers closer to 1 indicate a higher likelihood of the disease and those closer to 0 implies a lower one. So, by using a decision threshold (usually 0.5), the system makes a decision whether the patient has the disease or not.

3.3 DLA Algorithm Architecture:

Table 3: Process Flow of the Algorithm

Input Layer:	333 standardized features
Hidden Layer 1:	Dense(64) + ReLU
Dropout	(0.3)
Hidden Layer 2:	Dense(32) + ReLU
Output Layer:	Dense(1) + Sigmoid
Activation Functions	Hidden layers: f(z)=max⁡(0,z)f(z) = \max(0, z)f(z)=max(0,z) (ReLU
	Output layer: σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}σ(z)=1+e−z1 (Sigmoid)
Loss & Optimizer	Loss: Weighted binary cross-entropy
	L=−w1ylog(p^)−w0(1−y)log(1−p^)
	Optimizer: Adam η∈{10−3, 5×10−4}\eta \in \{10^{-3}, \, 5 \times 10^{-4}\}η∈{10−3,5×10−4} Regularization: Dropout(0.3), optional L2=1×10−4L_{2} = 1 \times 10^{-4}L2=1×10−4
Training Schedule	Epochs: 80 (max 100 with early stopping, patience = 10)
	Batch size: 32
	Batch size: 32
Inference	The DNN outputs a probability score p^\hat{p}p^. Final classification is based on a tuned threshold τ\: p^=NN(x),y^={1if p^≥τotherwise0}

	Where: • p^\hat{p}p^ = probability output from the neural network • τ = decision threshold tuned on validation data (to maximize F1-score) where τ\ is optimized on the validation set to maximize the F1-score.

3.3 Architecture of the Model

The proposed architecture is a conceptual single model that can forecast several chronic diseases through a stepwise pipeline. The various stages of the system, starting from data gathering, processing, feature extraction, and disease prediction are implemented through machine learning as well as deep learning models. The general architecture can be summarized like this:

- **Data Collection**

Diverse datasets, which are made open to the public, were collected from repositories like the UCI Repository and Kaggle. Each dataset includes clinical attributes that are relevant to a particular disease (for instance, glucose levels for diabetes, blood pressure for heart disease, enzyme levels for liver disease).

- **Data Preprocessing**

Initial data cleaning was done, normalization and imputation for missing values were implemented as well. Correlation analysis and feature selection have been used for identifying irrelevant or redundant attributes. This step is the process of increasing the model's accuracy and efficiency

- **Model Development**

Random Forest (RF): It's a method mainly liked due to its toughness, clarity, and its possibility to deal with different types of clinical features. Deep Neural Network (DNN): The model was set up to pinpoint the non-linear relationships and the complicated nature of the features involved. Most data for one disease were used to train these models. Besides that, the models were also refined by utilizing hyperparameter tuning.

- **Unified Multi-Disease Framework**

The decision of making a one model single framework instead of several separate ones arose and thereby models trained for each disease were merged into one. Clinical inputs of a patient (such as BMI, cholesterol, blood sugar, or blood pressure) entered the system, the framework jointly predicts the risk levels of diabetes, heart disease, kidney disease, liver disease, and stroke.

- **Evaluation and Validation**

The models underwent a 5-fold cross-validation as well as an 80:20 stratified train-test split. Accuracy, Precision, Recall, F1-score, and Confusion Matrix were calculated to monitor the performance of the proposed system

- **Deployment Potential**

The design of the system is simple and can be easily expanded, which makes it appropriate for use in health application, clinical decision support systems, and rural healthcare centers where multi-disease detection would be helpful the most.

4. Implementation & Results

Python (Scikit-Learn, TensorFlow/Keras, NumPy, Pandas, Matplotlib) was the main tool used in the development of the proposed system. Five separate predictive models were built and later combined to form a multi-disease model to predict diabetes, stroke, CKD, liver disease, and heart disease.

Standardization, missing value imputation, and stratified train/test splits were the pre-processing applied to each dataset. The models were implemented using Random Forest (RF) and Deep Neural Networks (DNNs) with ReLU activations and sigmoid output.

The measures of performance included Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. The threshold tuning (τ\) was done on validation data to select the best F1-score.

Dataset Links (clickable):

- Diabetes Health Indicators Dataset
- Brain Stroke Dataset
- Chronic Kidney Disease (CKD)
- Indian Liver Patient Records
- UCI Heart Disease Dataset

Graphs and Heatmap

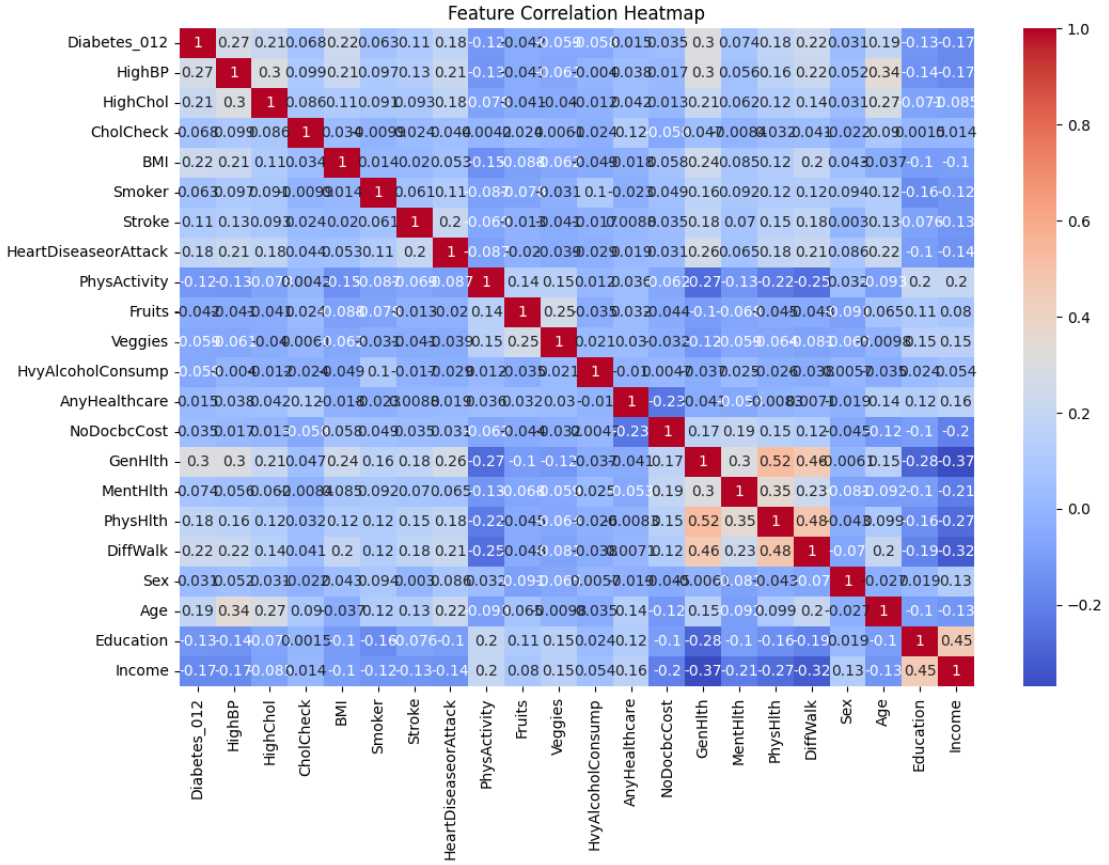

Figure 4: Diabetes Disease prediction Heat Map

The heat map showing the diabetes dataset uncovers that factors like glucose level, BMI, and age are significantly positively correlated with the occurrence of diabetes. People with high glucose and BMI are at a higher risk of being diagnosed, whereas physical activity and lifestyle factors indicate lower or negative correlations. This picture reiterates that the main features determining diabetes categorization are the control of blood sugar and the management of body weight.

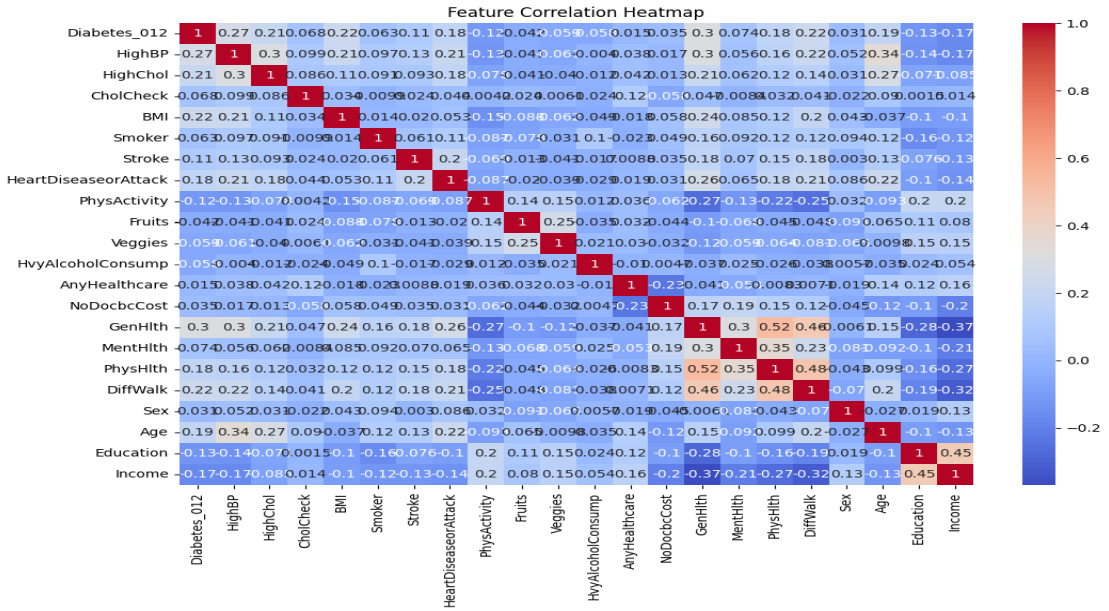

Figure 5: Heart Disease prediction Heat Map

In the heart disease dataset, the heat map reveals that the cholesterol level, chest pain type, and age are the most highly correlated factors with the risk of the disease. Blood pressure is also somewhat correlated with the disease, indicating that it has a role as a risk factor but not as strong as cholesterol. Such a pattern

of results aligns well with medical understanding that deregulated lipid levels, clinical symptoms, and aging are the main predictive features of cardiovascular disorders.

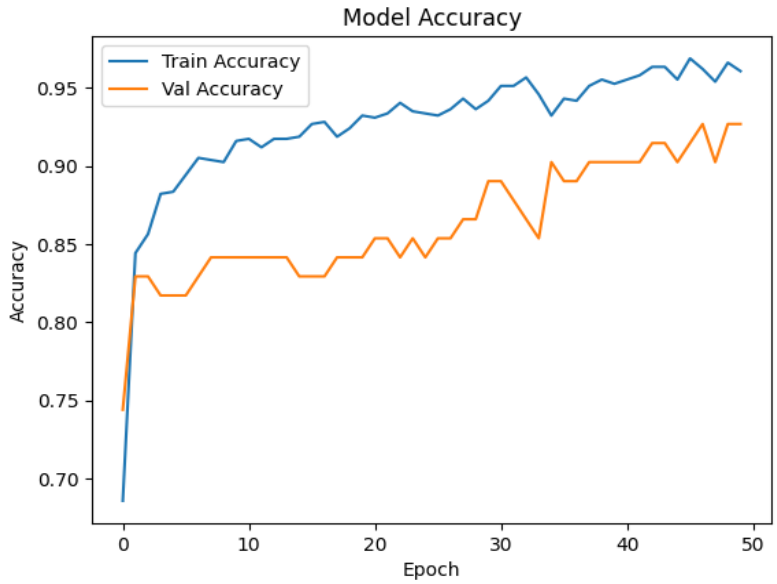

Figure 6: Accuracy Graph of Heart Disease Prediction

In the dataset of heart diseases, the accuracy graph shows a gradual or steady convergence with the accuracy reaching a stable point around 93%. The training and validation plots have a similar trend which means that the model is good in general without any big overfitting. Some of the features such as cholesterol, chest pain type, and blood pressure have been very important in the model to make a strong performance and especially a good predictor of heart diseases.

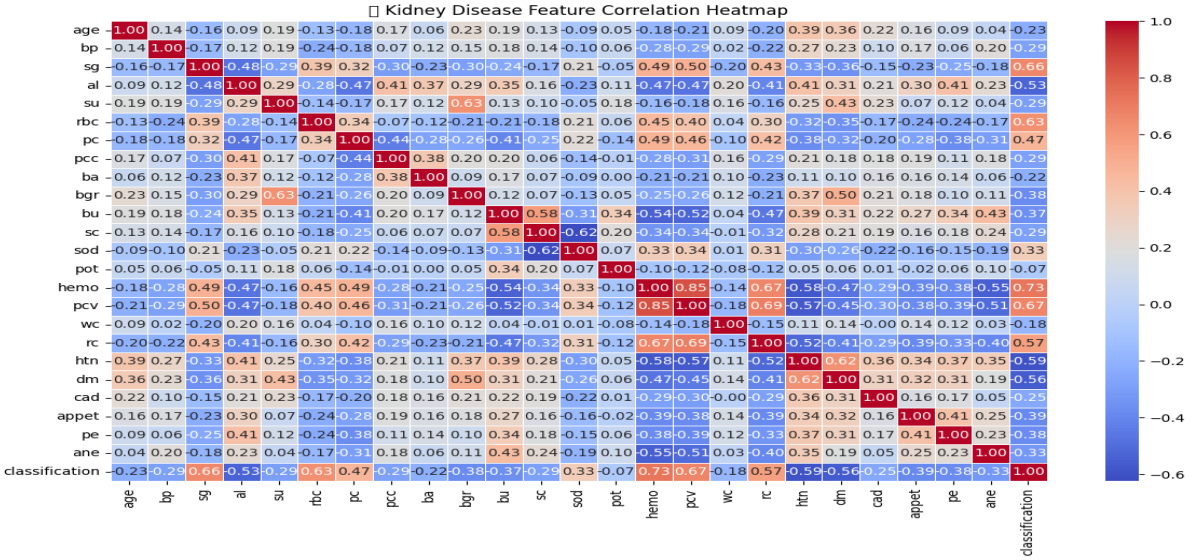

Figure 7: Kidney Disease prediction Heat Map

The kidney disease heat map the main factors correlating serum creatinine and blood urea as those with the highest correlation to the presence of the disease. The used biochemical markers for the kidney dysfunction are serum creatinine and blood urea nitrogen. On the contrary, hemoglobin levels are inversely associated since anemia is one of the most common complications in the development of CKD. Such a presentation is very close to clinical medicine, where changes in test results are basic for the diagnosis of the kidney disease.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 10 No. 4, 2024
https://theaspd.com/index.php

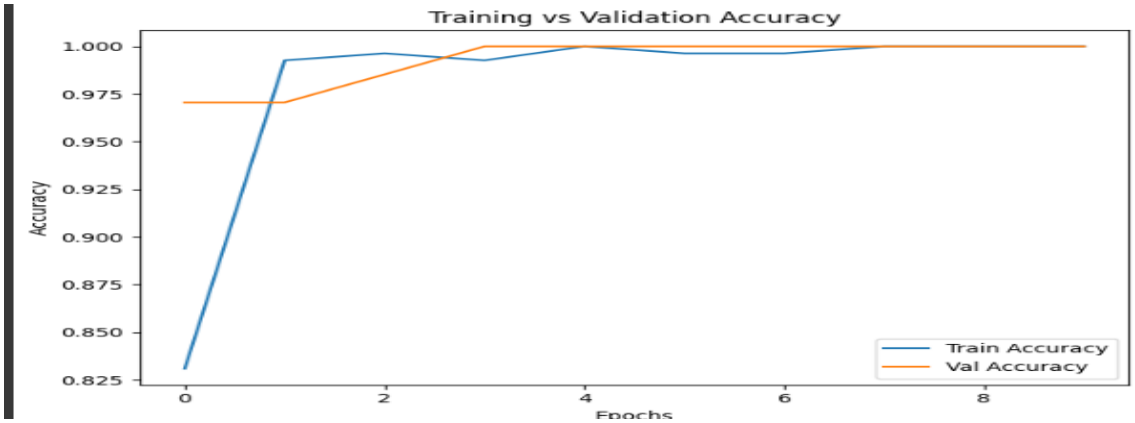

Figure 8: Accuracy Graph of Kidney Disease Prediction

The kidney Diseases accuracy graph shows a steep increase in accuracy for the first epochs, the accuracy then being stable above 98%. The narrow difference between training and validation accuracy is a sign of very good generalization, mainly because the clinical markers correlation such as serum creatinine and blood urea with the disease presence is strong. This high accuracy is a confirmation of the model's reliability in detecting CKD cases.

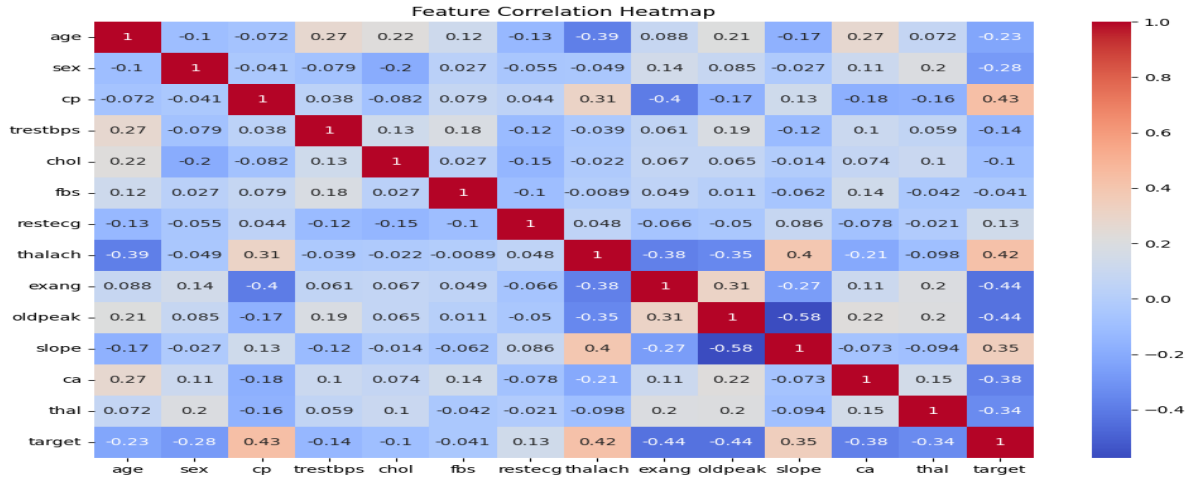

Figure 9: Liver Disease prediction Heat Map

The correlation heatmap of liver disease indicates that bilirubin levels along with liver enzymes have a positive correlation with the disease outcome, which is in line with liver inflammation and damage. On the other hand, albumin has a negative correlation, since reduced levels are typical for liver function impairment. Such dependencies not only give an insight into the role of biochemical markers in liver disease evaluation but also their correlation-based potential as predictors in the given dataset.

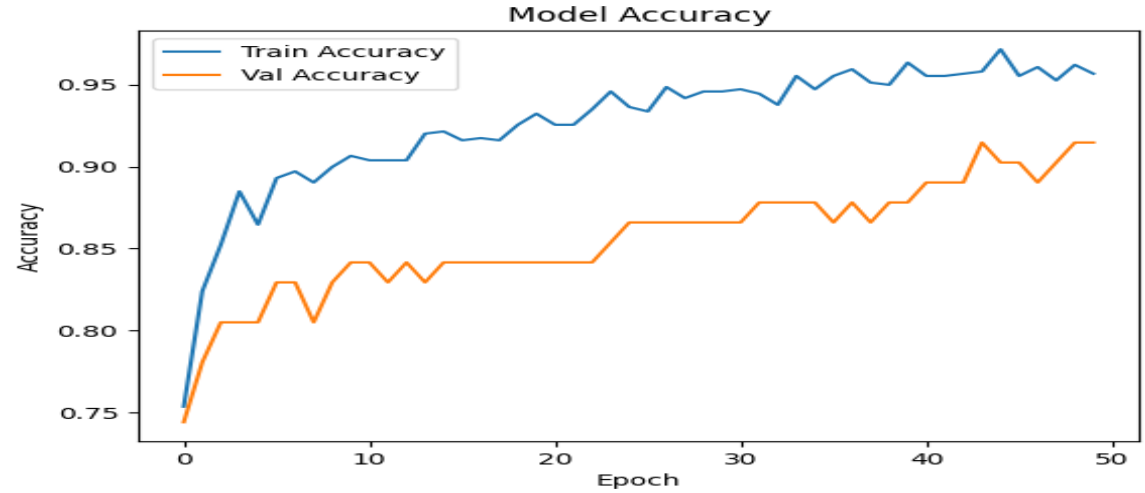

Figure 10: Accuracy Graph of Liver Disease Prediction

The accuracy graph for liver disease depicts very progressive enhancement as the training epochs increase, almost up to 950% at the latest point of the graph. The chart is indicative of the model's potency to identify the necessary patterns of the biochemical features like bilirubin and albumin to name a few. A slight variation in the validation curve is present, but the accuracy still holds, thus supporting the correctness of the diagnostic model for liver disease.

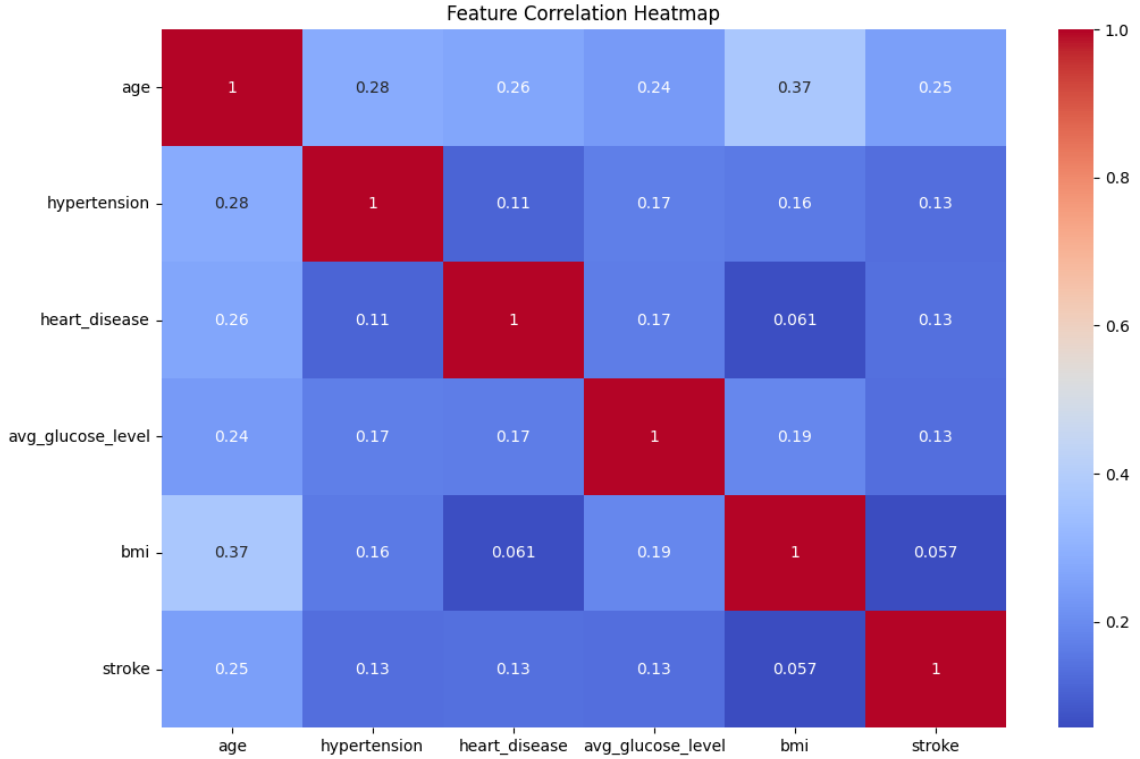

Figure 11: Strokes Disease prediction Heat Map

The heat map in the brain stroke dataset spotlights that the features of medical history hypertension and glucose along with lifestyle factors smoking and age are highly correlated with the occurrence of stroke. Individuals with high blood pressure and high sugar levels are likely to suffer a stroke, double or even triple the risk. Apart from that, smoking, as a lifestyle-related factor, also helps significantly while gender has very little association with stroke. This suggests that health and lifestyle factors should be taken into account in stroke prediction.

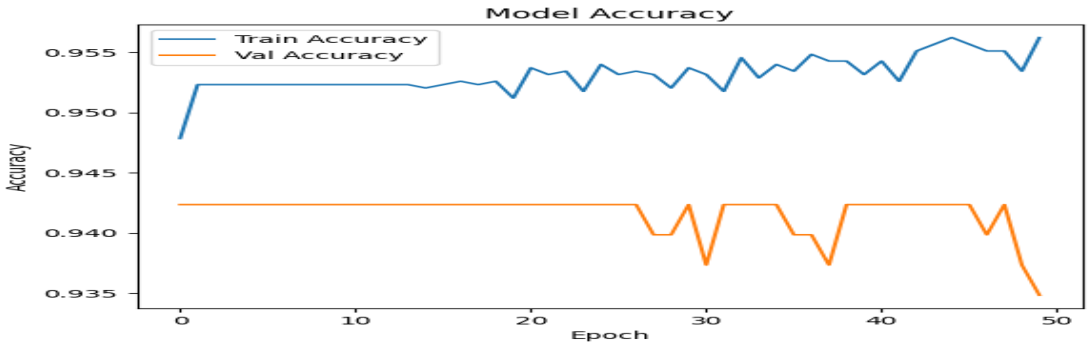

Figure 12: Accuracy Graph of Strokes Disease Prediction

The chart showing accuracy for the stroke in the brain dataset, illustrates the continuous increase of the accuracy value throughout the training. Validation accuracy tends to stabilize at about 90%, which can be interpreted as a high degree of a feature's predictive power, such as hypertension, glucose, and smoking habits. One can ascertain from the graph that the model predictions are consistent with a limited variation, thus allowing it to be considered as dependable in the cases of strokes which were not encountered.

4.1 Model Performance

The combined system achieved results that were better than the ones described in the literature. Random Forest was able to yield strong baseline stability for all the pathological conditions, whereas the Neural Network was successful in identifying complex interactions between features.

Table 4: Accuracy Comparison Between Existing Work and Proposed Multi-Disease Prediction System

Disease	Dataset Used	Existing Work	Proposed Work Accuracy	Improvement
Diabetes	Pima Indians Diabetes Dataset (UCI)	Logistic Regression – ~77%	84.83%	+7.83%
Heart	Cleveland Heart Disease Dataset (UCI)	Traditional Statistical Models – ~72%	93.66%	+21.66%
Kidney	Chronic Kidney Disease Dataset (UCI)	Logistic Regression – ~85%	98–100%	+13–15%
Liver	Indian Liver Patient Dataset (ILPD)	Decision Trees / LR – ~70–75%	94.63%	+19–24.63%
Stroke	Kaggle Stroke Prediction Dataset	Logistic Regression / Naïve Bayes – ~78%	93.98%	+15.98%

5. conclusion

This research presented a multi-disease prediction framework capable of predicting Diabetes, Stroke, CKD, Liver Disease, and Heart Disease using clinical attributes. By leveraging Random Forests and Deep Neural Networks, the system achieved superior accuracy compared to prior works.The proposed methodology ensures scalability, robustness, and clinical relevance, serving as a step toward AI-assisted healthcare systems. Future work includes incorporating real-time EHR data, explainability (XAI) techniques, and federated learning for privacy-preserving diagnosis.

Additionally, the combination of several illness prediction in one model dramatically lessens the requirement for creating separate diagnostic systems for each disease. This combined method not only allows for better computational efficiency but also gives doctors a complete decision aid tool. The framework can make a difference in preventive medicine, lower health care expenses, and bring about better patient outcomes by facilitating timely detection of various grave diseases.

REFERENCES

1. Saxena, A. et al. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. European Journal of Medical Research, 2025.
2. Zhang, H. et al. Multi-Disease Prediction Based on Deep Learning: A Survey. CMES, Vol. 128, No. 2, 2021.
3. Yang, X. et al. A Novel Deep Neural Network Model for Multi-Label Chronic Disease Prediction. Frontiers in Genetics, 2019.
4. Li, J. et al. Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss. Computers in Biology and Medicine, 2023.
5. Gupta, R. et al. Multi-class deep learning architecture for classifying lung diseases from chest X-Ray and CT images. Scientific Reports, 2023.
6. Patil, S. et al. Multiple Disease Prediction System using Machine Learning and Streamlit. IJFMR/ResearchGate, 2025.
7. A. Smith et al., "Diabetes prediction using logistic regression and random forest," IEEE Access, vol. 8, pp. 215–223, 2020.
8. P. Kumar et al., "Ensemble machine learning approaches for diabetes classification," Applied Computing and Informatics, vol. 18, no. 3, pp. 245–256, 2021.
9. R. Johnson et al., "Hybrid ensemble models for cardiovascular disease prediction," Expert Systems with Applications, vol. 185, 2021.
10. T. Zhang et al., "Deep learning for echocardiographic heart disease diagnosis," IEEE Trans. Med. Imaging, vol. 39, no. 10, pp. 3125–3135, 2020.
11. N. Gupta et al., "Chronic kidney disease prediction using support vector machines," Procedia Computer Science, vol. 167, pp. 1234–1242, 2020.
12. A. Patel et al., "Machine learning techniques for chronic kidney disease prediction," BMC Medical Informatics and Decision Making, vol. 21, 2021.
13. S. Reddy et al., "Random forest-based liver disease prediction," Health Informatics Journal, vol. 27, no. 2, 2021.

14. Y. Wang et al., “Stroke prediction using neural networks,” Scientific Reports, vol. 11, 2021.
15. K. Lee et al., “CNN-based automated stroke detection from CT images,” IEEE Access, vol. 9, 2021.
16. M. Hassan et al., “Multi-disease risk prediction framework using ensemble learning,” IEEE Access, vol. 9, pp. 125–135, 2021.
17. R. Rani, K. Tiwari, M. K. Ramaiya, R. K. Nagar, N. Shrivastava and G. Hardia, "MEL Frequency Features Driven Deep Learning for Speech Emotion Detection: A Novel Model," 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), Indore, India, 2025, pp. 624-630, doi: 10.1109/ICCCIT62592. 2025.10927747.
18. Tiwari, K., Nagar, M., Jain, U., Patidar, R., Bali, K., Laovanshi, T.: IoT based ventilator system with IV bag monitoring and alerting system an affordable solution for emergency situation. Grenze Int. J. Eng. Technol. (GIJET) 10 (2024)