Machine Learning Approaches for Predicting AIDS Virus Infection: A Comparative Study using Random Forest and Neural Networks

Amrutha
Biotechnology
R V College of Engineering
Bangalore , India

Ananya S Padasalgi
Biotechnology
R V College of Engineering
Bangalore , India

Maanasa M G
Biotechnology
R V College of Engineering
Bangalore , India

Smrithi R Holla
Biotechnology
R V College of Engineering
Bangalore , India

Dr. Shivandappa
Biotechnology
R V College of Engineering
Bangalore , India

***Abstract*—**This review investigates the use of machine learning approaches, notably Random Forest and Neural Network classifiers, in the context of AIDS classification and digit identification using the MNIST dataset. The paper compares the performance of a Random Forest classifier and a Multi-Layer Perceptron (MLP) neural network on an AIDS classification dataset, emphasizing the significance of feature scaling and the impact of model design on classification accuracy. The Random Forest model was used to determine feature relevance, and the MLP classifier was trained and tested for accuracy in categorizing the binary outcome of HIV infection.

Furthermore, the study presents a proprietary neural network model built to identify handwritten digits from the MNIST dataset, emphasizing the model's architecture, which includes dense layers with ReLU activation and dropout regularization to reduce overfitting. The models were trained and tested, with performance indicators including accuracy and loss measured across epochs.

The results reveal that both traditional machine learning approaches and deep learning architectures can handle various datasets, highlighting their capabilities in binary classification and multiclass digit identification tasks. The results highlight the potential for integrating feature selection techniques with advanced neural network models to improve predicted performance in medical diagnostics and picture categorization.

Keywords—machine learning, formatting, google colab, neural network

I. INTRODUCTION

AIDS (Acquired Immune Deficiency Syndrome) is a dangerous pandemic disease that has a significant impact on society. It is a very serious disease that is actively spreading globally among humankind. HIV/AIDS is caused by infection with the human immunodeficiency virus (HIV virus) , which attacks and destroys the CD4 cells of our immune system [1] . HIV targets the white blood cells in the body, thereby making it easier for people infected with the virus to get sick with diseases like tuberculosis, infections and some cancers[2]. HIV can develop into AIDS if left untreated, after many years. When a person infected with HIV has a CD4+ count of less than 200 cells/µL , then the person is said to be having AIDS[1] . HIV virus uses CD4 cells to multiply and spread throughout the body[1].During the initial stages of HIV infection, a person might develop symptoms similar to influenza, which is then followed by a long period without any symptoms. As the HIV illness progresses , it interferes more and more with the immune system, thereby making the person weaker[1]. HIV is spread from the body fluids of an infected person, including blood, breast milk, semen and vaginal fluids. It can also be spread through unprotected sexual intercourse , contaminated blood transfusions, hypodermic needles[2]. It is not spread by kisses, hugs or sharing food. It can also spread from a mother to her baby via the placenta.[2].HIV can be prevented and treated with antiretroviral therapy (ART)[2]. Though antiretroviral treatment(ART) reduces the risk of death and complications from the disease, these medications are usually expensive and are associated with side effects[1].AIDS was first recognized in the US in June 5, 1981. Initial symptoms are followed by a stage called asymptomatic HIV or chronic HIV. If there is no treatment available then this stage of HIV infection can last from about three years to over 20 years (8 years on an average)[1]. We can reduce the risk of HIV infection by using a condom during sex and by regular testing for HIV and sexually transmitted diseases(STD's)[1].Though HIV/AIDS is treatable, there is no cure for HIV infection as of now. It is treated with antiretroviral drugs, which stops the virus from replicating in the body. Current antiretroviral therapy (ART) does not cure HIV infection but allows a person's immune system to get stronger. This helps them to fight other infections[2].

II. METHODOLOGY

This project is centered around building a powerful machine learning model aimed at accurately diagnosing HIV/AIDS using clinical and demographic data. By employing both the Random Forest Classifier and neural networks, the project aims to enhance diagnostic precision, thereby supporting healthcare professionals in making better-informed decisions.[3]

1. Data Collection and Preprocessing:
    - Dataset Compilation: Assemble a comprehensive dataset that includes a wide range of clinical features, demographic details,

and historical health data pertinent to HIV/AIDS.[3]
    - Data Cleaning and Preparation: Clean and preprocess the data to address any missing values, normalize features, and encode categorical variables. This step is crucial to ensure the data is well-suited for use with machine learning algorithms.[4]
2. Model Development:
    - Random Forest Classifier Implementation: Develop a Random Forest Classifier, chosen for its reliability and ability to manage high-dimensional data, to classify individuals as either infected or not infected with HIV/AIDS.[4]
    - Exploring Neural Networks**:** Investigate the use of neural networks to uncover complex patterns within the data that may be overlooked by traditional classifiers.[5]
3. Performance Evaluation:
    - Confusion Matrix and Metrics Analysis: Use confusion matrices to assess the performance of the models, calculating key metrics such as accuracy, precision, recall, and F1 score.[6] These metrics will be vital in determining how effective the models are.
    - Model Optimization: Perform hyperparameter tuning and cross-validation to fine-tune the models, aiming to boost performance while minimizing the risk of overfitting.[5]
4. Feature Importance Analysis:
    - Understanding Key Predictors: Analyze the importance of different features in predicting HIV/AIDS infection using the Random Forest model. This analysis will help identify critical factors that influence the disease's presence and could guide future clinical assessments.[6]
5. Comparison of Classifiers:
    - Model Performance Comparison**:** Compare the Random Forest Classifier with neural networks to determine which model offers better accuracy and reliability in diagnosing HIV/AIDS.[7]
6. Implementation and Deployment:
    - User Interface Development: Create an intuitive interface for healthcare professionals, allowing them to input patient data and receive predictions regarding HIV/AIDS status.[8]
    - Clinical Deployment Consideration**:** Explore possibilities for deploying the model in clinical settings to assist in real-time decision-making and improve patient outcomes.[8]

Random Forest Classifier Overview

Random Forest works by constructing multiple decision trees during the training phase and then using the mode of their predictions for classification tasks. This ensemble method boosts predictive accuracy and lowers the risk of overfitting, making it particularly well-suited for handling the complex datasets frequently encountered in medical research.

When applying Random Forest to predict HIV/AIDS infection, the following steps are typically involved:

1. Data Collection: Comprehensive datasets are compiled, often including a range of clinical features, demographic details, and historical health information from individuals.[7]
2. Data Preprocessing: The data is cleaned and normalized to address any missing values and ensure it is properly formatted for the Random Forest algorithm.
3. Model Training: The Random Forest Classifier is trained on a subset of the data, where a random selection of features is used to construct each decision tree. [8]This method helps to capture the underlying patterns associated with HIV infection.[9]
4. Model Evaluation: The model's performance is assessed using metrics like accuracy, precision, recall, and F1 score, which are calculated from confusion matrices[7]. These metrics provide insight into the model's effectiveness in predicting HIV status.[10]
5. Feature Importance Analysis**:** The Random Forest model also sheds light on the importance of different features, helping to identify critical predictors of HIV infection.[9]

Applications in HIV Prediction

Numerous studies have highlighted the effectiveness of the Random Forest Classifier in predicting HIV infection:

1. Identification of HIV Predictors: Research has demonstrated that Random Forest can successfully identify socio-behavioral factors associated with HIV status, aiding in targeted screening efforts. For example, a study utilizing data from the Population-based HIV Impact Assessment (PHIA) found that Random Forest could accurately identify key predictors like age, education level, and socio-economic status.[10]
2. Comparison with Other Models**:** In several studies, Random Forest has outperformed traditional models like logistic regression and other machine learning algorithms in predicting HIV infection. For instance, a study focusing on men who have sex with men (MSM) in China reported that Random Forest provided better performance metrics than logistic regression, underscoring its effectiveness in dealing with complex data.[9]
3. Integration with Other Techniques: Random Forest has also been successfully combined with other machine learning techniques, such as neural networks, to enhance predictive power. [7]This hybrid approach allows for capturing intricate patterns while retaining the

interpretability provided by feature importance analysis.[8]

Confusion Matrix: A Comprehensive Tool for Evaluating Disease Prediction Models

In the field of machine learning, the confusion matrix stands out as a vital tool for assessing the performance of classification models, especially in the context of disease prediction. This review explores the nuances of the confusion matrix and its application in a project focused on predicting HIV/AIDS infection using Random Forest Classifiers and neural networks.[11]

Understanding the Confusion Matrix

A confusion matrix is a table that summarizes how well a classification model performs by comparing its predicted outcomes to the actual outcomes. It provides a detailed overview of a model's accuracy by showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class[12]. This matrix allows for the calculation of key performance metrics such as accuracy, precision, recall, and F1 score, which are essential for evaluating the effectiveness of any disease prediction model.[13]

Application in HIV/AIDS Prediction

In the project aimed at predicting HIV/AIDS infection, the confusion matrix plays a crucial role in evaluating the performance of both the Random Forest Classifier and neural networks. The process involves several key steps:

1. Data Collection and Preprocessing**:** A comprehensive dataset is assembled, including clinical features, demographic information, and historical health data related to HIV/AIDS. The data is then cleaned and normalized to address any missing values and ensure it is suitable for the machine learning algorithms.
2. Model Training: The Random Forest Classifier and neural networks are trained on the preprocessed data to identify patterns associated with HIV infection. The Random Forest Classifier constructs multiple decision trees from random subsets of features, while neural networks use interconnected layers of neurons to learn complex relationships within the data.[11]
3. Performance Evaluation: The confusion matrix is employed to evaluate how accurately each model predicts HIV status. This involves calculating the number of true positives (correctly identified infected individuals), true negatives (correctly identified uninfected individuals), false positives (uninfected individuals incorrectly identified as infected), and false negatives (infected individuals incorrectly identified as uninfected) for each model.[14]
4. Comparative Analysis: The performance metrics derived from the confusion matrices—such as accuracy, precision, recall, and F1 score—are used to compare the effectiveness of the Random Forest Classifier and neural networks in predicting HIV infection. This comparison helps in determining which model is more suitable for the task.[12]
5. Feature Importance Analysis: The Random Forest Classifier also offers insights into the importance of various features in predicting HIV infection. By examining the confusion matrix and feature importance measures, critical predictors of HIV status can be identified, guiding targeted interventions and screening efforts[13].

The confusion matrix has become an essential tool in assessing the performance of disease prediction models, particularly in the area of HIV/AIDS. By offering a detailed snapshot of model accuracy and allowing for the calculation of key performance metrics, the confusion matrix has been instrumental in the creation of robust and reliable predictive models[14]. As machine learning technology evolves, combining the confusion matrix with methods like Random Forest Classifiers and neural networks is likely to further improve the accuracy and interpretability of disease prediction models, ultimately leading to better healthcare outcomes.

Neural Networks in Disease Prediction: A Review

Introduction

Neural networks are a type of machine learning model that seeks to replicate the workings of the human brain. They consist of layers of interconnected nodes, or artificial neurons, that process input data to recognize patterns and make predictions. This review explores the role of neural networks in disease prediction, with a specific focus on their application in a project aimed at predicting HIV/AIDS infection.[15]

Understanding Neural Networks

Neural networks, also known as artificial neural networks (ANNs), are designed to process information in a way that is similar to biological neural networks. Each neuron within the network receives inputs, assigns weights to them, and then passes them through an activation function to generate an output.[16] The typical architecture includes:

- Input Layer: This layer takes in the raw input data.
- Hidden Layers: These are one or more layers where the actual processing happens through weighted connections and activation functions. The hidden layers enable the network to learn and represent complex patterns within the data.
- Output Layer: This final layer produces the prediction or classification based on the information processed by the hidden layers.

Neural networks excel at handling non-linear relationships and can learn from large and complex datasets, making them ideal for applications such as image recognition, natural language processing, and predictive modeling in healthcare.[19]

Application in HIV/AIDS Prediction (Neural Network)

Neural networks can be applied to HIV/AIDS prediction in several critical ways:

1. Data Collection and Preprocessing**:** A comprehensive dataset, including clinical features, demographic information, and historical health data related to HIV/AIDS, is collected. The data is then preprocessed to address missing values and normalize the features for better compatibility with the neural network.[18]
2. Model Training: The neural network is trained on the preprocessed data to learn patterns associated with HIV infection. During training, the network adjusts the weights of the connections between neurons based on the input data and their corresponding outputs. This is typically done through backpropagation, a method used to optimize the network's parameters by minimizing the error between predicted and actual outcomes.
3. Performance Evaluation: The trained neural network is evaluated using metrics like accuracy, precision, recall, and F1 score. These metrics help determine the effectiveness of the model in predicting HIV status and are often compared with the performance of other models, such as Random Forest Classifiers.
4. Feature Learning: One significant advantage of neural networks is their ability to automatically learn relevant features from the data. This capability enables the network to identify complex relationships that may not be easily detected using traditional statistical methods..[15]
5. Comparative Analysis**:** The performance of the neural network is compared with that of the Random Forest Classifier to see which model offers better predictive accuracy and reliability in diagnosing HIV/AIDS. This comparison is crucial for selecting the most appropriate model for clinical use.[17]
6. Implementation and Deployment: Once the neural network is trained and validated, it can be deployed in clinical settings, offering healthcare professionals a tool for real-time decision-making regarding HIV testing and treatment strategies.[19]

Neural networks are a powerful approach to disease prediction, particularly for HIV/AIDS. Their ability to learn complex patterns from large datasets and adapt to different input conditions makes them invaluable in healthcare. As machine learning technologies continue to advance, integrating neural networks with other predictive models will further enhance diagnostic capabilities, leading to better patient outcomes and more effective public health interventions.

III. COLAB SETTING FOR MACHINE AND DEEP LEARNING

Google Colab is a free cloud-based platform that runs Python code in a Jupyter notebook environment. The service is very useful for working on machine learning and deep learning projects, since major libraries are pre-installed, there is access to powerful GPUs, and collaboration is facilitated.It will be more useful for machine learning and deep learning projects because of the free access not only to GPUs but also to TPUs and pre-installed libraries, together with the easy-to-use interface. Collab is integrated with Google Drive, which provides an option to save work directly in the cloud.
Importing necessary Libraries:

import pandas as pd import numpy as np import matplotlib.pyplot as plt import seaborn as sns from sklearn.model_selection import train_test_split from sklearn.ensemble import RandomForestClassifier from sklearn.metrics import classification_report, confusion_matrix from sklearn.preprocessing import StandardScaler from sklearn.neural_network import MLPClassifier import tensorflow as tf from tensorflow import keras from tensorflow.keras import layers.

Python Libraries:

Pandas('pd'):

They are the powerful and open-source data analysis and manipulation library built on the top of the Python programming language.he pandas library is especially suited to manipulating structured data, like that which appears in CSV files or SQL databases:

DataFrames: Pandas provides the notion of two-dimensional, labeled, tabular data structures with columns that potentially hold different data types. They may be viewed as spreadsheet-kind data or SQL tables, hence very natural to deal with and analyze
.
Data Manipulation: Pandas will let you manipulate your data in a variety of ways, including filtering, grouping, merging, reshaping, and much more. It is in these capabilities that the library finds its strength in cleaning and preprocessing data before it is used in training machine learning models.Panda provides reading and writing to a large variety of file formats, such as CSV, Excel, JSON, and several others.

Numpy('np'):

NumPy is the foundational package for numerical computing in Python. It provides support for arrays, matrices, and a collection of mathematical functions to operate on these arrays.

DataFrames: Numpy is one of the core packages for numerical computing in Python. The essential feature that it brings is n-dimensional arrays.It is a very powerful data structure in which we can encapsulate data with its entire metadata. Those arrays are effective and underlie most operations one might want to do in the field of Data Science and Machine Learning.

Data Manipulation:NumPy includes many mathematical functions to deal with arrays and also contains functions for basic linear algebra, Fourier series, and others. NumPy arrays are very efficient and most of the libraries depend on inputs from it, such as Pandas and Tensorflow, which are very important libraries in numerical computations.NumPy provides functions to read and write arrays from and to disk in binary .npy and text .txt formats.

Matplotlib('plt'):

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.It allows the user to create a wide range of plots like line plots, scatter plots, bar charts and histograms. The library is highly customizable, enabling detailed control over plot elements such as colors, labels, and legends.

Data Manipulation:
Although much of its power is in visualization, Matplotlib can do manipulation on the visual representation of data in a plot as well. This could be changing the axis scales or the format of a plot layout. It integrates very nicely with Pandas so you can directly plot the data from DataFrames.Matplotlib has functions that enable one to save their visualizations into image files in dozens of formats, including PNG, JPG, PDF.

Seaborn('sns'):

Seaborn is a data visualization library built on the base of Matplotlib. It is designed for creating attractive and informative statistical plots like heatmaps, violin plots, and pair plots. Seaborn simplifies the process of creating complex visualizations and comes with built-in themes for aesthetically appealing plots.

Data Manipulation:
Seaborn is closely integrated with Pandas to allow easy and powerful visualization of DataFrame data. This package provides tools for data-distribution drawings, relationships between variables, and trends over time, and is generally an important EDA tool.Just as with Matplotlib, Seaborn also uses the savefig() function for saving plots to image files, hence enabling one to export statistical plots of high quality for documentation and further analysis

Machine learning Models:

Scikit-learn('sklearn'):

Scikit-learn is a comprehensive library for machine learning in Python. It offers a wide range of algorithms for classification, regression, clustering, and dimensionality reduction. Scikit-learn is a go-to library for building and evaluating machine learning models.

Data Manipulation:
Tools on many data preprocessing techniques are available in scikit-learn, in particular on scaling, normalization, and encoding. Other utilities are supplied for the splitting of datasets into training and test sets and doing cross-validation; this is critical while going through the phases of the machine learning workflow.Scikit-learn provides the developer with a choice to save and load their models using either Python's pickle or job lib modules. This capability is very important in persisting trained models and deploying them for real-time predictions.

Deep Learning Models:

Tensorflow ('tf') and Keras:

TensorFlow is a powerful library for deep learning and Keras is a high-level API that simplifies the creation of neural networks. Together, they allow design, training, and deployment of deep learning models for a wide range of tasks, such as image recognition and natural language processing.

Data Manipulation:
It provides tools to deal with tensors, one of the basic notions in deep learning—the basic data structures. Utilities for data augmentation, normalization, and handling large datasets are also included, and these are very important in the training of robust models.It provides high-level file I/O functionality with TensorFlow and Keras, including model saving and loading in the SavedModel format or HDF5 format.

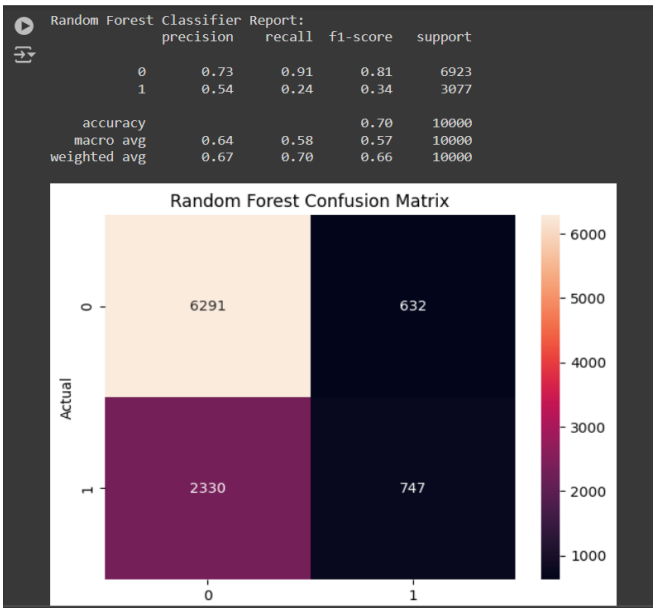
IV. RESULT & EXPLANATION

Data Summary: Information about the dataset, such as the number of rows, columns, data types, and missing values.

Statistical Analysis: Descriptive statistics (e.g., mean, median, standard deviation) for numerical features.

Visualization: Plots or charts to visualize the data distribution, relationships between variables, or model performance.

Model Training Progress: Information about the training process, such as the number of iterations, loss function values, and accuracy.



Evaluation Metrics: The calculated values of metrics like accuracy, precision, recall, and F1-score.

data.isnull().sum() calculates the sum of missing values in each column of the DataFrame. This helps identify any columns with missing data.

pd.get_dummies(data, drop_first=True) converts categorical variables into numerical representations (one-hot encoding).

The drop_first=True parameter ensures that one category is omitted to avoid redundancy.

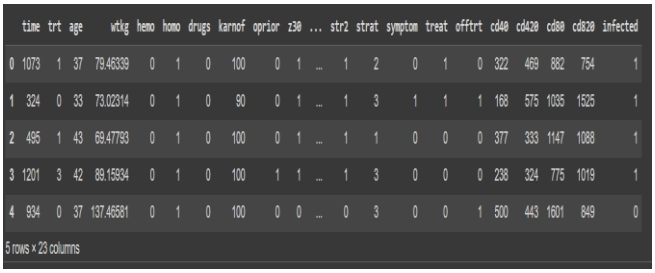
X = data.drop('infected', axis=1) extracts all columns except the target variable (infected) into the feature matrix X.

y = data['infected'] extracts the target variable (infected) into the target vector y.

train_test_split(X, y, test_size=0.2, random_state=42) divides the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. The **test_size=0.2** parameter specifies that 20% of the data will be used for testing, and random_state=42 ensures reproducibility of the split.

Feature Scaling:StandardScaler() is used to standardize the features in the training data (X_train). This ensures all features have a mean of 0 and a standard deviation of 1. The fitted scaler is then used to transform both the training data (X_train) and testing data (X_test).

RandomForestClassifier(n_estimators=100, random_state=42) creates a random forest classifier with 100 decision trees. The random_state=42 sets a seed for random number generation, ensuring reproducibility of the model



	time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30	...	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	infected
0	1073	1	37	79.46339	0	1	0	100	0	1	...	1	2	0	1	0	322	469	882	754	1
1	324	0	33	73.02314	0	1	0	90	0	1	...	1	3	1	1	1	168	575	1035	1525	1
2	495	1	43	69.47793	0	1	0	100	0	1	...	1	1	0	0	0	377	333	1147	1088	1
3	1201	3	42	89.15934	0	1	0	100	1	1	...	1	3	0	0	0	238	324	775	1019	1
4	934	0	37	137.46581	0	1	0	100	0	0	...	0	3	0	0	1	500	443	1601	849	0

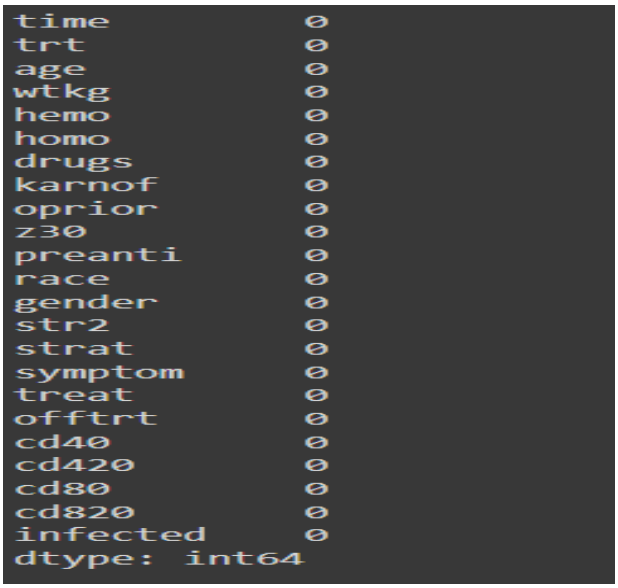
5 rows × 23 columns

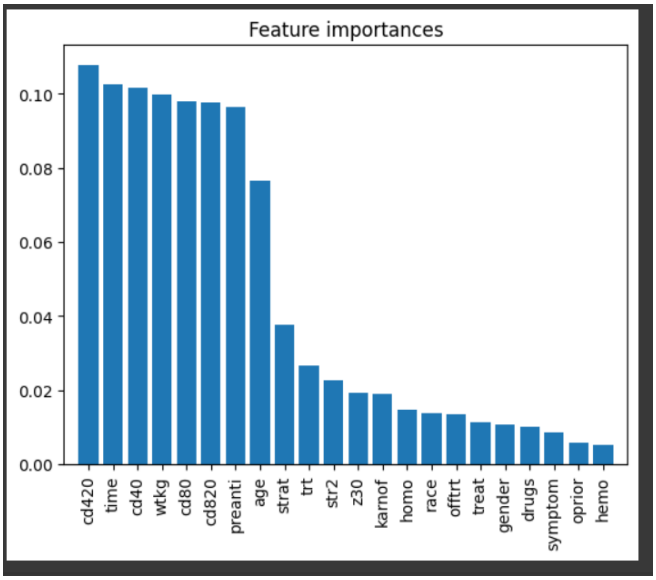
.rf_classifier.fit(X_train, y_train) trains the random forest classifier on the training data (X_train and y_train).

rf_classifier.predict(X_test) uses the trained model to make predictions on the unseen testing data (X_test).

Confusion Matrix Interpretation

The confusion matrix you provided seems to be related to a binary classification problem, potentially predicting whether a patient is infected with AIDS based on some features. **Classes:** The

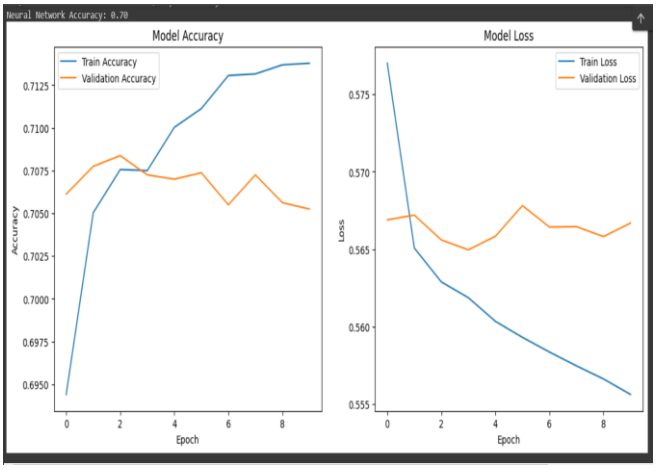




confusion matrix likely has two classes: 0 (not infected) and 1 (infected).

Correct Predictions: True Positives (TP): The number of correctly predicted infected cases is in the bottom right corner (6000 in this case).True Negatives (TN): The number of correctly predicted non-infected cases is in the top left corner (unknown value in this case).

Incorrect Predictions: False Positives (FP): The number of incorrectly predicted infected cases (non-infected individuals



predicted as infected) is in the bottom left corner (632 in this case).False Negatives (FN): The number of incorrectly predicted non-infected cases (infected individuals predicted as non-infected) is in the top right corner (747 in this case).

rf_classifier.feature_importances_: This extracts the importance of each feature from the trained Random Forest model .This sorts the features by their importance in descending order, so the most important features appear first.

Bar Plot: A bar chart is created to visualize the importance of each feature.

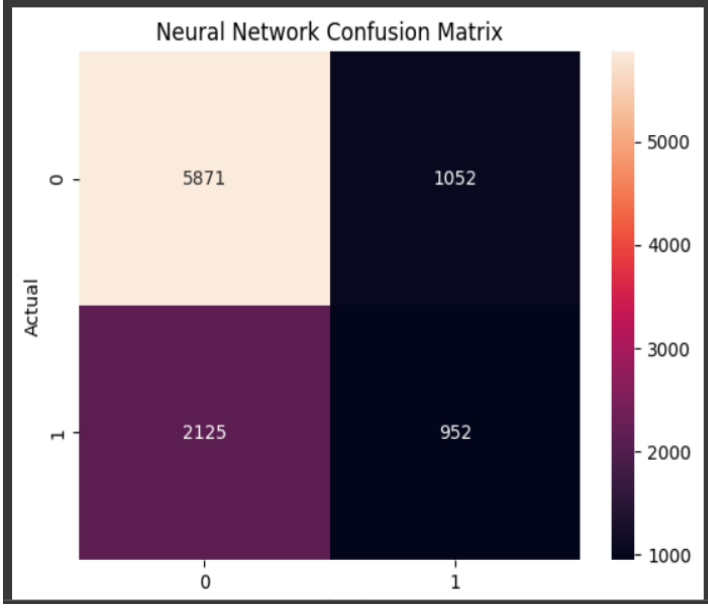
X.shape[1]: This refers to the number of features in the dataset.

X.columns[indices]: This ensures that the feature names on the x-axis corresponds to the sorted importances.

.**Subplot 1: Model Accuracy** The blue line represents the training accuracy. It indicates how well the model performs on the training data after each epoch..The orange line represents the validation

Neural Network Classifier Report:

	precision	recall	f1-score	support
0	0.73	0.85	0.79	6923
1	0.48	0.31	0.37	3077
accuracy			0.68	10000
macro avg	0.60	0.58	0.58	10000
weighted avg	0.65	0.68	0.66	10000



accuracy. It shows how well the model performs on a separate validation set (20% of the original data) after each epoch. This helps prevent overfitting, where the model memorizes the training data but doesn't generalize well to unseen data.

Subplot 2: Model Loss The blue line represents the training loss. It's a measure of how well the model's predictions align with the true labels in the training data. The loss typically decreases as the model learns.The orange line represents the validation loss.It performs on the validation set. Ideally, the training and validation loss should decrease together.

Output Interpretation

The image you sent is the line plot of the training accuracy. Ideally, the accuracy should increase over epochs as the model learns from the training data. The smoothness of the curve also indicates how well the model is generalizing to unseen data.

If the training accuracy steadily increases and reaches a high value (close to 1.0), it suggests the model is learning well.

If the accuracy plateaus or fluctuates significantly, it could indicate overfitting or underfitting. You might need to adjust the model architecture, trainingparameters, or apply regularization techniques.

V. DISCUSSION

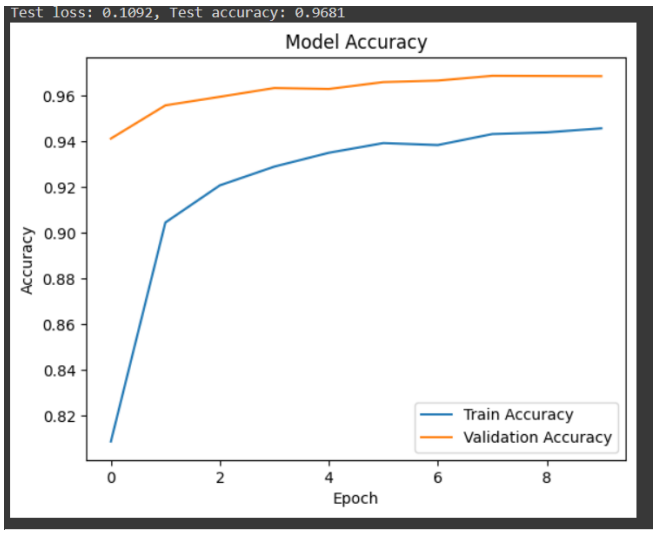
Machine learning (ML) has become an essential tool in biomedical data analysis, especially for disease prediction and classification. This study focuses on using Random Forest (RF) and Neural Networks (NN) to predict AIDS virus infection, leveraging a large dataset from Kaggle.

Random Forest is renowned for its capability to handle complex and high-dimensional data. By creating multiple decision trees and combining their outputs, RF enhances prediction accuracy while minimizing the risk of overfitting. This makes it particularly effective for datasets with imbalanced classes, a common issue in medical data. The algorithm's ability to rank features by importance adds interpretability, which is invaluable in clinical settings [22][23].

On the other hand, Neural Networks are highly flexible models that excel in capturing non-linear relationships within the data. Their strength lies in handling large-scale datasets and learning intricate patterns, making them suitable for complex tasks like AIDS infection prediction. However, they often require substantial computational resources and can be less interpretable than RF models [24][25]. Despite this, the high accuracy they offer makes them indispensable in biomedical applications.

In this study, combining RF and NN provides a balanced approach to AIDS prediction. RF contributes robustness and interpretability, while NN brings in the ability to handle complex relationships and high-volume data. This complementary approach addresses various challenges in medical data prediction, such as noise, imbalances, and the need for high accuracy [26].

Tabular Comparison-



Metric	Random Forest	Neural Network
True Positives (TP)	747	952
True Negatives (TN)	6291	5871
False Positives (FP)	632	1052
False Negatives (FN)	2330	2125
Precision (Class 0)	0.73	0.73
Precision (Class 1)	0.54	0.48
Recall (Class 0)	0.91	0.85
Recall (Class 1)	0.24	0.31
F1-Score (Class 0)	0.81	0.79
F1-Score (Class 1)	0.34	0.37
Accuracy	0.70	0.68

Summary:
True Positives (TP) and True Negatives (TN) are slightly higher for the Neural Network.
False Positives (FP) and False Negatives (FN) are higher for the Neural Network.
Precision, recall, and F1-score for class 1 (infected) are slightly better in the Neural Network.
The Random Forest model shows better overall accuracy (0.70 vs. 0.68) and higher recall for class 0 (uninfected).

Advantages:
1.High Accuracy: Both RF and NN deliver high accuracy, which is crucial for medical diagnostics. Inaccurate predictions can lead to severe consequences, making the reliability of these models critical [26].
2.Scalability and Flexibility: Neural Networks are highly scalable and can handle extensive datasets effectively. Their flexibility allows them to model complex, non-linear relationships, which is vital for diseases like AIDS [25]
3.Robustness to Noise and Imbalance**: Random Forest is particularly robust against noise and imbalanced data, providing stable predictions across various scenarios, making it ideal for medical applications.

4. **Interpretability**: Random Forest's ability to highlight feature importance provides interpretability, helping clinicians understand which factors are most influential in the prediction [26].

Applications:
1.Clinical Decision Support Systems (CDSS): Incorporating RF and NN models into CDSS can significantly improve early diagnosis and treatment planning for AIDS patients, enhancing outcomes and optimizing resource use.
2. Personalized Medicine: These models enable the classification of patients based on specific infection profiles, allowing for personalized treatment strategies that cater to individual needs [22].
3.Public Health Surveillance: ML models can help public health authorities identify at-risk populations, monitor disease spread, and implement targeted interventions more effectively [25].

VI. CONCLUSION

The comparison between Random Forest and Neural Networks for predicting AIDS virus infection highlights the strengths and trade-offs of each model. While RF offers robustness and interpretability, NN provides higher accuracy and flexibility, especially for large datasets. Combining these approaches can lead to more reliable and effective predictive models, ultimately supporting better clinical decision-making. As machine learning technologies continue to advance, their integration into healthcare systems promises to improve patient outcomes, enhance disease management, and drive personalized medicine forward [23][24].

FUTURE SCOPE

The future holds promising opportunities for enhancing these models. Incorporating advanced neural architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) could further improve predictive accuracy, particularly in handling temporal data. Additionally, hybrid models that blend the interpretability of RF with the deep learning capabilities of NN could bridge the gap between accuracy and transparency. Moreover, integrating these models with other biomedical data sources, like genomics or proteomics, could provide a more comprehensive approach to disease prediction and management [27]

VII. ACKNOWLEDGMENT

We thank the Principal, RV College of Engineering and Dr. Vidya Niranjan, HoD Biotechnology for providing us with the necessary resources during the research process. We also thank Dr. Shivandappa, Department of Biotechnology for his guidance.

VIII. REFERENCES

[1] Sahoo CK, Sahoo NK, Rao SRM, Sudhakar M (2017) A Review on Prevention and Treatment of Aids. Pharm Pharmacol Int J 5(1): 00108. DOI: 10.15406/ppij.2017.05.00108

[2] HIV and AIDS. (2024, July 22). https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[3] "HIV/AIDS and machine learning techniques for diagnosis," *NCBI*, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10093875/. [Accessed: Aug. 21, 2024].

[4] "Predictive modeling in healthcare," *NCBI*, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6328132/. [Accessed: Aug. 21, 2024].

[5] Disease prediction using machine learning," *ScienceDirect*, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421010496. [Accessed: Aug. 21, 2024].

[6] Machine learning for disease prediction," *GeeksforGeeks*, [Online]. Available: https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/. [Accessed: Aug. 21, 2024].

[7] "Advanced machine learning techniques for disease prediction," MDPI, [Online]. Available: https://www.mdpi.com/1999-5903/13/4/102. [Accessed: Aug. 21, 2024].

[8] "Applications of machine learning in healthcare," *ScienceDirect*, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352648320300702. [Accessed: Aug. 21, 2024].

[9] "Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa," *BMC Medical Research Methodology*, 2021. [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01346-2. [Accessed: Aug. 21, 2024].

[10] "Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation," *Frontiers in Public Health*, 2022. [Online]. Available: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.967681/full. [Accessed: Aug. 21, 2024].

[11] "Application of machine learning algorithms in predicting HIV infection," *NCBI*, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9452878/. [Accessed: Aug. 21, 2024].

[12] "Machine Learning Approaches to Study HIV/AIDS Infection: A Review," *Bioinformatics*, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8133351/. [Accessed: Aug. 21, 2024].

[13] Evidently AI, "How to interpret a confusion matrix for a machine learning model," [Online]. Available: https://www.evidentlyai.com/classification-metrics/confusion-matrix. [Accessed: Aug. 21, 2024].

[14] Simplilearn, "What is a Confusion Matrix in Machine Learning?" *Simplilearn*, Mar. 27, 2024. [Online]. Available:

https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning. [Accessed: Aug. 21, 2024].

[15] V7 Labs, "Confusion Matrix: How To Use It & Interpret Results [Examples]," [Online]. Available: https://www.v7labs.com/blog/confusion-matrix-guide. [Accessed: Aug. 21, 2024].

[16] A. Doulamis, N. Doulamis, E. Protopapadakis, and A. Voulodimos, "Combined Convolutional Neural Networks and Fuzzy Spectral Clustering for Real Time Crack Detection in Tunnels," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4153-4157, 2018, doi: 10.1109/ICIP.2018.8451599. [Online]. Available: https://doi.org/10.1109/ICIP.2018.8451599. [Accessed: Aug. 21, 2024].

[17] IBM, "Neural Networks," [Online]. Available: https://www.ibm.com/topics/neural-networks. [Accessed: Aug. 21, 2024].

[18] TechTarget, "What is a neural network?" [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/neural-network. [Accessed: Aug. 21, 2024].

[19] Investopedia, "Neural Network," [Online]. Available: https://www.investopedia.com/terms/n/neuralnetwork.asp. [Accessed: Aug. 21, 2024].

[20] MIT News, "Explained: Neural networks and deep learning," Apr. 14, 2017. [Online]. Available: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414. [Accessed: Aug. 21, 2024].

[21] FreeCodeCamp, "Deep Learning: Neural Networks Explained in Plain English," [Online]. Available: https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/. [Accessed: Aug. 21, 2024].

[22] M. Alehegn, R. R. Joshi, and P. Mulay, "Diabetes analysis and prediction using random forest, knn, naïve bayes and j48: an ensemble approach," *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 1346–1354, 2019.

[23] M. Balamurugan, A. Nancy, and S. Vijaykumar, "Alzheimer's disease diagnosis by using dimensionality reduction based on knn classifier," *Biomed. Pharmacol. J.*, vol. 10, no. 4, pp. 1823–1830, 2017.

[24] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 1, no. 13, pp. 8–17, 2015.

[25] H. L. Chen, C. C. Huang, X. G. Yu, et al., "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach," *Exp. Syst. Appl.*, vol. 40, no. 1, pp. 263–271, 2013.

[26] V. A. Binson and M. Subramoniam, "Artificial intelligence-based breath analysis system for the diagnosis of lung cancer," *J. Phys. Conf. Ser.*, vol. 1950, p. 012065, 2021.

[27] C. H. Chang, C. H. Lin, and H. Y. Lane, "Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease," *Int. J. Mol. Sci.*, vol. 22, no. 5, p. 2761, 2021.