Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Hybrid Deep Learning Models for AIDS Prediction

Hari Krishnan Andi[1]

[1]Centre for Postgraduate Studies, Asia Metropolitan University, Malaysia.

E-mail: [1]hari.andi@amu.edu.my

Abstract

Acquired immunodeficiency syndrome (AIDS) consistently ranks as a leading cause of mortality. Effective prevention methodologies include early detection techniques. Controlling infectious diseases is important due to their potential to cause epidemics or pandemics, emphasizing the importance of early diagnosis. This necessity has prompted researchers to develop models aimed at improving disease diagnosis. Traditional clinical prediction models rely on patient-specific characteristics. For infectious illnesses, sources other than the patient, such as previous patient characteristics and seasonal variables, may increase prediction performance. This study predicts infectious diseases by optimizing the settings of deep learning algorithms while taking into account big data, which includes social media data. The collected findings indicate the proposed LSTM model achieves the highest accuracy rate of 92%.

Keywords: AIDS, Infectious Diseases, Machine Learning, Deep Learning, Performance Metrices.

1. Introduction

HIV/AIDS outbreaks necessitate prompt study and development of appropriate intervention programmes and approaches. Behavioural and socio-demographic features are major contributors to the spread of AIDS and necessitate research on the nature and impact of the AIDS pandemic in a given community.

Infectious illnesses provide substantial challenges to worldwide health, with AIDS populations being especially vulnerable due to their growing immune systems and distinct physiological features. To effectively treat infectious infections, suitable drugs must be

DOI: https://doi.org/10.36548/jscp.2024.2.008
Received: 13.05.2024, received in revised form: 02.06.2024, accepted: 18.06.2024, published: 29.06.2024

carefully selected and administered, taking into account aspects such as resistance to pathogens, demographics of patients, and clinical presentation. However, determining the best appropriate medications for paediatric infectious disorders may be difficult and time-consuming, often dependent on empirical research, clinical recommendations, and individual patient reactions[1,11].

In recent years, developments in computational approaches, notably machine learning as well as deep learning, have opened up new possibilities for enhancing medication prediction in infectious illnesses. Researchers and healthcare workers may use large-scale datasets including varied patient groups, pathogens, and treatment results to construct prediction models that are capable of identifying appropriate drugs for paediatric infectious disease. These models have the potential to improve therapeutic efficacy, reduce negative side effects, and optimize the use of resources in clinical settings. Treatment of infectious diseases, especially in pediatric populations, continues to be a significant issue in healthcare due to variables such as pathogen diversity, changing resistance patterns, and unique patient responses. Identifying effective pharmacological therapy customized to individual infectious pathogens and patient characteristics is essential for improving treatment results while minimizing side effects [12,13]. In the past few years, deep learning models have come to be as promising medication prediction tools, with the potential to improve the precision and efficiency of therapy choices in infectious diseases.

This research explores the use of machine learning and deep learning algorithms for medication prediction in AIDS [15,16]. The limitations of traditional medication selection approaches are explored and the opportunity provided by computational modelling tools are emphasised in the proposed study. This research seeks to study the use of deep learning models in medication prediction for infectious disease, with a particular emphasis on AIDS populations [18]. The core concepts of deep learning architectures, their application in analyzing various forms of data related to infectious disease, and their potential impact on clinical decisions will be examined. Furthermore limitations and potential associated with using deep learning in medication prediction tasks by a thorough analysis of existing research and approaches are also highlighted

2. Related Work

The Table.1 presents the short overview of the existing hybrid deep learning networks.

Table 1. Related Works of the Existing Hybrid Deep Learning Networks

Year	Inference	Limitations
2022[3]	Analysing data from the BARDA Paediatric COVID-19 Data Challenge, the researchers created a deep learning algorithm to predict hospitalisation and severe complication risks for paediatric COVID-19 patients that outperformed previous machine learning methods.	Predicting hospitalization risk for children infected with COVID-19, as well as serious consequences for hospitalized paediatric COVID-19 patients
2021[4]	A deep learning model can predict the beginning of CLABSI in hospitalised children with central venous lines 48 hours before specimen collection.	Data from a particular pediatric health system may have limited applicability to other situations. The study's retrospective structure may limit its ability to capture real-time prediction performance.
2022[5]	The research examined current machine-learning algorithms for infectious illness detection, emphasising the significance of choosing the right ML approach depending on the dataset and intended results. Furthermore, combining deep learning with NLP can increase the performance of ML diagnostic models.	The study's shortcomings, as noted in the publication, include limited datasets, the necessity to combine methodologies to extract additional features, the absence of real-time performance evaluation, the lack of prospective validations, and the heterogeneity of the evaluated data.
2022[6]	The study contains enhanced MIDDM prediction results for diagnosing infectious diseases, the effect of	Due to a small sample size and lack to account for national alterations to diagnostic criteria, the model's

	sample data on prediction accuracy, and MIDDM's superiority over other models in multi-classification of infectious disorders.	input features had unstable accuracy.

[7]A geometric deep learning (GDL) technique is suggested for predicting HIV medication resistance and viral-drug interactions. The data set within the SMILES representation was first translated to a molecular representation, followed by a graph representation that the GDL model could comprehend. The Message Passing Neural Network (MPNN) moves the node feature vectors to a new area where the training process may take place. The GDL technique outperforms predicting treatment resistance in HIV by 93.3%.

[8] In this study, every acceptable combination of convolutional neural network (CNN), transformer (TF) models and long short-term memory (LSTM), is tested. These hybrid models were also compared to single CNN, LSTM, and TF models using various types of optimizers.

[14] This article presents a way to predict cardiovascular disorders utilizing machine learning, neuro-fuzzy, and statistical techniques. It demonstrates that the new methodology outperformed established methods, achieving a high prediction accuracy of more than 90 percent.

[17] This research proposes a machine learning model for predicting seasonal diseases like dengue, malaria, pneumonia, and typhoid using real-time data from Madurai district. The model, incorporating Antlion Optimization for feature selection and Random Forest with XG-Boost for classification, demonstrates superior efficiency compared to other methods, achieving high precision and recall rates.

[19] This research focuses on creating a hybrid dataset named "Sathvi" from commonly used heart disease prediction datasets to improve CVD risk prediction models. Utilizing machine learning classifiers like Naive Bayes, XGBoost, k-NN, MLP, SVM, and CatBoost, the study achieves accuracies ranging from 88.67% to 98.11%, demonstrating robust predictive performance validated through 10-fold cross-validation with a mean accuracy of 94.34%

The results in [20] demonstrate that an interpretable model may identify patterns through learning filters that are thought to be the most crucial characteristics for predicting virus sequences, and it can also function as a recommendation system for additional analysis of the raw sequences classified as "unknown" by alignment-based techniques.

3. Proposed Methodology

The development of an LSTM (Long Short-Term Memory) model for analysing a dataset such as the AIDS Clinical Trials Group Study 175 (ACTG 175) entails multiple processes, spanning from data preparation to model validation. ACTG 175 is a clinical study that examined the effectiveness of various antiretroviral medication combinations in HIV-positive individuals. The information generally contains variables such as baseline CD4 count, virus load, treatment protocol, and patient demographics, with an emphasis on outcomes such as CD4 count changes, suppression of viruses, or clinical progression. The Figure 1 depicts the block diagram of the proposed.

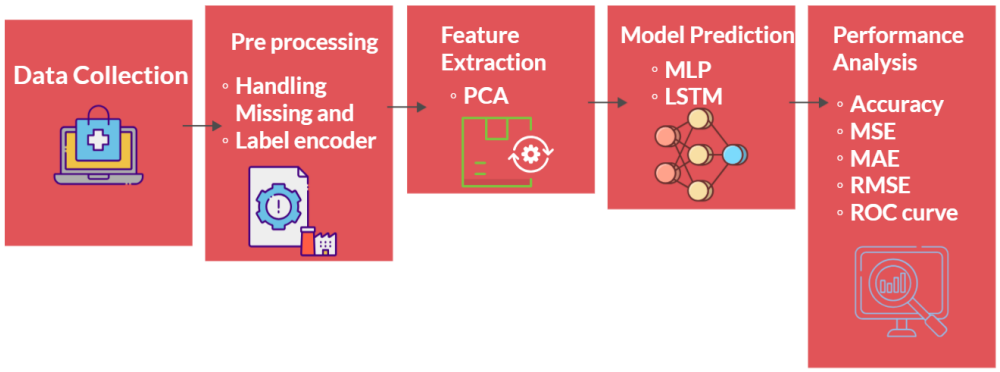

Figure 1. Block Diagram of the Proposed Methodology

3.1 Data Collection

The AIDS Clinical Trials Group Study 175 Dataset (Figure 2), is a complete collection of medical statistics and categorical information about AIDS patients. This dataset was produced primarily to compare the efficacy of two types of AIDS treatments: zidovudine (AZT) vs didanosine (ddI), AZT plus ddI, and AZT plus zalcitabine (ddC). The prediction objective for this dataset is to determine if every individual died within a specific time frame.It consists of 2139 rows and 24 columns.

	time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30	zprior	preanti	race	gender	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	label
0	948	2	48	89.8128	0	0	0	100	0	0	1	0	0	0	0	1	0	1	0	422	477	566	324	0
1	1002	3	61	49.4424	0	0	0	90	0	1	1	895	0	0	1	3	0	1	0	162	218	392	564	1
2	961	3	45	88.452	0	1	1	90	0	1	1	707	0	1	1	3	0	1	1	326	274	2063	1893	0
3	1166	3	47	85.2768	0	1	0	100	0	1	1	1399	0	1	1	3	0	1	0	287	394	1590	966	0
4	1090	0	43	66.6792	0	1	0	100	0	1	1	1352	0	1	1	3	0	0	0	504	353	870	782	0
5	1181	1	46	88.9056	0	1	1	100	0	1	1	1181	0	1	1	3	0	1	0	235	339	860	1060	0
6	794	0	31	73.0296	0	1	0	100	0	1	1	930	0	1	1	3	0	0	0	244	225	708	699	1
7	957	0	41	66.2256	0	1	1	100	0	1	1	1329	0	1	1	3	0	0	0	401	366	889	720	0
8	198	3	40	82.5552	0	1	0	90	0	1	1	1074	0	1	1	3	1	1	1	214	107	652	131	1
9	188	0	35	78.0192	0	1	0	100	0	1	1	964	0	1	1	3	0	0	1	221	132	221	759	1
10	1073	2	34	95.256	0	0	0	100	0	1	1	897	1	0	1	3	0	1	0	471	468	770	620	1
11	1175	3	38	76.4316	0	1	0	100	0	1	1	461	0	1	1	3	0	1	0	340	230	660	510	0
12	1203	2	25	68.04	0	1	0	90	0	1	1	852	0	1	1	3	0	1	0	540	590	1590	1330	0
13	1065	1	34	62.8236	0	1	0	90	0	1	1	342	0	1	1	2	0	1	1	212	190	1094	1180	0
14	357	2	49	79.38	0	1	0	90	0	1	1	402	0	1	1	3	0	1	1	120	140	800	1090	1

Figure 2. AIDS Clinical Trials Group Study 175 Dataset

From the dataset it is analysed that

- The dataset includes a variety of clinical and demographic characteristics such as age, weight (wtkg), therapy type (trt), and numerous metrics from HIV/AIDS clinical studies.
- The dataset has no missing values, as all columns possess the same count with the total number of records.
- The goal variable appears to be 'label', which suggests a binary classification problem.

3.2 Data Preprocessing

This process involves identifying outliers or abnormalities in the data, normalizing numerical characteristics, and encoding categorical variables as needed.The numerical characteristics in the dataset were effectively standardised. The outlier identification technique found 344 outliers in the sample. The outliers have been removed from the dataset.

The next process is to check for categorical variables:

3.3 Feature Extraction

Principal Component Analysis, is a popular technique used for dimensionality reduction and feature extraction. It's particularly handy when dealing with datasets with many variables, as it helps in simplifying the data while preserving its important characteristics. Before using

PCA, it is critical to standardise the data such that all variables have a mean of zero and a normal deviation. This step is required to give every variable equal weight. PCA computes the covariance matrix of standardised data. This matrix determines the link between each pair of variables in the dataset. PCA separates the covariance matrix into its eigenvectors and eigenvalues. Eigenvectors describe the directions (principal components) of maximal variation in the data, and eigenvalues show the quantity of variance along these paths. PCA conducts the eigenvectors in descending order according to their respective eigenvalues. Finally, PCA identifies the top k eigenvectors (principal components) using the explained variance ratio or a specified threshold. PCA successfully decreases the dataset's dimensionality while maintaining as much variance as feasible.

The dataset was visualised using Principal Component Analysis (PCA), which reduced it to two main components (Figure 3):

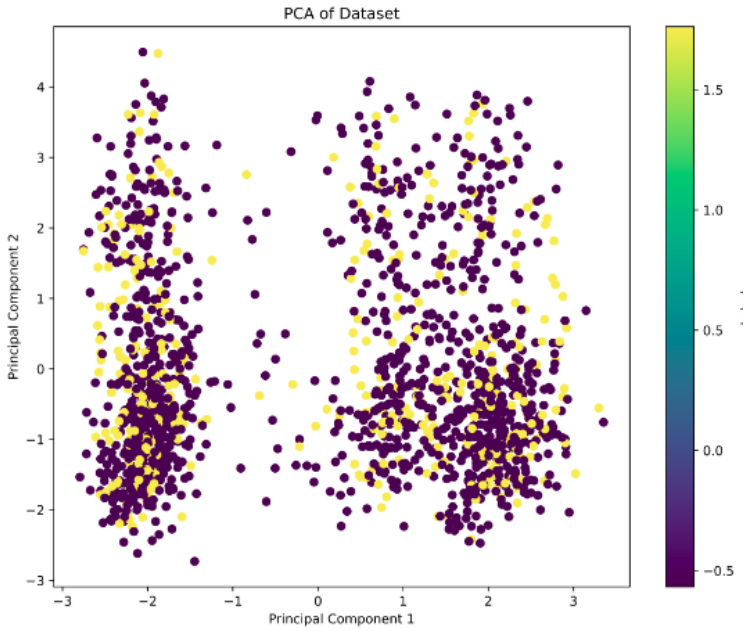

Figure 3. PCA – Feature Scaling Analysis

3.4 Model Training

3.4.1 MultiLayer Perceptron

MLPs are a sort of feedforward neural network made up of several layers of neurons, each coupled to the next. They are effective at processing complicated patterns in data, but they

need careful tweaking of factors such as the number of layers as well as the number of neurons per layer.

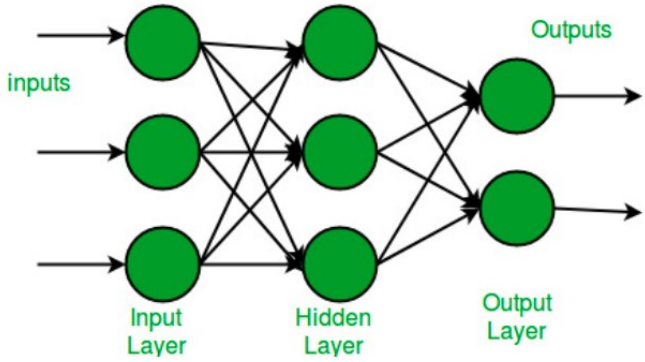

Figure 4. MLP Model Architecture [9]

In the multi-layer perceptron model is shown in the above figure, there are three inputs, resulting in three input nodes, and the hidden layer contains three nodes. The output layer generates two outputs, hence there are two output nodes. The nodes in the input layer accept data and transmit it for further processing. In the Figure 4 above, the nodes in the input layer forward their output to each of the three nodes in the hidden layer, and the hidden layer processes the information before passing it to the output layer. Every node in the multilayer perception employs a sigmoid activation function.

3.4.2 LSTM Model

Long short-term memory (LSTM), shown in Figure 5, is a variant on the RNN model. An RNN can only remember short-term information, but LSTM is able to handle complex time series data. Furthermore, an RNN model suffers from the problem of vanishing gradients for lengthy sequence data; LSTM can avoid this issue during training. An LSTM model can recall past long-term time-series information and has automated control over whether to keep useful features or reject irrelevant characteristics in the cell state. An LSTM model includes three gates to govern features: input, forget, and output. The input gate allows new information to enter the cell state. The forget gate deletes past irrelevant information from the cell state.

The output gate governs the recovered information from the cell state and determines the next hidden state. These gates allow an LSTM model to automatically save or erase stored memory.

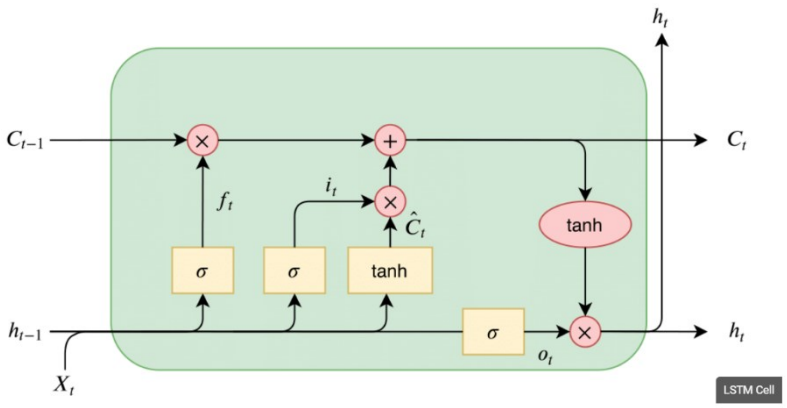

Figure 5. LSTM Model Architecture [10]

4. Evaluation Results

The dataset has been prepared for further analysis or deep learning model training:

Training set size: 1436 samples

Testing set size: 359 samples

Performance Metrics Analysis:

Accuracy: This is the most basic statistic, indicating the proportion of correctly predicted cases out of all predicts made.

AUC Curve: This statistic is utilised with the Receiver Operating Characteristic (ROC) curve. It calculates an overall measure of performance for every classification criteria.

Mean Square Error (MSE): This calculates the average of the squared errors, or the mean square difference between estimated and actual values.

Root Mean Square Error: This is the square root of the average of the squared errors. RMSE is an accurate measure of how well the model predicts the response, and it is the most relevant fit criteria if the model's primary function is prediction.

4.1 Outcomes of MLP Model

ROC curve.

The ROC curve corresponding to the MLP model (Figure 6) after converting continuous scores to binary labels and using the median as a threshold:

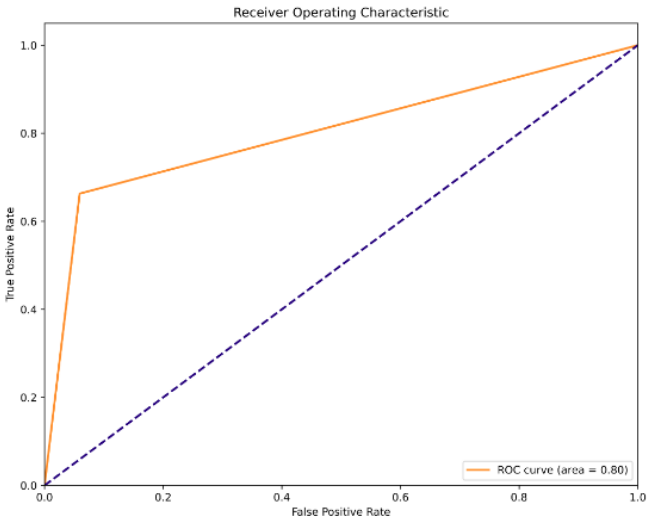

Figure 6. ROC Curve of MLP Model

Mean Squared Error (MSE): 0.6923333336858726

Root Mean Squared Error (RMSE): 0.8320657027458064

Mean Absolute Error (MAE): 0.743881603827356

The model was trained over 50 epochs.

The final training accuracy reached approximately 84.51%.

The final validation accuracy was about 79.94%.

4.2 Outcomes of LSTM Model

ROC curve analysis of the LSTM model (Figure 7):

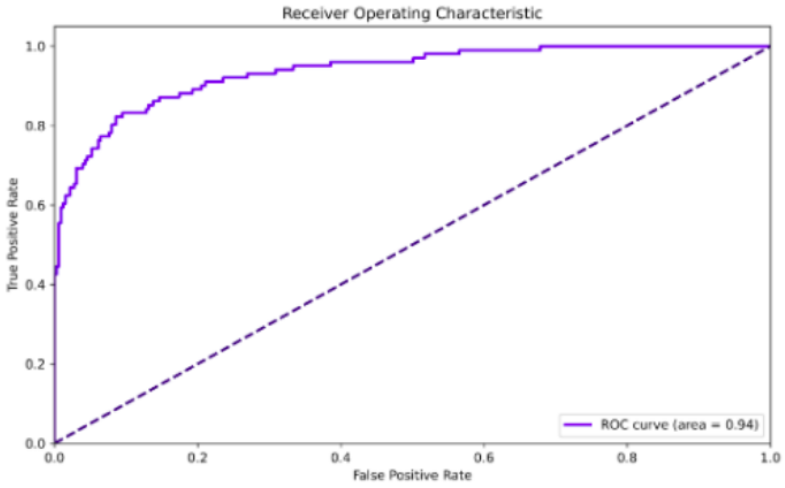

Figure 7. ROC Curve of LSTM Model

Mean Squared Error (MSE): 0.0743

Root Mean Squared Error (RMSE): 0.2726

Mean Absolute Error (MAE): 0.1594

Area Under the ROC Curve (AUC): 0.9367

The model was trained over 50 epochs.

The training and validation accuracy improved over time.

The final training accuracy reached approximately 92.46%.

The final validation accuracy was about 89.49%.

Metrics

5. Conclusion

This research proposes Hybrid MLP and LSTM models for the early detection and classification of AIDS. In the entire modelling, the statistical and categorical information are collected from the available open-source dataset. The collected dataset is pre-processed and normalized with scalar encoding and PCA is used for feature extraction. Since the data are categorical the best-suited model of deep learning algorithms such as MLP and LSTM are used and compared with the performance metrics such as Accuracy, AUC curve, and regression

metrics like MSE, RMSE, and MAE. On the complete analysis of the dataset, LSTM provides a better accuracy rate of 92.46% and an AUC rate of 0.9367. In future work, the model performance can be further improved by hyper-tuning of parameters.

References

[1] Ketu, Shwet, and Pramod Kumar Mishra. "A Hybrid Deep Learning Model for COVID-19 Prediction and Current Status of Clinical Trials Worldwide." Computers, Materials & Continua 66, no. 2 (2021).

[2] Shah, Jaimin, Darsh Vaidya, and Manan Shah. "A comprehensive review on multiple hybrid deep learning approaches for stock prediction." Intelligent Systems with Applications 16 (2022): 200111.

[3] Mahmud, Sajid, Elham Soltanikazemi, Frimpong Boadu, Ashwin Dhakal, and Jianlin Cheng. "Deep Learning Prediction of Severe Health Risks for Pediatric COVID-19 Patients with a Large Feature Set in 2021 BARDA Data Challenge." ArXiv (2022).

[4] Tabaie, Azade, Evan W. Orenstein, Shamim Nemati, Rajit K. Basu, Gari D. Clifford, and Rishikesan Kamaleswaran. "Deep learning model to predict serious infection among children with central venous lines." Frontiers in pediatrics 9 (2021): 726870.

[5] Alqaissi, Eman Yahia, Fahd Saleh Alotaibi, and Muhammad Sher Ramzan. "Modern machine-learning predictive models for diagnosing infectious diseases." Computational and mathematical methods in medicine 2022 (2022).

[6] Wang, Mengying, Zhenhao Wei, Mo Jia, Lianzhong Chen, and Hong Ji. "Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records." BMC medical informatics and decision making 22, no. 1 (2022): 41.

[7] Das, Bihter, Mucahit Kutsal, and Resul Das. "Effective prediction of drug–target interaction on HIV using deep graph neural networks." Chemometrics and Intelligent Laboratory Systems 230 (2022): 104676.

[8] Salman, Diaa, Cem Direkoglu, Mehmet Kusaf, and Murat Fahrioglu. "Hybrid deep learning models for time series forecasting of solar power." Neural Computing and Applications (2024): 1-18

[9] https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/

[10] https://thorirmar.com/post/insight_into_lstm/

[11] Gill, O. Noel, DanielaDe Angelis, C. L. R. Bartlett, N. E. Day, RoyM Anderson, and GordonT Stewart. "AIDS predictions." The Lancet 341, no. 8855 (1993): 1286-1288.

[12] Kareem, Sameem Abdul, S. Raviraja, Namir A. Awadh, Adeeba Kamaruzaman, and Annapurni Kajindran. "Classification and regression tree in prediction of survival of aids patients." Malaysian Journal of Computer Science 23, no. 3 (2010): 153-165.

[13] Li, Zeming, and Yanning Li. "A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS." BMC medical informatics and decision making 20 (2020): 1-13.

[14] Taylan, Osman, Abdulaziz S. Alkabaa, Hanan S. Alqabbaa, Esra Pamukçu, and Víctor Leiva. "Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods." Biology 12, no. 1 (2023): 117.

[15] Zhai, Xuanpei, Wenshuang Li, Fengying Wei, and Xuerong Mao. "Dynamics of an HIV/AIDS transmission model with protection awareness and fluctuations." Chaos, Solitons & Fractals 169 (2023): 113224.

[16] Saha, Ramesh, Lokesh Malviya, Akshay Jadhav, and Ramraj Dangi. "Early stage HIV diagnosis using optimized ensemble learning technique." Biomedical Signal Processing and Control 89 (2024): 105787.

[17] Indhumathi, K., and K. Satheshkumar. "Prediction of seasonal infectious diseases based on hybrid machine learning approach." Multimedia Tools and Applications 83, no. 3 (2024): 7001-7019.

[18] Marcus, Julia L., Whitney C. Sewell, Laura B. Balzer, and Douglas S. Krakower. "Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic." Current HIV/AIDS Reports 17 (2020): 171-179.

[19] Kanagarathinam, Karthick, Durairaj Sankaran, and R. Manikandan. "Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset." Data & Knowledge Engineering 140 (2022): 102042.

[20] Dasari, Chandra Mohan, and Raju Bhukya. "Explainable deep neural networks for novel viral genome prediction." Applied Intelligence 52, no. 3 (2022): 3002-3017.