Multimodal Disease Prediction Using Hybrid Machine Learning Algorithms

J. Grace Arputha Rajakumari[1], N.Balajiraja[2]

[1]Research Scholar J.J college of Arts and Science (Autonomous) Affiliated to by Bharathidasan university, Tiruchirapalli Pudukkottai, India, Email: rajigrace81@gmail.com
[2]Assistant Professor PG and Research Department of Computer Science J.J college of Arts and Science(Autonomous) Affiliated to by Bharathidasan university, Tiruchirapalli Pudukkottai, India, Email: nbalajiraja@gmail.com

Received: 18.04.2024 Revised : 16.05.2024 Accepted: 23.05.2024

ABSTRACT
Machine learning (ML) and deep learning (DL) are derivatives of artificial intelligence (AI) that have proven effective in various fields, including healthcare, and are now routinely integrated into patients' daily activities. Machine learning techniques are increasingly used to improve disease prediction. In this paper, we propose a multi-disease prediction system that uses ML and DL algorithms to predict the probability of several common diseases. Although there are many algorithms and techniques for predicting diseases, there are not enough systems that can identify multiple diseases in a single system. Therefore, this paper focuses on the prediction of various diseases using ML and DL algorithms. Our goal is to build a model that effectively predicts diseases such as kidney, heart, diabetes, and malaria using ML and DL algorithms. This helps to better predict disease. For accurate predictions, we will use stacking and assembling models, which helps improve model accuracy. We will implement all these models in a Flask web application.

Keywords: multimodal machine learning, disease prediction, data heterogeneity, data fusion, learning model.

1. INTRODUCTION

In modern times, modern computing is growing rapidly in the healthcare field. As digital technologies emerge and health data continues to grow, researchers are exploring ways to use machine learning to improve health outcomes. One of the key areas of research is the use of machine learning algorithms to predict a variety of diseases. In this paper, we propose a new method to predict multiple diseases simultaneously using machine learning and deep learning techniques. Our approach uses a variety of patient data, including demographic information, medical history, and laboratory test results, to train machine learning models to predict the occurrence of multiple diseases in a given patient. The ability to predict multiple diseases in patients may provide a way for health care professionals to understand disease. Additionally, our approach helps identify patients with multiple diseases for early intervention and potentially lower healthcare costs.

In this paper, we describe methods for predicting multiple diseases using machine learning and present the results of our experiments on real-world medical datasets. We also discuss the potential impact of our approach on improving healthcare outcomes, as well as future research directions in this area. This article focuses on the prediction of various diseases that have a greater impact on human life. Here are some statistics from the World Health Organization (WHO) on diabetes-related deaths. Diabetes is one of the dangerous diseases globally, killing around 1.6 million people. There were an estimated 229 million cases of malaria worldwide in 2019, with an estimated 409,000 cases of malaria. , 17.9 million people died from malaria and cardiovascular diseases in 2019, accounting for 32% of global deaths.Heart disease is the leading cause of death in the United States, accounting for approximately 655,381 deaths in 2018. Therefore, it is important to identify the type of disease early to prevent serious illness. We used different models to predict different types of diseases. To train the model, we used various classification techniques such as decision trees, support vector machines, k-nearest neighbors, random forests, and convolutional neural networks, which use assembly models such as bagging to achieve higher accuracy.

2. REVIEW OF LITERATURE

Rahim et.al., [2023] proposed a cardiovascular disease diagnosis framework based on machine learning to improve the accuracy of cardiovascular disease prediction, and stated that using various feature selection and classification methods to improve prediction accuracy has attracted research The utmost attention to personnel. . However, handling missing values and class imbalance issues have received less attention and accuracy has improved.

Zhu et al. [2023] proposed a machine learning model that uses speech pauses as effective biomarkers in dementia detection, aiming to reduce the confidence of the detection model by adding perturbations to speech pauses. Talk about test samples.

Compared to other learning algorithms, Feng et al., [2023] used SVM for regression and classification of training datasets. They also developed relationships between input and output data to improve the overall accuracy of healthy learners and predict concrete strength using these given input variables.

Tarawneh et al., [2022] conducted a study using a hybrid data mining classifier approach to predict heart disease. The dataset comes from the UCI Machine Learning Repository and contains 303 records and 76 attributes.

Banu et al. al., [2022] focused on developing AIS-based devices, acquiring knowledge of classifiers for clinical analysis, and studying the functionality of the proposed classifiers. The proposed classifier successfully advances the identification methods of thyroid diseases.

Maleki et al., [2022] proposed a coronavirus recovery and confirmed case prediction model for epidemic control and effective healthcare resource management. Statistical methods have been used for accurate time-indexed data prediction. Autoregressive two-part mixed normal distribution (TP-SMN-AR) models are a family of various symmetric/asymmetric and light-tailed/heavy-tailed models that have been used to predict Covid-19 cases.

3. MATERIALS AND METHODS

3.1 Data collection

3.1.1 Diabetes

The Pima Indians Diabetes Dataset is commonly used for binary classification to predict diabetes (positive/negative). It contains features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, and Age, with the target variable, Outcome (0 or 1). Available through sources like the UCI Repository, this dataset is frequently applied in machine learning models like logistic regression, decision trees, SVMs, and neural networks, making it a popular choice for diabetes prediction projects.

3.1.2 Kidney

The Chronic Kidney Disease (CKD) Dataset is used for binary classification to predict the presence of CKD. Key features include Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Serum Creatinine, Hemoglobin, and Sodium levels, with the target variable being Class (1 = CKD, 0 = not CKD). Machine learning models such as logistic regression, decision trees, and SVMs commonly apply this dataset, which is available from sources like the UCI Repository, for kidney disease prediction.

3.1.3 Heart disease

The heart disease dataset is widely used for classification in machine learning to predict the presence of heart disease. It contains features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, maximum heart rate, fasting blood sugar, exercise-induced angina, and the number of major vessels. The target variable represents the diagnosis of heart disease. Pre-processing typically involves handling missing values and scaling continuous features to improve model performance.

3.1.4 Malaria

The Malaria dataset is used to classify red blood cells as infected or uninfected through image-based analysis. Attributes include pixel values from microscopic images, with features like texture and intensity variations. The target variable identifies whether a cell is infected (1) or not (0). Convolutional neural networks (CNNs) are typically applied for this task. The dataset is available from sources such as NIH and Kaggle, often with preprocessed images for machine learning applications.

3.2 Data Cleaning

Data cleaning is crucial for preparing the datasets for analysis. Each dataset may have inconsistencies such as missing values, outliers, or irrelevant features. The key steps include:

- Handling Missing Values: Missing data can be imputed using techniques such as mean, median, or mode imputation, or by using advanced methods like k-nearest neighbors (KNN) or regression-based imputation.
- Outlier Detection: Outliers are detected and removed or adjusted using statistical methods like Z-scores, interquartile ranges (IQR), or domain-specific rules.
- Normalization/Standardization: Continuous features such as blood pressure or glucose levels are normalized (scaling between 0 and 1) or standardized (mean = 0, standard deviation = 1) to bring all features into a comparable range.
- Removing Irrelevant Features: Attributes that do not contribute to disease prediction, such as IDs or redundant information, are removed.
- Encoding Categorical Features: Convert categorical features (e.g., gender, chest pain type) into numerical form using one-hot encoding or label encoding.

3.3 Feature Selection

Feature selection helps in identifying the most relevant features for disease prediction. It reduces the dimensionality of the data and improves model performance by focusing on the most impactful variables.

- Correlation Analysis: Identify and remove features that are highly correlated with each other, as they may introduce redundancy. The Pearson or Spearman correlation coefficient is used for this purpose.
- Statistical Tests: Use statistical methods like Chi-Square tests, ANOVA (for categorical vs. continuous features), or Mutual Information to assess feature importance.
- Dimensionality Reduction: Techniques like Principal Component Analysis (PCA) or Recursive Feature Elimination (RFE) are applied to reduce the number of features while retaining the most informative ones.

3.4 Weight Factor Analysis

Weight factor analysis involves assigning importance scores to each feature based on its contribution to the prediction of a disease. This step ensures that the most influential features get higher priority in the prediction process.

- Logistic Regression Coefficients: In logistic regression models, the coefficients of each feature indicate its influence on the target variable. Features with higher absolute values of coefficients have more weight.
- Decision Tree Feature Importance: In tree-based models like Random Forests, feature importance can be derived from how frequently a feature is used in decision splits.
- SHAP or LIME Values: SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can be used to explain the model's predictions by attributing the contribution of each feature to the prediction.

3.5 Ranking

After determining the weight factors, features are ranked based on their importance for disease prediction. This ranking helps to identify the most critical factors for each disease.

- Ranking by Feature Importance: Features with the highest weights or importance scores are ranked at the top. For example, glucose levels may be ranked highest for diabetes prediction, while chest pain type could be most important for heart disease.
- Top-N Features: Often, the top-N features are selected for the final model to avoid over fitting and ensure generalization. This reduces model complexity while retaining predictive power.

3.6 Model Selection and Training

Multiple machine learning algorithms are applied to build prediction models for each disease. Common algorithms used for disease prediction include:

- Logistic Regression: A simple, interpretable model useful for binary classification tasks.
- Decision Trees/Random Forests: These models are great for handling non-linear relationships and ranking features by importance.
- Support Vector Machines (SVM): SVMs are effective for small to medium-sized datasets and work well with clear margins of separation.
- Neural Networks: Especially for the Malaria dataset, Neural Networks (NNs) are used for classification.
- Ensemble Methods: Techniques like Gradient Boosting, XGBoost, or stacking can combine the strengths of multiple models for improved performance.

3.7 Evaluation

The performance of each model is evaluated using various metrics:

- Accuracy: The proportion of correctly predicted cases.
- Precision: The number of true positives divided by the number of predicted positives, indicating the model's ability to avoid false positives.
- Recall (Sensitivity): The number of true positives divided by the actual positives, indicating the model's ability to detect the disease.
- F1-Score: The harmonic mean of precision and recall, providing a balanced measure of both.
- AUC-ROC Curve: Measures the trade-off between true positive rate and false positive rate, especially useful for binary classification tasks.

3.8 Machine Learning Techniques

3.8.1 KNN Algorithm

K-Nearest Neighbors is a simple method that maintains all existing examples and uses a similarity measure (such as a distance function) to classify new examples. KNN has been used as a non-parametric method in statistical estimation and pattern recognition since the early 1970s. KNN (K Nearest Neighbors) is a classification technique based on how "similar" data (one vector) is to another vector. It is one of several algorithms (supervised learning) used in data mining and machine learning.

The KNN's steps are as follows:

- Get info that isn't classified;
- Calculate the distance (Euclidian, Manhattan, Murkowski, or Weighted) between the new data and all previously categorized data
- Calculate the distance (Euclidian, Manhattan, Murkowski, or Weighted) between the new data and all previously categorized data.
- Examine the list of classes with the lowest distance between them and count how many of each class appear;
- Takes the class with the most appearances as the proper class

3.8.2 SVM Algorithm

SVM is a powerful classification algorithm. Based on the distribution of training samples, find the best hyper plane that classifies samples into different categories. In many cases, it is difficult to construct such a hyper plane in the original feature space. To map samples to a new space with higher dimensions, where hyper planes are easy to construct, kernel methods are used. We can set its class based on which side of the hyper plane it lies.

3.8.3 Random Forest Algorithm

One of the versatile machine learning algorithms is random forest. This algorithm can perform both classification and regression methods. The random forest algorithm consists of multiple decision trees. It can be said that a set of decision trees jointly produces a forest. Each decision tree is unique to the random forest, reducing the overall variance of the random forest classifier. Random forest algorithm can make better decisions than decision trees. In other words, the decision-making of random forests is more stable and reliable than trees. Especially when each tree is not related to each other. Random forests collect the decisions of individual trees through a voting scheme such as majority voting, i.e. for each observation, each tree decides on a set of votes and selects the group with the highest number of votes. Random forests have some tuning parameters that can be optimized. These include the number of trees, the number of randomly selected predictors at each node, the proportion of observations in each decision tree, and the minimum number of observations at the final node of the decision tree.

3.8.4Decision Tree

DT is a relatively simple classification algorithm used as a predictive model in medical diagnosis and biomarker detection. Different from the previous three algorithms, the classification principle is easier to understand, which the biggest advantage of DT. is By learning from training samples, a tree is constructed. This tree provides a completely open classification procedure for test samples. This provides the opportunity to discover classification principles. Therefore, it is considered a white-box algorithm. In addition to tree structures, DTs can also be represented using a set of rules. Each rule shows the path from the root to an individual leaf. For multiple classes, these rules can suggest multiple patterns.

3.8.5 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the human brain, consisting of interconnected neurons organized in layers. They process input data through forward propagation, adjusting weights using back propagation to minimize prediction errors. Common applications include image recognition, natural language processing, and time series forecasting. Various types of ANNs exist, such as feed forward, convolutional, and recurrent networks, each tailored for specific tasks. Their ability to learn complex patterns has led to significant advancements across multiple fields.

4. RESULTS AND DISCUSSIONS

The machine learning process for predicting multiple diseases, such as kidney disease, heart disease, diabetes, and malaria, using datasets from the UCI Repository involves a structured pipeline. This pipeline ensures that the datasets are properly cleaned, features are selected and weighted, and predictive models are developed and evaluated. Below is a detailed description of each step in this process:
Here's a stepwise procedure based on the provided code for the hybrid algorithm that combines Artificial Neural Networks (ANN) and Random Forest (RF) for classification tasks:

4.1 Proposed Hybrid Algorithm RBHML

Step 1: Data Preprocessing
1. Define the function `preprocess_data(data)`:
- Handle missing values using an appropriate method (e.g., imputation or removal).
- Normalize or standardize features to ensure consistent scaling.
- Return the cleaned and scaled dataset.

Step 2: Train Random Forest for Feature Selection
2. Define the function `train_random_forest(X_train, y_train)`:
- Initialize the Random Forest model with parameters (e.g.`n_estimators=100,max_depth=10`).
- Fit the Random Forest model to the training data (`X_train`, `y_train`).
- Extract feature importance scores from the trained Random Forest model.
- Select the top N important features based on the feature importance scores.
- Return the trained Random Forest model and the list of selected features.

Step 3: Train Artificial Neural Network with Selected Features
3. Define the function `train_ann(X_train, y_train)`:
- Initialize an ANN model using a sequential approach.
- Add an input layer with a specified number of neurons (based on selected features) and activation function (e.g., `ReLU`).
- Add one or more hidden layers with activation functions (e.g., `ReLU`).
- Add an output layer for binary classification (using `sigmoid` activation).
- Compile the ANN model with the specified optimizer and loss function (e.g., `binary_crossentropy`).
- Train the ANN model on the training data (`X_train`, `y_train`) for a specified number of epochs and batch size.
- Return the trained ANN model.

Step 4: Ensemble Strategy - Model Stacking
4. Define the function `ensemble_predictions(rf_model, ann_model, X_test)`:
- Generate predictions from the Random Forest model using the test set (`X_test`).
- Generate predictions from the ANN model using the same test set (`X_test`).
- Combine the predictions from both models using an averaging approach.
- Apply a threshold (e.g., 0.5) to convert the average predictions into binary class labels (0 or 1).
- Return the final combined predictions.

Step 5: Evaluation
5. Define the function `evaluate_model(y_test, final_preds)`:
- Calculate evaluation metrics such as accuracy, precision, recall, and F1-score using the true labels (`y_test`) and the final predictions.
- Print or return the evaluation metrics.
- Main Function - Complete Hybrid Algorithm
- 6. Define the function `hybrid_ann_rf(data)`:
- Call `preprocess_data(data)` to preprocess the dataset, obtaining features `X` and labels `y`.
- Split the dataset into training and testing sets using `train_test_split`.
- Call `train_random_forest(X_train, y_train)` to train the Random Forest model and obtain selected features.

- Create new training and testing datasets using the selected features.
- Call `train_ann(X_train_selected, y_train)` to train the ANN model using the selected features.
- Call `ensemble_predictions(rf_model, ann_model, X_test_selected)` to get final predictions from the ensemble of models.
- Call `evaluate_model(y_test, final_preds)` to evaluate the performance of the hybrid model.

4.2 Prediction

This process enables the prediction of multiple diseases using machine learning models applied to datasets from the UCI Repository. Data cleaning, feature selection, weight factor analysis, and ranking ensure that the most relevant features are used. Models like logistic regression, decision trees, SVMs, and ANNs are employed to predict diseases like kidney disease, heart disease, diabetes, and malaria with high accuracy. Evaluation and tuning steps ensure robust performance across diverse datasets.

This paper focuses on the forecast of various illnesses, with a special focus on the diabetes prediction task. Figure 1 shows the performance measurements of different AI calculations on this task. Based on the research results, it is clear that neural networks and random forest algorithms show superior performance in disease prediction.

Table 4.1. Performance Metrics for ML Model

	MCC	F1 Score	Accuracy
KNN	0.664	0.849	0.85
SVM	0.841	0.924	0.926
Decision Trees	0.664	0.842	0.851
Random Forest	0.978	0.99	0.99
ANN	0.962	0.99	0.99
Proposed RBHML	0.979	0.99	0.99

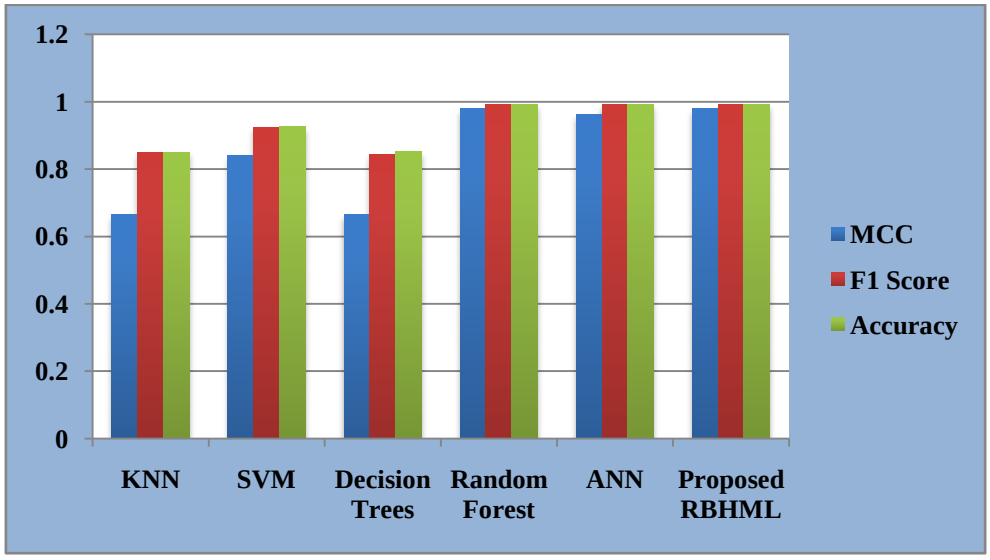

Figure 4.1. Performance Evaluation for Proposed with Machine Learning Model

6. CONCLUSION

Multimodal disease prediction using machine learning and deep learning algorithms has the potential to revolutionize the healthcare industry by improving diagnosis, treatment, and patient outcomes.Our approach demonstrates that the integration of machine learning and deep learning techniques can lead to more accurate and reliable disease predictions for a variety of diseases.By leveraging large data sets and advanced algorithms, we can identify disease symptoms and risk factors with high accuracy, as well as predict disease progression and treatment response.Our results highlight the potential of machine learning and deep learning to improve early disease detection and personalized treatment planning, ultimately leading to better patient outcomes and lower healthcare costs.

REFERENCES

[1] Marcos-Zambrano, Laura Judith, KanitaKaraduzovic-Hadziabdic, TatjanaLoncarTurukalo, PiotrPrzymus, Vladimir Trajkovik, Oliver Aasmets, MagaliBerland et al. "Applications of machine

learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment." Frontiers in microbiology 12 (2021): 313.

[2] Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis." Health and Technology 11 (2021): 87-97.

[3] Elngar, Ahmed A., Rishabh Kumar, Amber Hayat, and PrathameshChuri. "Intelligent system for skin disease prediction using machine learning." In Journal of Physics: Conference Series, vol. 1998, no. 1, p. 012037. IOP Publishing, 2021.

[4] Goel, Rati. "Heart disease prediction using various algorithms of machine learning." Proceedings of the International Conference on Innovative Computing & Communication (ICICC). 2021.

[5] Gupta, Jay Prakash, Ashutosh Singh, and Ravi Kant Kumar. "A computer-based disease prediction and medicine recommendation system using machine learning approach." Int J Adv Res EngTechnol (IJARET) 12.3 (2021): 673-683.

[6] Lakshmanarao, A., M. Raja Babu, and T. Srinivasa Ravi Kiran. "Plant disease prediction and classification using deep learning ConvNets." 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV).IEEE, 2021.

[7] Sudharsan, M., and G. Thailambal. "Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA)." Materials Today: Proceedings (2021).

[8] Nadeem, Muhammad Waqas, et al. "Fusion-Based Machine Learning Architecture for Heart Disease Prediction." Computers, Materials & Continua 67.2 (2021).

[9] Patel, Bharati, and AakankshaSharaff. "Rice crop disease prediction using machine learning technique." International Journal of Agricultural and Environmental Information Systems (IJAEIS) 12.4 (2021): 1-15.

[10] Chowdhury, Mohammed NowshadRuhani, Ezaz Ahmed, Md Abu Dayan Siddik, and AkhlakUzZaman. "Heart disease prognosis using machine learning classification techniques."In 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1-6.IEEE, 2021.

[11] Joshi, Ram D., and Chandra K. Dhakal. "Predicting type 2 diabetes using logistic regression and machine learning approaches." International journal of environmental research and public health 18.14 (2021): 7346.

[12] Hasan, Ruby. "Comparative analysis of machine learning algorithms for heart disease prediction." ITM Web of Conferences.Vol. 40.EDP Sciences, 2021.

[13] Kibria, HafsaBinte, Abdul Matin, NusratJahan, and Sanzida Islam. "A comparative study with different machine learning algorithms for diabetes disease prediction." In 2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), pp. 1-8. IEEE, 2021.

[14] Krishna, P. Gopala, and D. StalinDavid. "An Effective Parkinson's Disease Prediction Using Logistic Decision Regression and Machine Learning with Big Data." Turkish Journal of Physiotherapy and Rehabilitation 32.3 (2021): 778-786.

[15] Wu, Chia-Tung, Guo-Hung Li, Chun-Ta Huang, Yu-Chieh Cheng, Chi-Hsien Chen, Jung-YienChien, Ping-Hung Kuo, Lu-Cheng Kuo, and Feipei Lai. "Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study." JMIR mHealth and uHealth 9, no. 5 (2021): e22591.

[16] Srivastava, Asmit, and Ashishkumar Singh. "Heart Disease Prediction using Machine Learning." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE).IEEE, 2022.

[17] Riyaz, Lubna, Muheet Ahmed Butt, MajidZaman, and OmeeraAyob. "Heart disease prediction using machine learning techniques: a quantitative review." In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 3, pp. 81-94. Springer Singapore, 2022.

[18] Bhavekar, Girish S., and Agam Das Goswami. "A hybrid model for heart disease prediction using recurrent neural network and long short term memory." International Journal of Information Technology 14.4 (2022): 1781-1789.

[19] Raju, K. Butchi, Suresh Dara, AnkitVidyarthi, V. MNSSVKR Gupta, and Baseem Khan. "Smart heart disease prediction system with IoT and fog computing sectors enabled by cascaded deep learning model." Computational Intelligence and Neuroscience 2022 (2022).

[20] Boukhatem, Chaimaa, HebaYahia Youssef, and Ali BouNassif. "Heart disease prediction using machine learning." 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE, 2022.

[21] Pallathadka, Harikumar, Malik Mustafa, Domenic T. Sanchez, GunaSekharSajja, SanjeevGour, and MohdNaved. "Impact of machine learning on management, healthcare and agriculture." Materials Today: Proceedings 80 (2023): 2803-2806.

[22] Khan, Arsalan, MoizQureshi, Muhammad Daniyal, and KassimTawiah. "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction." Health & Social Care in the Community 2023 (2023).
[23] Abdulsalam, Ghada, SouhamMeshoul, and HadilShaiba. "Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach." Intell.Autom. Soft Comput 36 (2023): 761-779.
[24] Gupta, Gaurav, Shakir Khan, VandanaGuleria, AbrarAlmjally, Bayan IbrahimmAlabduallah, TamannaSiddiqui, Bader M. Albahlal, Saad Abdullah Alajlan, and Mashael Al-Subaie. "DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms." Diagnostics 13, no. 6 (2023): 1093.
[25] Yousaf, Fatima, SajidIqbal, Nosheen Fatima, TanzeelaKousar, and MohdShafryMohd Rahim. "Multi-class disease detection using deep learning and human brain medical imaging." Biomedical Signal Processing and Control 85 (2023): 104875.