UtilitasMathematica
ISSN 0315-3681 Volume 121, 2024

AN ENSEMBLE MODEL FOR MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Ramasamy R[1]Dr G Pattabirani[2] and Dr S Parasuraman[3]
[1]Research Scholar, Department of Information Technology, Annamalai University, Chidambaram, Tamil Nadu, India.
rams.ms01@gmail.com
[2]Assistant Professor, Department of Information Technology, Annamalai University, Chidambaram, Tamil Nadu, India.
prvijayaraja@gmail.com
[3]Co-Guide and Professor, Department of ECE,
Karpaga Vinayaga College of Engineering and Technology, Chengalpattu
parasuhodece78@gmail.com

ABSTRACT

Multiple Disease Prediction using Machine Learning, Deep Learning is a comprehensive project aimed at predicting various diseases including diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. This project leverages machine learning algorithms such as TensorFlow with Keras, Support Vector Machine (SVM), Logistic Regression and Proposed Hybrid algorithm. The models are deployed using Streamlit Cloud and the Streamlit library, providing a user-friendly interface for disease prediction. The primary goal is to develop a comprehensive system that can predict the likelihood of multiple diseases based on patient data. By using both machine learning (ML) and deep learning (DL) methods, the system can be designed to analyze various types of medical information such as clinical data to detect early signs of disease, provide accurate diagnoses, or even forecast the progression of these conditions. Once the models are trained, they must be evaluated for their predictive performance using metrics such as accuracy, precision, recall and F1-score. This helps in determining how well the model can predict various diseases. Cross-validation techniques can be applied to ensure that the model generalizes well across different patient populations. The research focuses on building an intelligent, multi-disease prediction system by combining traditional machine learning models with deep learning approaches for handling different types of medical data. The research holds immense potential in improving personalized healthcare and reducing the burden of chronic diseases worldwide. The high accuracies achieved by the different models demonstrate the effectiveness of the employed machine learning algorithms in disease prediction.

Keywords: *Machine Learning, Streamlit, Tensor Flow, Keras, SVM, Logistic Regression, Diabetes, Heart Disease, Kidney Disease, Parkinson's Disease, Breast Cancer.*

I. INTRODUCTION

In recent years, machine learning has seen significant progress and applications across various sectors, notably in healthcare. Predicting multiple diseases simultaneously using machine learning models holds immense promise for revolutionizing medical diagnostics and enhancing patient outcomes. This study delves into the utilization of ML to predict the presence of three prevalent diseases: heart disease, diabetes, and Parkinson's disease. These conditions pose substantial challenges to individuals and healthcare systems worldwide,

making early detection and accurate diagnosis crucial for better patient prognosis and cost effective treatment. Machine learning, with its capacity to analyze extensive datasets and discern intricate patterns, offers promising avenues for multi-disease prediction. ML as robust supervised learning models, are extensively employed for classification tasks[1]. They seek to identify an optimal solution that effectively separates different classes in the data, maximizing the margin between them. The versatility of ML in handling both linear and nonlinear relationships between input features and target variables makes them suitable for various medical diagnostic applications.

This research aimed to develop a multi-disease prediction framework using ML and assess its performance in predicting heart disease, diabetes, and Parkinson's disease [2]. By leveraging publicly available datasets and employing appropriate feature engineering techniques, a comprehensive dataset encompassing relevant demographic, clinical, and biomarker information was constructed. The ML model was trained on this dataset to discern the intricate relationships between input features and the presence of the three diseases.

Accurate disease prediction using machine learning models can facilitate early interventions, personalized treatment plans, and targeted disease management strategies. It holds promise for assisting healthcare providers in making informed decisions, improving patient care, and optimizing resource allocation within healthcare systems. Additionally, it offers potential for population-level disease surveillance, enabling prompt detection of disease outbreaks and implementation of preventive measures. The findings of this research contribute to the growing literature on machine learning-based disease prediction, specifically focusing on the application of ML for multi-disease prediction.

The evaluation and analysis of the ML model's performance in predicting heart disease, diabetes, and Parkinson's disease shed light on the feasibility and effectiveness of using machine learning algorithms in complex medical diagnose[3]. In summary, this research underscores the potential of ML as a valuable tool in the multi-disease prediction domain. Leveraging machine learning can bring us closer to achieving more accurate, timely, and personalized healthcare interventions, ultimately leading to improved patient outcomes and more efficient healthcare systems.

II. LITERATURE REVIEW

Abbasi, E. Y*et al.,*(2024) used the deep learning of Endoscopic Ultrasonography (EUS) images to predict whether the malignant potential of gastrointestinal stromal tumours. First let the EUS image be resized in format through Lanczos interpolation. The deep learning part uses 20 CNN kernels for the first layer and 50 for the second layer. Suitable for using time relationship improves the performance of the model for detecting sudden heart failure in a short observation window of 12–18 months.

Ahsan, M. M *et al.,* (2022) where Multi diseases that have been recognized harm greenhouse or field-grown tomatoes. Machine learning was used to identify different diseases in tomato plant leaves. The project aimed to have the deep learning algorithms run in real-time on the robot. As a result, the robot can detect plant illnesses while traveling on the field or in the greenhouse, either manually or automatically.

Allegra, A., *et. al.,*(2022) presented a comparative study of traditional machine learning methods and deep learning methods in the early detection of AD and progression of MCI to AD with the use of a deep learning 16 studies out of which 4 were using deep learning approaches and traditional machine learning together and 12 studies were using only deep learning approaches.

Arya, A. D., *et al.,* (2024) developed an ensemble of ML and deep learning (DL) models to predict the disease with accuracy rate of 88.70%. This study employed a total of six classification algorithms developed a stacking ensemble model after applying SVM, NB, and KNN with a 10-fold cross-validation synthetic minority oversampling technique (SMOTE) in order to balance out imbalanced datasets

Aslam, N*et al.,* (2022) incorporated ML algorithms used on raw walkway data to distinguish between MS patients and healthy controls. They focused on constructing a series of novel features to enhance standard parameters which in turn improves the model's performance. Hence, they used an instrumented walkway to generate rich data that are usually unnoticed by clinicians.

Bhardwaj, P *et. al.,* (2024) compared two classification models, SVM and Naive Bayes, to classify if a patient had Dengue or not. To determine the best SVM hyper parameters, a Grid Search was performed changing the gamma parameter and the cost. Despite executing the Grid Search, neither the best configuration nor the configuration of the Naive Bayes model was detailed.

Chandrasekhar, N *et al.*(2023) conducted a comparative assessment of widely used machine learning and deep learning algorithms to predict heart disease prevalence. The classification methods are based on the Cleveland Dataset, which is freely available. Different models compared, and their ability to predict cardiac disease was evaluated.

III. METHODOLOGY

The methodology followed in this research work is data set creation, preprocessing and preparing a data set for application machine learning based on classification of multi disease datasets.

The methodology followed in this research work is detailed in Figure 1.

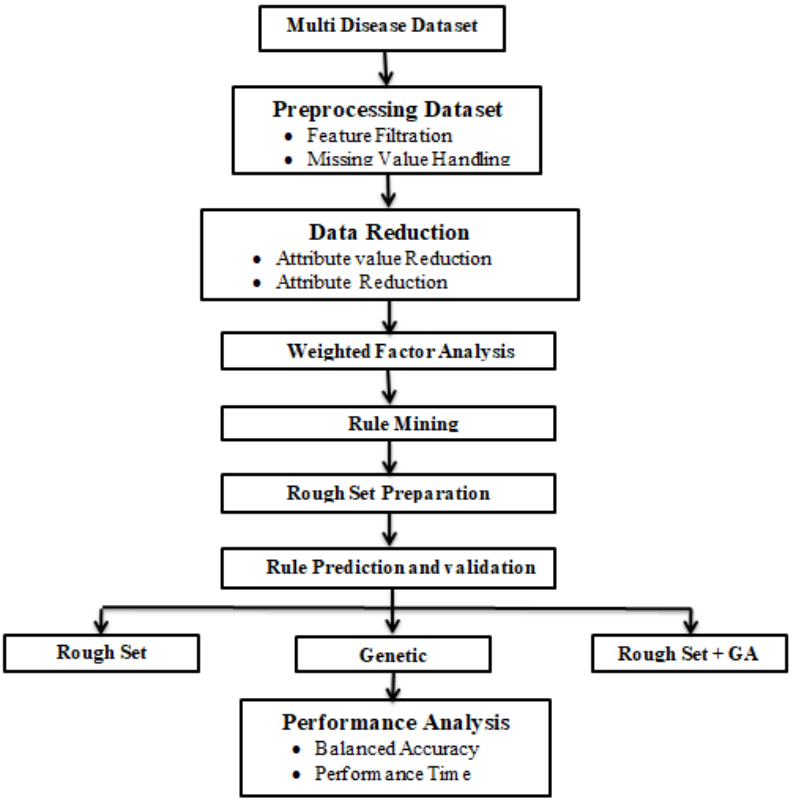

Figure3.1: Research Methodology

To build a machine learning or deep learning model for multi-disease prediction using various conditions like diabetes, Parkinson's disease, kidney disease, heart disease, and breast cancer, in this work to follow a structured process for dataset preparation.

3.1 Data Collection

To obtain datasets related to each disease from trusted sources like public health databases, research institutions, or open-source platforms (e.g., Kaggle, UCI Machine Learning Repository, etc.). Diabetes Dataset: Typically includes features like glucose levels, blood pressure, BMI, age, insulin, etc. Parkinson's Dataset: Contains motor symptoms, voice samples, or diagnostic features like tremors and rigidity [4].Kidney Disease Dataset: Includes attributes like creatinine levels, blood pressure, GFR (Glomerular Filtration Rate), etc. Heart Disease Dataset: Contains risk factors like cholesterol, age, blood pressure, chest pain type, etc. Breast Cancer Dataset: Often includes mammography results, tumor size, patient history, and sometimes genetic markers.

Hence, this research work was implemented using a data set called Multi-Disease Dataset (MDD) that was created from real life patient's details. The database consists of Thousand Twenty-Four patient's data on generic, symptoms and disease related factors. The created MDD dataset of this research work is depicted in Figure 2, while Table 1 lists the Attributes of the MDD dataset.

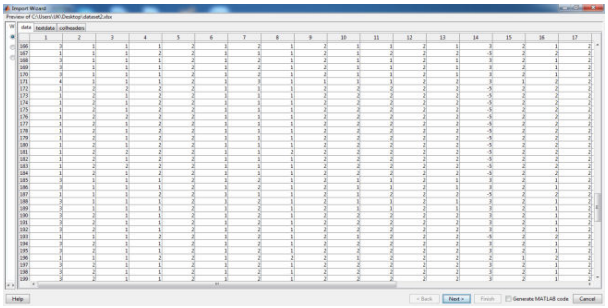

Figure 3.2: MDD Dataset

3.2 Data Integration

After collecting disease-specific datasets, the goal is to combine them into a unified dataset for multi-disease prediction. Steps include:

- **Standardization of Features:**
  - Ensure that similar features (e.g., age, gender, BMI) across datasets are consistent in units and meaning.
  - Rename columns to maintain uniformity. For example, "Age" in the diabetes dataset should match "Age" in the heart disease dataset.
- **Feature Engineering:**
  - If the datasets contain different features, need to either engineer new features or handle missing data.
  - Common features across datasets (e.g., age, gender, BMI) can serve as the primary input features. For unique features (e.g., glucose for diabetes or creatinine for kidney disease), this work use one-hot encoding, scaling, or imputation techniques if applicable.
- **Add new labels (target) to distinguish between diseases.** For example, `Disease Label` can represent:
  - 0 for Healthy/No Disease
  - 1 for Diabetes

- 2 for Parkinson's
- 3 for Kidney Disease
- 4 for Heart Disease
- 5 for Breast Cancer

- **Handling Missing Data:**
  - Some datasets may have missing values for specific features. We using techniques like mean/median imputation, KNN imputation, or dropping rows with excessive missing values.
  - For features missing in one disease dataset but available in others (e.g., glucose might not exist in the Parkinson's dataset), handle these with missing values or use zero-padding where appropriate.

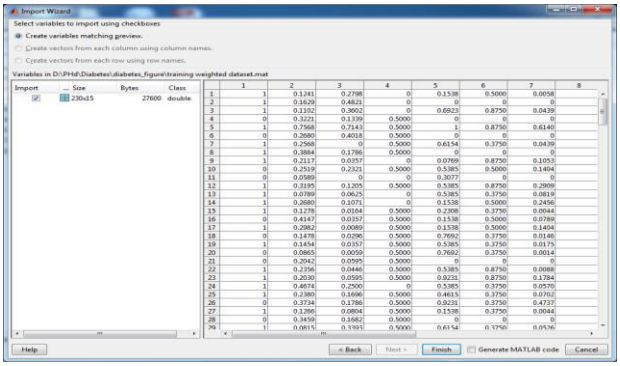

Figure 3.3: Weighted Dataset

The captured values and attributes are exhaustive and complicated; hence this research work implemented a weighted factor analysis on the attributes for reducing the size of the data set. Common traits possessed by patients and selected for the reduction process[5].The objective of this process was to remove unrelated value attributes or neutralize outliers by filling a mean mode value. Each attribute was set using a weight, based on the importance of the attribute. The weighted data set, a reduced version of the MDD dataset can be stored easily. This data set reduced processing time and was aimed to improve accuracy and efficiency. Figure 3 depicts the weighted dataset.

3.3 Ranking (Multiclass Classification)

Multi-label Setup: Since you are predicting multiple diseases, each sample should have a target class (disease Rank). This can be done by assigning a unique rank for each disease.
Example Ranking:

- Healthy (No disease): Label = 0
- Diabetes: Label = 1
- Parkinson's: Label = 2
- Kidney Disease: Label = 3
- Heart Disease: Label = 4
- Breast Cancer: Label = 5

In case of patients with multiple diseases (which may occur in real-world data), use multi-label classification techniques, where a patient can be marked with multiple diseases.

3.4 Balancing the Dataset

- Real-world medical datasets can be imbalanced, meaning some diseases (like heart disease) may have more data points than others (like Parkinson's).

- Use oversampling (e.g., SMOTE) or under sampling techniques to balance the dataset, ensuring no class is underrepresented.
- Also generate synthetic samples using techniques like data augmentation in case of highly imbalanced datasets.

3.5 Dataset Partitioning

- Split the dataset into Training, Validation, and Testing sets (e.g., 70% training, 15% validation, 15% testing).
- That each partition has a similar distribution of disease classes (stratified splitting).

3.6 Feature Selection

Reduce the dimensionality of the dataset by removing irrelevant or redundant features. The following techniques are used:

- Correlation Analysis is used toremove features that are highly correlated to avoid multicollinearity.
- Recursive Feature Elimination (RFE) to select features that are most important for the prediction task.
- PCA (Principal Component Analysis) are useful for reducing feature dimensions, especially for imaging or genetic data.

Table 3.1 Example of Dataset Structure

Age	Gender	BMI	Glucose	Creatinine	Blood Pressure	Tremors	Cholesterol	Tumor Size	Disease_Label
45	Male	25	150	0.9	120/80	0	200	NaN	1 (Diabetes)
62	Female	28	100	1.1	140/90	1	220	NaN	2 (Parkinson)
38	Male	30	NaN	3.5	110/70	0	190	NaN	3 (Kidney)
50	Female	27	130	NaN	130/85	0	240	NaN	4 (Heart)
55	Female	22	NaN	NaN	NaN	0	NaN	2.5	5 (Breast)
...	...	...	...	...	...	...	...	...	...

Once the dataset is prepared, you can proceed with training machine learning and deep learning models like Random Forest, SVM, CNN, or LSTM for disease prediction.

IV. RESULTS AND DISCUSSIONS

The proposed novel algorithm Rough Set Based Machine Learning Algorithm (RSMLA) for detection of MDD was implemented and evaluated on Python with MATLAB on inteli5 system with 16gb ram running windows[6].The proposed algorithm takes the weighted dataset as an input and generates a decision rule set for core attributes based on weightage of the attributes. Algorithm 1 lists RSMLA

Scheme Evaluation

Data: MDD Patients Data Set
Result: The mean performance values
Step 1: M=12 :No of Data Set
Step 2: i=1;
Step 3: while i<=M do
Step 4: Read Patients Data Set D[i];
Step 5: Split Data set Intances using % split;
Step 6: Train[i]=60% of D; % Training Data;

Step 7: Learning(Train[i],scheme);
Step 8: Test Data=D[i]-Train[i];% Test Data;
Step 9: Result=TestClassifier(Test[i],Learner);
Step 10: end

Prediction Algorithm
Input: MDD Data set with attributes
Output: Multiple disease prediction accuracy rate measure
Step 1: Initialize MDD data set
Step 2: Analysis Label information
Step 3: Generate random wise to input weight of attributes
Step 4: Calculate weights based on the rule prediction
Step 5: Classify prediction tool based on max weight
Step 6: Regression classification data with weighted value
Step 7: populate weighted matrix for selection process
Step 8: Find maximum fitness of weighted attributes.
Step 9: Crossover weighted population and regenerate population
Step 10 : Classify attributes with weight to predict diabetics
Step 11: Using confusion matrix to measure accuracy
The dependencies like partial or total dependencies in data are identified in RSMLA and redundant data is eliminated, thus clearing the data for rule generation. This refined data is then used for rough set creation and then a machine learning algorithm is executed on the rough set to clearly predict MDD in patients from collected attributes[7].Table 2 lists the accuracy table of the proposed RSMLA, while Figure 4 shows the time complexity and accuracy plot of RSGBA

Table 4.1Accuracy table of RSMLA

iteration	**Training time**	**Testing time**	**Avgtraining accuracy**	**Avg Testing Accuracy**	**Elapsed Time**
10	0.0072	0.0001	0.8780	0.8795	16.632
20	0.0031	0.0019	0.8868	0.8757	32.884
30	0.0140	0.0006	0.9004	0.8636	49.44
40	0.0112	0.0006	0.9049	0.8641	65.34
50	0.0144	0.0016	0.9127	0.8668	81.582
60	0.0184	0.0028	0.9169	0.8607	98.118
70	0.0337	0.0037	0.9206	0.8615	115.43
80	0.0309	0.0025	0.9237	0.8560	131.98
90	0.0437	0.0028	0.9297	0.8481	148.951
100	0.0415	0.0025	0.9330	0.8459	165.93

.

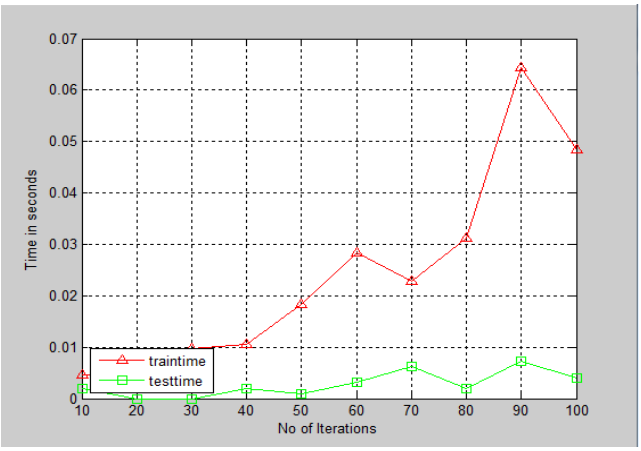

Figure 4.1:RSMLA Time Complexity Plot.

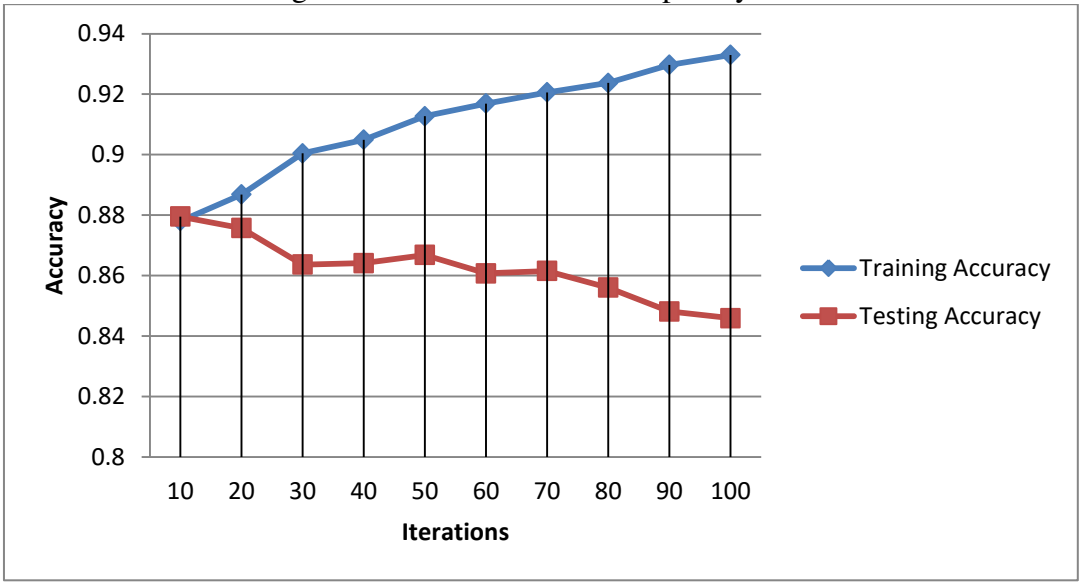

Figure 4.2: RSMLA Accuracy Plot.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

Patient clinical charts and online guidelines are the most important available knowledge sources for physicians. They make future predictions regarding specific observations of the patient. Automatic prediction systems can help physicians to understand patient and make proper decisions.This work designed a hybrid Rough set reasoning model from generated rules in the decision making process. RSMLA reduces generation and understandable decision rules extraction using ARM algorithm from the original MDD dataset. Experimental results for the prediction model reveal that performance of the model is above 95% in classifying multi-disease prediction [8].Correlation-based trends analysis results suggest insights of the patient conditions to physicians in an appropriate way and assisting them in controlling risky stages.Although the proposed hybrid model performs well, further experiments and improvements are required. Additionally, the proposed model can be applied to prediction problems in other fields. Thus, it can be concluded that RSMLA has performed decently with less processing time for prediction of MDD data set.

REFERENCES

[1] Abbasi, E. Y., Deng, Z., Ali, Q., Khan, A., Shaikh, A., Al Reshan, M. S., ...&Alshahrani, H. (2024). A machine learning and deep learning-based integrated multi-omics technique for leukemia prediction. *Heliyon*, *10*(3).

[2] Allugunti., Singla, P., Sharma, R., Kukreja, V., &Bansal, A. (2022, October). Deep learning based multi-classification model for rice disease detection. In 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-5). IEEE.

[3] Ahsan, M. M., Luna, S. A., &Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541).MDPI.

[4] Allegra, A., Tonacci, A., Sciaccotta, R., Genovese, S., Musolino, C., Pioggia, G., &Gangemi, S. (2022). Machine learning and deep learning applications in multiple Disease prediction prognosis, and treatment selection.Cancers, 14(3), 606.

[5] Arya, A. D., Verma, S. S., Chakarabarti, P., Chakrabarti, T., Elngar, A. A., Kamali, A. M., &Nami, M. (2024). A systematic review on machine learning and deep learning techniques in the effective diagnosis of Multi disease.Brain Informatics, 10(1), 17.

[6] Aslam, N., Khan, I. U., Bashamakh, A., Alghool, F. A., Aboulnour, M., Alsuwayan, N. M., & Al Ghamdi, K. (2022). Multiple sclerosis diagnosis using machine learning and deep learning: challenges and opportunities. Sensors, 22(20), 7856.

[7] Bhardwaj, P., Kumar, Y., & Mishra, S. (2024, January). Machine Learning-Based Approaches for the Prognosis and Prediction of Multiple Diseases.In *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1-6).IEEE.

[8] Chandrasekhar, N., &Peddakrishna, S. (2023). Enhancing Multi disease prediction accuracy through machine learning techniques and optimization.Processes, 11(4), 1210.

[9] Da Silva Neto, S. R., Tabosa Oliveira, T., Teixeira, I. V., Aguiar de Oliveira, S. B., Souza Sampaio, V., Lynn, T., & Endo, P. T. (2022). Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review.PLoS neglected tropical diseases, 16(1), e0010061.

[10] El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2024). Multi-label active learning-based machine learning and deep learning model for heart disease prediction. Sensors, 22(3), 1184.

[11] Hemachandran, K., Alasiry, A., Marzougui, M., Ganie, S. M., Pise, A. A., Alouane, M. T. H., &Chola, (2023). Performance analysis of deep learning algorithms in diagnosis of Multi disease.Diagnostics, 13(3), 534.

[12] Hussain, I., &Nazir, M. B. (2024). Empowering Healthcare: AI, ML, and Deep Learning Innovations for Brain and Heart Health. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(4), 167-188.

[13] Islam, M. M., Adil, M. A. A., Talukder, M. A., Ahamed, M. K. U., Uddin, M. A., Hasan, M. K., &Debnath, S. K. (2023). Deep Crop: Machine learning-based crop disease prediction with web application. Journal of Agriculture and Food Research, 14, 100764.

[14] Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023, January). A review of machine learning and deep learning approaches on Multi disease prediction diagnosis. In Healthcare (Vol. 11, No. 3, p. 285).MDPI.

[15] Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., &Tavera Romero, C. A. (2022). Early-stage Alzheimer's disease prediction using machine learning models.Frontiers in public health, 10, 853294.

[16] Khandakar, S., Al Mamun, M. A., Islam, M. M., Hossain, K., Melon, M. M. H., &Javed, M. S. (2024). Unveiling Early Detection And Prevention Of Multi. Machine Learning And Deep Learning Approaches. Educational Administration: Theory and Practice, 30(5), 14614-14628.
[17] Khandakar, S., Al Mamun, M. A., Islam, M. M., Hossain, K., Melon, M. M. H., &Javed, M. S. (2024). Unveiling Early Detection And Prevention Of Cancer: Machine Learning And Deep Learning Approaches. *Educational Administration: Theory and Practice*, *30*(5), 14614-14628.
[18] Kumar, D. V. S., Chaurasia, R., Misra, A., Misra, P. K., &Khang, A. (2023). Multi disease prediction using machine learning.In Data-centric AI solutions and emerging technologies in the healthcare ecosystem (pp. 205-214).CRC Press.
[19] MunishKhanna, Singh, L. K., &Garg, H. (2024). A novel approach for human diseases prediction using nature inspired computing & machine learning approach. *Multimedia Tools and Applications*, *83*(6), 17773-17809.
[20] Reddy, T. S. E., Sripathi, S. R., Akula, D., Palaniswamy, S., &Subramani, R. (2022, December). Multi disease prediction using machine learning and deep learning.In 2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) (pp. 1-5).IEEE.
[21] Sahid, M. A., Hasan, M., Akter, N., &Tareq, M. M. R. (2022, July). Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning.In 2022 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6).IEEE.
[22] Sandhiya, S., Rao, G. V. V., Prabhu, V., Mohanraj, K., &Azhagumurugan, R. (2022, December). Disease Prediction Using Machine Learning Algorithm.In 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (pp. 1-5).IEEE.
[23] Sarra, R. R., Dinar, A. M., Mohammed, M. A., Ghani, M. K. A., &Albahar, M. A. (2022). A robust framework for data generative disease prediction based on efficient deep learning models.Diagnostics, 12(12), 2899.
[24] Saxena, M. K., Sharma, R., Kumar, R., & Kumar, R. (2022). Disease Prediction Using Machine Learning and Deep Learning.proceeding on International Journal for Research in Applied Science & Engineering Technology, 10.
[25] Tiwari, H. (2022). Early prediction of Multi disease using deep learning approach.In Deep learning for medical applications with unique data (pp. 107-122).Academic Press.
[26] Vayadande, K., Golawar, R., Khairnar, S., Dhiwar, A., Wakchoure, S., Bhoite, S., &Khadke, D. (2022, May). Heart disease prediction using machine learning and deep learning algorithms.In *2022 international conference on computational intelligence and sustainable engineering solutions (CISES)* (pp. 393-401).IEEE.
[27] Vetrithangam, V., Devi, A., &Aggarwal, S. (2022, October). Multi disease prediction based on combined deep reinforcement Boltzmann machines. In AIP Conference Proceedings (Vol. 2555, No. 1).AIP Publishing.
[28] Yu, Z., Wang, K., Wan, Z., Xie, S., &Lv, Z. (2023). Popular deep learning algorithms for disease prediction: a review. Cluster Computing, 26(2), 1231-1251.