International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2s, 2025

Multiple Disease Prediction Using Machine Learning Algorithm (XGBoost Algorithm)

Shaik Sihab[1] , Syed Akber Quadri[2] , Mohammed Abdul Ibrahim[3], Mrs.T Anita[4].

[1,2,3]B.E Students, Department Of CSE, ISL Engineering College HYD India.

[4] Assistant Professor, Department Of CSE, ISL Engineering College HYD India.

ABSTRACT:

Every day, many individuals encounter different illnesses. The prognosis of a disease is the most pivotal part of treatment. The enormous increase in healthcare and medical data has enabled accurate medical data analysis, which aids in early sickness discovery and proactive patient care. This study focuses on analyzing extensive medical data by employing supervised classification algorithms, with a primary focus on the XGBoost (Extreme Gradient Boosting) Classifier. The proposed model anticipates the most probable disease based on symptoms and predicts the likelihood of whether a person might be suffering from a particular illness. XGBoost is known for its high performance, scalability, and ability to handle imbalanced and sparse data, making it well-suited for complex healthcare datasets. By combining predictions with XGBoost as the core model, the system achieves improved diagnostic accuracy and reduced false positives compared to traditional individual models. This study enhances the swiftness of clinical decision-making and assists healthcare organizations in providing timely and precise early patient care. It also supports medical professionals in formulating more effective patient treatment strategies.

1. INTRODUCTION:

In today’s rapidly evolving healthcare landscape, early diagnosis and timely treatment are critical to improving patient outcomes. With millions of individuals experiencing various illnesses daily, the accurate prognosis of a disease becomes a cornerstone of effective treatment planning. The surge in healthcare data, driven by advancements in digital health records and medical technologies, has paved the way for intelligent data-driven decision-making in clinical environments. Analyzing this massive and complex medical data using machine learning techniques enables early detection of illnesses, better resource allocation, and enhanced patient monitoring. Among the many algorithms available, supervised classification methods have shown remarkable promise in predictive healthcare analytics. This study emphasizes the use of the Extreme Gradient Boosting (XGBoost) Classifier for disease prediction based on patient symptoms. XGBoost has emerged as a leading machine learning algorithm due to its exceptional performance, scalability, and robustness in handling sparse and imbalanced datasets—common challenges in healthcare data. By leveraging XGBoost, the proposed system can predict the likelihood of specific diseases with greater accuracy and efficiency. The model utilizes symptom-based inputs to forecast the most probable disease and determine the risk level associated with it. Compared to conventional machine learning models, the integration of XGBoost enhances prediction precision while minimizing false positives. This contributes to faster, more reliable clinical decision-making. Furthermore, the system supports healthcare professionals by offering a data-backed, intelligent diagnostic tool that complements their expertise. It enables more accurate, timely, and personalized patient care. Ultimately, this approach improves diagnostic reliability and fosters more strategic treatment planning. The study aims to bridge the gap between data science and clinical medicine through innovative computational methodologies.

2. LITERATURE SURVEY

Title:
Machine Learning for the Multiple Disease Prediction System

Author:
Kumar Bibhuti B. Singh, Ashutosh Sharma, Ashish Verma, Ranjeet Maurya
Year: 2022
Description: Disease prediction, which aims to identify individuals who are at risk of contracting specific diseases, is a crucial component of healthcare. Recently, machine learning algorithms have shown to be effective tools in the fight against illness prediction due to their superior ability to sort through large datasets in search of complex patterns. The development of Machine Learning (ML) in the contemporary healthcare period has created new opportunities for the diagnosis and treatment of chronic illnesses. In this paper we are proposes a complete Multiple Disease Prediction System that makes accurate predictions of diabetes, cancer, and heart disease using machine learning algorithms. The system's purpose is to analyse intricate medical datasets and find trends and risk factors related to these illnesses. The system uses cardiovascular data

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP117-123

analysis and logistic regression to detect heart disease and provide a probabilistic evaluation of heart health. Convolutional Neural Networks, which evaluate medical imaging to find malignancies with high precision, are used to simplify cancer detection. Finally, Support Vector Machines are used to predict diabetes by taking into account a variety of metabolic and genetic indicators to evaluate. Making it simpler for people to detect their own health issues with just their symptoms and exact vital signs is the aim of this project. The proposed approach improves both the predictive power and precision of sickness.

Title: Multiple Disease Prognostication Based on Symptoms Using Machine Learning Techniques

Authors:
K. Patil, S. Pawar, P. Sandhyan, J. Kundale
Year: 2024
Description: This paper presents a machine learning-based system for predicting multiple diseases based on user-reported symptoms. The approach addresses challenges posed by overlapping symptoms and limited healthcare resources. By inputting symptoms, users receive immediate prognoses and health maintenance suggestions. The system employs various machine learning algorithms to ensure rapid and reliable predictions. The study emphasizes the importance of such tools in enhancing healthcare accessibility and accuracy, particularly in regions with a low doctor-patient ratio.

.
Title: A Novel Method for the Detection and Classification of Multiple Diseases Using Transfer Learning-Based Deep Learning Techniques with Improved Performance
Authors:
K. Natarajan, S. Muthusamy, M.S. Sha, et al.
Year: 2024
Description: This study introduces an advanced deep learning framework employing transfer learning to classify multiple diseases from medical images. Utilizing architectures like VGG16, ResNet50, InceptionV3, and EfficientNetB4, the model processes diverse imaging modalities, including chest X-rays, skin lesions, MRI scans, and retinal fundus images. The integration of a channel attention mechanism enhances the model's focus on critical features, improving diagnostic precision. Data augmentation techniques are applied to bolster model robustness against image quality variability. Among the models tested, EfficientNetB4 achieved the highest accuracy of 94.04%. The research underscores the potential of deep learning in facilitating early and accurate disease diagnosis across various organ systems.

3.METHODOLOGIES:

The approach ensured that the platform remains scalable, user-centric, and aligned with real-world healthcare needs.The following methodological steps were undertaken to develop a disease prediction system using the XGBoost classifier:

1. Data Collection

Collected a comprehensive dataset containing patient symptoms and corresponding diagnosed diseases. Sources of data included publicly available medical datasets (e.g., from Kaggle, UCI Machine Learning Repository) and synthetic data where necessary. Ensured data privacy and compliance with healthcare data standards.

2. Data Preprocessing

Data Cleaning: Removed or imputed missing values, handled duplicates, and corrected inconsistencies.
Label Encoding: Converted categorical symptom and disease names into numerical format suitable for machine learning.
Feature Selection: Identified the most relevant symptoms for each disease to reduce dimensionality and improve model accuracy.

3. Model Selection And Training

Selected XGBoost Classifier for its robustness, high accuracy, and ability to handle missing and sparse data.
Tuned hyperparameters using Grid Search or Randomized Search with cross-validation to optimize performance.
Trained the model on the training set using supervised learning principles.

4. Model Evaluation

Evaluated the model performance using the testing set based on the following metrics:
Accuracy, Precision, Recall, F1-score and
Compared results with traditional classifiers (e.g., Decision Tree, Random Forest, Logistic Regression) to highlight XGBoost’s superiority

5. Disease Prediction System Design

Integrated the trained XGBoost model into a user-interactive interface (e.g., web or desktop application).
Users input the symptoms, and the model predicts the most probable disease.

6. Testing and Validation

Testing included:

- Functionality Testing: Across desktop and mobile environments.

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP117-123

- Usability Testing: To ensure intuitive design and user flow.

- Performance Testing: To evaluate response times and system load under real conditions.

4.REQUIREMENTS ENGINEERING:

We can see from the results that on each database, the error rates are very low due to the discriminatory power of features and the regression capabilities of classifiers. Comparing the highest accuracies (corresponding to the lowest error rates) to those of previous works, our results are very competitive.

Hardware Requirements

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It should what the system do and not how it should be implemented.

• PROCESSOR	: INTEL i5 11 GEN
• RAM	: 8GB SSD RAM
• HARD DISK	: 512 GB SSD.

Software Requirements

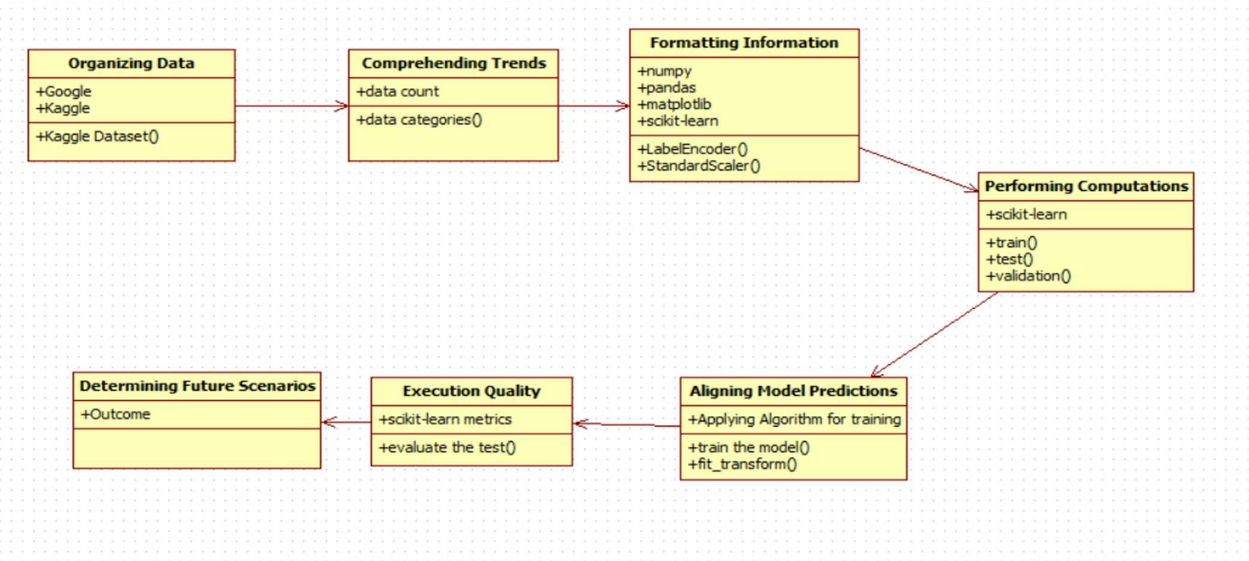
The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

• Operating System	:Windows 10
• Platform	: Spyder3
• Programming Language	: Python
• Front End	: Spyder3

5.DESIGN ENGINEERING :

Design Engineering deals with the various UML [Unified Modelling language] diagrams for the implementation of project. Design is a meaningful engineering representation of a thing that is to be built. Software design is a process through which the requirements are translated into representation of the software. Design is the place where quality is rendered in software engineering.

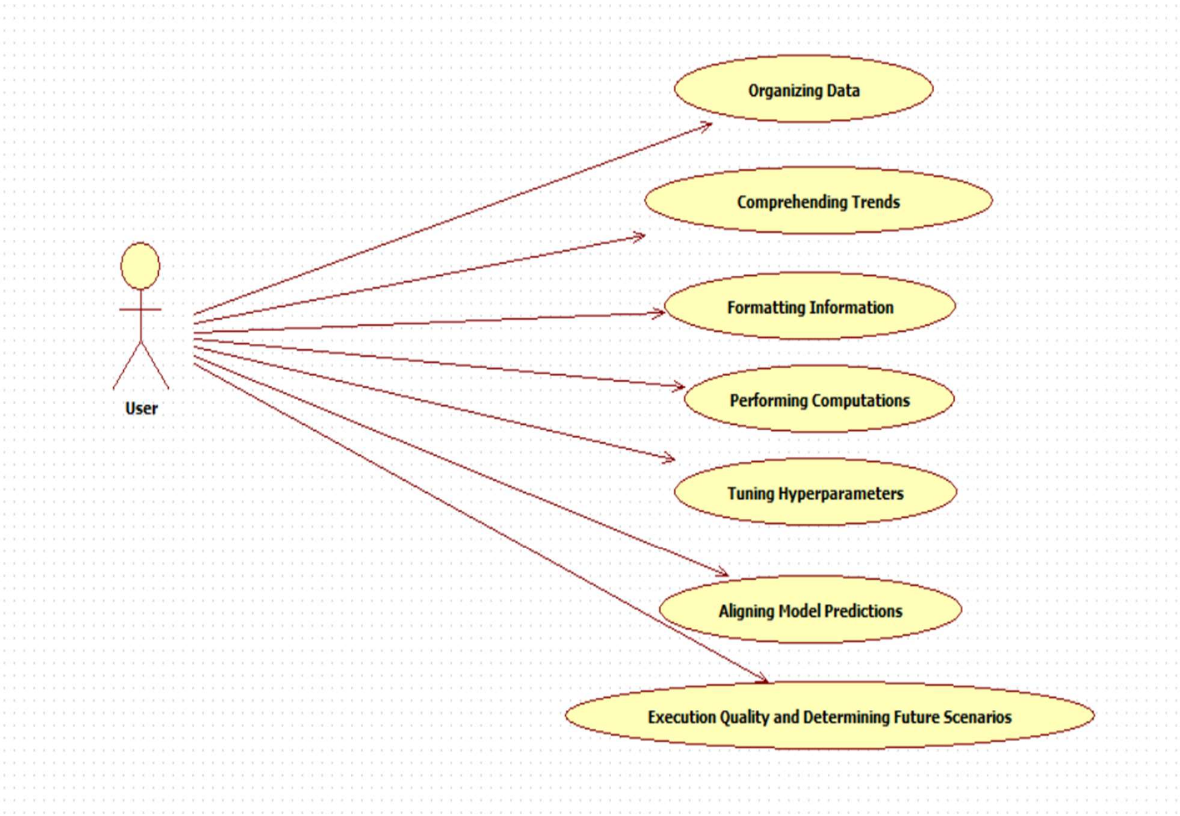
CLASS DIAGRAM:



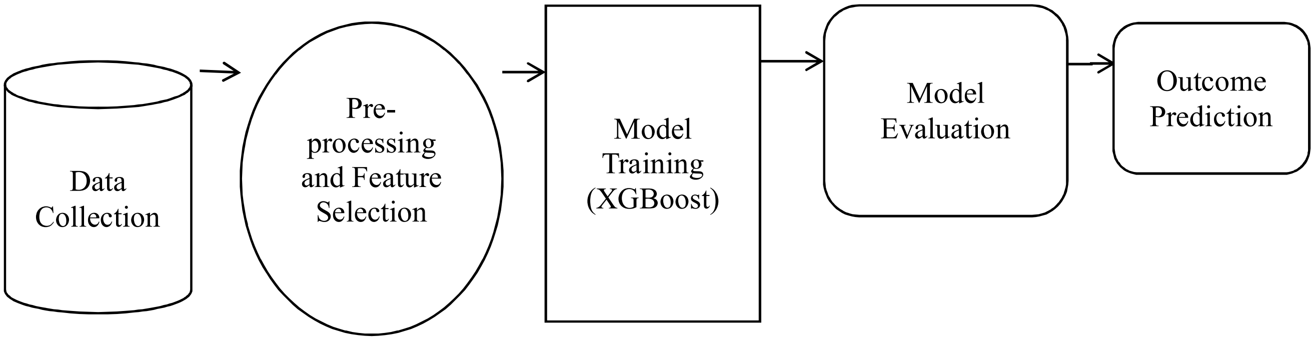
EXPLANATION:

In this class diagram represents how the classes with attributes and methods are linked together to perform the verification with security. From the above diagram shown the various classes involved in our project.

USECASE DIAGRAM :



SYSTEM ARCHITECTURE:



EXPLANATION:

The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. The above diagram consists of user as actor. Each will play a certain role to achieve the concept.

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP117-123

6.IMPLEMENTATION :

Introduction:
The implementation phase of this project involves translating design specifications into a functional website. This section details the process of developing, integrating, and deploying this website and focusing on the implementation of key features and technologies.
It's implementation spans across various phases, ensuring a comprehensive and multifaceted approach to health management. Each phase/stage contributes uniquely to the platform's overall functionality, enhancing the user experience and addressing diverse health needs.
The implementation of the Multiple Disease Prediction System is carried out in several modular stages, ensuring a clean, scalable, and reproducible machine learning workflow. The core algorithm utilized is the *XGBoost (Extreme Gradient Boosting) Classifier*, chosen for its robustness, scalability, and superior performance on structured and imbalanced healthcare datasets.

1. Data Collection And Preparation:
The system utilizes publicly available and cleaned datasets for three major diseases: Diabetes,
Heart Disease, and
Parkinson's Disease.
Each dataset contains structured health-related features such as glucose level, blood pressure, BMI, heart rate, voice frequency measurements, etc.
*Diabetes Dataset: Contains features like glucose level, insulin, age, BMI, etc.
*Heart Disease Dataset: Includes chest pain type, resting ECG, maximum heart rate, cholesterol, etc.
*Parkinson's Dataset: Includes voice frequency and amplitude-related features like MDVP\:Fo(Hz), jitter, shimmer, etc.

Before model training, the data undergoes preprocessing including:

* Handling missing values (if any)
* Feature scaling using *StandardScaler*
* Label encoding (if required)
* Train-test split (80:20 ratio)

2. Model Building Using XGBoost:
Each disease is treated as an independent binary classification problem. The XGBoost Classifier is trained separately for each disease due to differences in features and diagnosis criteria.

3. Training Process:
* The model is trained on the training set and validated on the test set
* Accuracy and classification metrics are computed to evaluate performance
* Trained models are saved using joblib for deployment or further prediction

4. Disease Prediction Module:
A dedicated prediction script is implemented to accept input features from the user and return the prediction result using the corresponding trained XGBoost model.
Input is passed as a list or vector matching the model's expected feature dimensions
The system outputs a binary prediction (e.g., 0 = No Disease, 1 = Disease Detected)
Predictions are real-time and optimized for speed

5. Evaluation Metrics:
Each trained model is evaluated based on:
*Accuracy Score: Correct predictions out of total predictions
*Confusion Matrix: True positives, false positives, false negatives, true negatives
*Precision, Recall, F1-Score: To handle class imbalance and false positives
These metrics validate that the XGBoost model outperforms traditional classifiers in terms of both precision and robustness, especially on complex medical datasets.

7. FUTURE ENHANCEMENTS:

In the future, this disease prediction system can be significantly enhanced to offer more advanced and comprehensive healthcare support. One key improvement would be integrating real-time symptom tracking using wearable health devices, enabling continuous health monitoring. The system can also be connected to Electronic Health Records (EHR) for more detailed patient analysis and personalized predictions. Incorporating Natural Language Processing (NLP) would allow the model to interpret free-text symptom inputs, making it more user-friendly. Expanding the dataset to include rare and emerging diseases will further improve its diagnostic coverage and accuracy. To reach a broader audience, multi-language support can be introduced. A feedback mechanism could also be added to allow the system to learn from real-world usage and enhance its performance over time. Mobile app integration would improve accessibility, especially for remote or rural areas. Collaborating with hospitals for clinical validation could ensure reliability and encourage real-world adoption. Additionally, implementing advanced data privacy measures, such as blockchain technology, would further strengthen data security and user trust.

DOI: https://doi.org/10.62647/IJITCE2025V13I2sPP117-123

7. CONCLUSION :

In conclusion, the proposed disease prediction system using the XGBoost algorithm offers a powerful and efficient approach to early diagnosis based on user-reported symptoms. By leveraging machine learning on large-scale medical datasets, the system enhances diagnostic accuracy while minimizing false positives. Its ability to handle imbalanced and sparse data makes it highly suitable for complex healthcare scenarios. The model supports both patients and healthcare providers by offering quick, reliable, and data-driven insights. With its user-friendly interface and scalability, the system can be integrated into various healthcare platforms, including mobile and web applications. It not only assists in timely medical interventions but also helps in formulating effective treatment plans. Furthermore, the solution reduces the workload on healthcare professionals through automated screening. Overall, this project demonstrates the potential of AI-driven tools in improving healthcare quality, accessibility, and efficiency. It lays the foundation for future innovations in intelligent, personalized, and technology-assisted medical care..

8.REFERENCES :

[1] Md R. Hoque & M. Sajedur Rahman. (2020). Predictive modelling for chronic disease: Machine learning approach. In Proceedings of the 4th International Conference on Compute and Data Analysis (pp. 97–101). CA, USA.

[2] Perwej, Y., Ahamad, F., Khan, M. Z., & Akhtar, N. (2021). An Empirical Study on the Current State of Internet of Multimedia Things (IoMT). International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 8(3), 25–42. https://doi.org/10.1617/vol8/iss3/pid85026

[3] Littell, C. L. (1994). Innovation in medical technology: Reading the indicators. Health Affairs, 13(3), 226–235. https://doi.org/10.1377/hlthaff.13.3.226

[4] Perwej, Y., Alzahrani, M. Y., Mazarbhuiya, F. A., & Husamuddin, M. (2018). The State-of-the-Art Cardiac Illness Prediction Using Novel Data Mining Technique. International Journal of Engineering Sciences & Research Technology (IJESRT), 7(2), 725–739. https://doi.org/10.5281/zenodo.1184068

[5] Mobeen, A., Shafiq, M., Aziz, M. H., & Mohsin, M. J. (2022). Impact of workflow interruptions on baseline activities of the doctors working in the emergency department. BMJ Open Quality, 11(3).

[6] Ahmed, S., Szabo, S., & Nilsen, K. (2018). Catastrophic healthcare expenditure and impoverishment in tropical deltas: Evidence from the Mekong delta region. International Journal for Equity in Health, 17(1), 1–13.

[7] Roberts, M. A., & Abery, B. H. (2023). A person-centered approach to home and community-based services outcome measurement. Frontiers in Rehabilitation Science, 4.

[8] Perwej, Y. (2015). An Evaluation of Deep Learning Miniature Concerning in Soft Computing. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 4(2), 10–16. https://doi.org/10.17148/IJARCCE.2015.4203

[9] Miljkovic, D., et al. (2016). Machine Learning and Data Mining Methods for Managing Parkinson's Disease. Lecture Notes in Artificial Intelligence (LNAI), 9605, 209–220.

[10] Van Stiphout, M. A. E., Marinus, J., Van Hilten, J. J., Lobbezoo, F., & De Baat, C. (2018). Oral health of Parkinson's disease patients: A case-control study. Parkinson's Disease, Article ID 9315285, 8 pages. https://doi.org/10.1155/2018/9315285

[11] Perwej, Y. (2015). The Bidirectional Long-Short-Term Memory Neural Network Based Word Retrieval for Arabic Documents. Transactions on Machine Learning and Artificial Intelligence (TMLAI), 3(1), 16–27. https://doi.org/10.14738/tmlai.31.863

[12] Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. Healthcare, 9(12), Article 1712. https://doi.org/10.3390/healthcare9121712

[13] Anila, M., & Pradeepini, G. (2020). A Review on Parkinson's Disease Diagnosis Using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 9(6).

[14] Perwej, Y. (2022). Unsupervised Feature Learning for Text Pattern Analysis with Emotional Data Collection: A Novel System for Big Data Analytics. IEEE International Conference on Advanced Computing Technologies & Applications (ICACTA'22). https://doi.org/10.1109/ICACTA54488.2022.9753501

[15] Cao, J., Wang, M., Li, Y., & Zhang, Q. (2019). Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment. PLOS ONE, 14(4).

[16] Perwej, Y., Hann, S. A., & Akhtar, N. (2014). The State-of-the-Art Handwritten Recognition of Arabic Script Using Simplified Fuzzy ARTMAP and Hidden Markov Models. International Journal of Computer Science and Telecommunications (IJCST), 8, 26–32.

[17] Hamidi, H., & Daraee, A. (2016). Analysis of preprocessing and post-processing methods and using data mining to diagnose heart diseases. International Journal of Engineering, Transactions B: Applications, 29(7), 921–930.

[18] Maurano, M., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science, 337, 1190–1195. https://doi.org/10.1126/science.1222794

[19] Kumar, A. (2021). Disease Prediction and Doctor Recommendation System Using Machine Learning Approaches. International Journal for Research in Applied Science and Engineering Technology (IJRASET). https://doi.org/10.22214/IJRASET.2021.36234

[20] Perwej, Y., Parwej, F., & Akhtar, N. (2018). An Intelligent Cardiac Ailment Prediction Using Efficient ROCK Algorithm and K-Means & C4.5 Algorithm. European Journal of Engineering Research and Science (EJERS), 3(12), 126–134. https://doi.org/10.24018/ejers.2018.3.12.989

[21] Patil, K., Pawar, S., Sandhyan, P., & Kundale, J. (2022). Multiple Disease Prognostication Based on Symptoms Using Machine Learning Techniques. ITM Web of Conferences, 44, 03008. https://doi.org/10.1051/itmconf/20224403008

[22] Takura, T., Goto, K. H., & Honda, A. (2021). Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour. BMC Medicine, 19(1), 1–16.

[23] Akhtar, N., Rahman, S., Sadia, H., & Perwej, Y. (2021). A Holistic Analysis of Medical Internet of Things (MIoT). Journal of Information and Computational Science (JOICS), 11(4), 209–222. https://doi.org/10.12733/JICS.2021/V11I3.535569.31023

[24] Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. BMC Medical Informatics and Decision Making, 16(3), 197–208.

[25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[26] Ravi, S. K., Chaturvedi, S., Rastogi, N., Akhtar, N., & Perwej, Y. (2022). A Framework for Voting Behavior Prediction Using Spatial Data. International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 10(2), 19–28. https://doi.org/10.55524/ijircst.2022.10.2.4

[27] Ren, Q., Cheng, H., & Han, H. (2017). Research on machine learning framework based on random forest algorithm. AIP Conference Proceedings, 1820, 080020.

[28] M.A.Bari & Shahanawaj Ahamad,” Process of Reverse Engineering of Enterprise InformationSystem Architecture” in International Journal of Computer Science Issues (IJCSI), Vol 8, Issue 5, ISSN: 1694-0814, pp:359-365,Mahebourg ,Republic of Mauritius , September 2011

[29] Dr.Abdul Bari ,Dr. Imtiyaz khan , Dr. Rafath Samrin , Dr. Akhil Khare , “ VPC & Public Cloud Optimal Perfomance in Cloud Environment “,Educational Administration: Theory and Practic,ISSN No : 2148-2403 Vol 30- Issue -6 June 2024

[30] Dr. Mohammed Abdul Bari,Arul Raj Natraj Rajgopal, Dr.P. Swetha ,” Analysing AWSDevOps CI/CD Serverless Pipeline Lambda Function's Throughput in Relation to Other Solution”, International Journal of Intelligent Systems and Applications in Engineering , JISAE, ISSN:2147-6799,