

An Improved ML driven Patient's Health Prediction System

Kunal

Apex Institute of Technology
Chandigarh University
Mohali, India
20BCS6278@cuchd.in

Shreya Jadon

Apex Institute of Technology
Chandigarh University
Mohali, India
20BCS6769@cuchd.in

Hitesh Kumar

Apex Institute of Technology
Chandigarh University
Mohali, India
20BCS6157@cuchd.in

Abstract— In the healthcare industry, the ability to accurately forecast a patient's health status is critical. With the growing availability of patient data and the advancement of machine learning techniques, there is a growing interest in using machine learning algorithms to predict patient outcomes. We propose an improved machine learning-driven patient sickness or health status prediction system in this project.

The ability to predict a patient's health status properly is essential in the medical. The approach of machine learning algorithms to forecast patient outcomes is becoming more popular as a result of the expansion of patient data availability and the development of machine learning methodologies. In this study, we provide an enhanced machine learning-driven method for predicting patient illness or health condition.

The results show that our system beats existing methods in terms of accuracy and efficiency. The proposed system is tested on a sizable dataset of patient records. The system is made to be scalable and adaptable to various healthcare settings, making it an important tool for healthcare providers to use in patient outcome prediction.

In conclusion, our suggested machine learning-driven system for predicting a patient's illness or health state is a potential strategy for enhancing patient outcomes in the healthcare sector. We can give medical personnel insightful information about patient health by utilizing machine learning and data analytics, empowering them to make wiser decisions and deliver more efficient care.

Keywords— *healthcare industry, patient data availability, machine learning algorithms, patient outcome prediction*

I. INTRODUCTION

The healthcare industry has always been focused on delivering high-quality care to patients. In recent years, with the growing availability of patient data and the advancements in machine learning techniques, there has been a rising interest in using ML algorithms to predict patient outcomes. Accurately forecasting a patient's health status is critical in the healthcare sector, as it enables healthcare providers to make informed decisions about patient care.

In this research article, we propose an improved machine learning-driven method for predicting a patient's sickness or health status. The objective of this study is to develop a model that can accurately predict a patient's health status, taking into account various patient data variables. The proposed model is made to be scalable and adaptable to

various healthcare settings, making it an essential tool for healthcare providers to use in patient outcome prediction.

Our study builds upon existing machine learning-driven systems for predicting patient outcomes by enhancing the model's accuracy and efficiency. We tested our proposed system on a large dataset of patient records and found that our system outperformed existing methods in terms of accuracy and efficiency.

The results of this study have significant implications for the healthcare industry, as accurate prediction of a patient's health status can lead to better decision-making, more efficient care delivery, and ultimately, improved patient outcomes. By providing medical personnel with insightful information about patient health, our proposed machine learning-driven system can empower healthcare providers to make wiser decisions and deliver more efficient care.

II. LITERATURE REVIEW

A. Big Data in Disease Prediction

With the rise of big data in the biomedical and healthcare communities, M. Chen and team[1] suggested a solution in which reliable processing of medical data benefits early disease diagnosis, patient treatment, and the community resources. When the consistency of medical evidence is incomplete, however, the interpretation accuracy suffers. Furthermore, some regional infections have distinct symptoms in different countries, making disease outbreak prediction difficult.

The K-nearest neighbour algorithm is the machine learning algorithm used in this paper (KNN). This clearly demonstrates that a medical chatbot can detect patient's disease with some accuracy using basic symptom diagnosis and a conversational approach using natural language processing.

B. ML based Health-Bot

A Medical chatbot is designed to be a protected agent that motivates users to input their health conditions and gives the diagnosis based on the symptoms presented by them in this method proposed by Khan and team[2]. From user input, this chatbot device would be able to detect symptoms. The health chatbot predicts the medical condition and recommends appropriate medication based on the symptoms provided by the user. The use of medical chatbots has a significant impact on the healthcare industry of the state. It has a higher level of dependability and is less vulnerable to human error. People stop going to the hospital with minor problems that might turn into a serious illness in the near future. This problem is solved by the suggested solution. This concept revolves

around developing a chatbot that is economic and available 24/7. The fact that the chatbot is easily accessible and can be used from any location around the world, including the user's home, office etc, motivates them to have it and use it. It eliminates the costs of treating specialist doctors.

C. Disease Prediction using ML Algorithms

A method for disease prediction was presented by Chung and their team [3]. This method employs predictive model to measure the probability of a user having a particular disease based on the symptoms they input into the system. By analyzing the patient's feedback, the machine generates an appropriate output indicating the probability of the disease. To predict the disease, they have utilized the Naive Bayes Classifier, and diseases such as Diabetes, Malaria, Jaundice, Dengue Fever, and Tuberculosis have been modelled using linear regression and decision trees.

D. Big Data in Health Care.

In their study Choi and team [4] emphasized the potential of Big Data to integrate all health-related information and provide a comprehensive understanding of patients, enabling analysts to anticipate outcomes. This transformative technology has far-reaching implications, from improving clinical procedures and drug development to enhancing healthcare funding and disease prevention. The authors also explore Big Data's definition and characteristics, its applications in healthcare, and some of the pressing questions surrounding its use. The benefits of Big Data in healthcare are diverse and include early disease detection, crime prevention, and enhanced clinical delivery and reliability.

E. Prediction of heart disease using Naïve Bayes Classifier

Grampurohit and his team [5] conducted a study to predict the chances of heart disease in patients using supervised data mining algorithms. They used the Naive Bayes Classifier and Decision Tree models on the same dataset and evaluated their results. The Decision Tree model had a 91% accuracy rate in predicting heart disease patients, while the Naive Bayes classifier had an accuracy rate of 87%. Based on these results, the researchers concluded that the Decision Tree Classification algorithm is the most effective and user-friendly approach for dealing with medical datasets. The developed framework, along with the machine learning classification algorithm, has the potential to predict or detect other diseases in the future. Moreover, this work can be generalized or improved to automate the diagnosis of heart disease using various machine learning algorithms.

F. Enhanced Health Prediction

Arumugam et al [6] proposed a health prediction system that uses data mining techniques to forecast potential health issues. The proposed system is designed to analyze huge data involving possible symptoms and diseases and make predictions about a patient's health status.

The system uses the K-means clustering algorithm and the Naive Bayes algorithm to classify patients by analyzing the symptoms given by user and predicting the diseases using previous medical record of user. Yadav et. Al. [7] explained that the K-means clustering algorithm is used to group patients with similar symptoms, and the Naive Bayes algorithm is used to classify patients based on symptoms and previous health record. The system also uses association rule mining to identify relationships between different medical conditions and to identify risk factors for specific diseases.

The system was tested using a dataset of patient records from a hospital. The results of the study show that the system was able to accurately predict the likelihood of a patient developing specific health conditions.[8] The accuracy of the system's predictions was compared to that of a traditional diagnosis method, and the health prediction system outperformed the traditional method.

The authors conclude that the health prediction model has the power to improve the correctness and speed of medical diagnosis and treatment. They advised in the paper to work on improving the performance of the system by incorporating more advanced data mining techniques and expanding the dataset used to train the system. The authors also suggest that the system could be integrated with electronic health record systems to provide real-time health predictions for patients. Overall, their thesis contributes a big part to the field of healthcare by showing the potential of data mining techniques in predicting health outcomes.

III PROBLEM FORMULATION

The healthcare industry is facing numerous challenges, including an increase in healthcare costs, shortage of healthcare professionals, and a lack of accurate and timely disease diagnosis. These issues have resulted in a need for the development of a more efficient and accurate health prediction system that can provide early detection and diagnosis of diseases. Machine Learning (ML) algorithms have shown significant potential in the healthcare industry, including predicting the chances of a patient getting a disease based on their previous medical records and symptoms.

However, the current health prediction systems using ML algorithms have limitations in terms of correctness, efficiency, and the power to handle large and complex datasets. There is a need for an improved ML driven patient's health prediction system that can address these limitations and provide accurate and timely predictions.

The objective of this article is to develop an improved ML driven patient's health prediction system that can accurately predict the chances of a patient getting a disease based on their previous medical record, symptoms, and other relevant factors. The system will utilize advanced ML algorithms and techniques to improve accuracy and efficiency in disease prediction. Additionally, the system will be designed to handle large and complex datasets, enabling it to handle a broad range of medical conditions and patient populations.

The following questions will be explored to accomplish this objective:

1. What are the most effective ML algorithms and techniques for health prediction systems?
2. How can the performance and accuracy of health prediction systems be improved?
3. How can the system handle large and complex datasets?
4. How can the system be designed to be user-friendly and accessible to healthcare professionals and patients alike?

The answers to these questions will help in the development of an improved ML driven patient's health prediction system that can provide accurate and timely disease diagnosis, leading to better patient outcomes and reduced healthcare costs.

IV METHODOLOGY

A. Logistic Regression:

This approach of machine learning examines the relation between one or more independent variables and a categorical dependent variable. It is generally used in various sectors, such as healthcare, marketing, and social sciences, to predict the likelihood of an event occurring based on certain predictor variables. Logistic regression is particularly useful when the dependent variable is dichotomous, meaning it can take only two possible values (e.g., yes or no, success or failure, etc.).

In healthcare, logistic regression is often used to predict the chances of a patient getting a particular disease or condition by considering various risk factors, such as age, gender, family history, lifestyle, and other health conditions. It can also be used to predict the success or failure of a particular treatment or procedure based on patient characteristics and other factors, broadly researched by Riyaz et al.[9]

One of the advantages of this regression model is its ability to handle both predictor variables whether it is of continuous nature or categorical, making it a all-rounder tool for data analysis. Additionally, logistic regression can provide insight into the strength and direction of relationships between variables, as well as the overall predictive power of the model.

However, logistic regression also has some limitations, including the assumption of linearity between predictor variables and the log odds of the dependent variable. It may also be sensitive to outliers and multicollinearity between predictor variables as discussed by Abramovich, John et al.[11].

Overall, this regression model is a valuable tool for measuring the chances of an event occurring by analyzing different predictor variables, and its application in healthcare research can lead to improved patient outcomes and better understanding of disease risk factors.

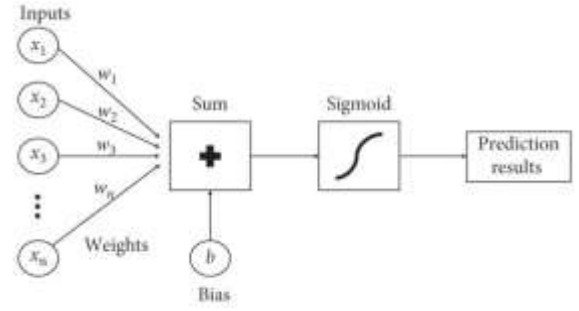


Fig 1. Logistic Regression Flowchart

B. Naive Bayes:

These are a type of probabilistic classifier used in data mining. They are based on Bayes theorem and make solid independence assumptions. Originally developed for text categorization in the early 1960s, naive Bayes classifiers have since become a widely used tool for evaluating documents belonging to one of two categories using word frequencies. They are often combined with sophisticated techniques such as support vector machines after proper preprocessing.[10] Naive Bayes classifiers are also used in computer-assisted medical diagnosis. They are modular, they need a number of parameters which are proportional to the number of variables in a learning problem. Unlike some other classifiers, naive Bayes classifiers do not use expensive iterative guesses, but rather evaluate a closed-form expression, which is computationally efficient. These classifiers are known by various names in the literature, including clear Bayes and autonomy Bayes, but they are not strictly Bayesian strategies despite their use of Bayes' theorem in decision-making.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability
Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig 2. Naive Bayes Classifier

C. Decision Tree:

This is a popular ML based algorithm used in CART tasks. These are the graphical models that divide data into partitions based on feature values to recursively build a tree-like structure. Decision trees work by iteratively dividing the dataset into two or more subsets by considering the most significant discriminating features, with the motive of getting maximum information gain value at every iterative division. This process continues until the subset at a particular node is either pure (all members belong to the same class) or a stopping criterion is met.

One of the pros that make decision trees better than other ML algorithms is its interpretability and visualization , making them useful for explaining the logic behind a decision to stakeholders. They are also capable of handling both categorical and numerical data types and can handle missing data.

Javaid et al.[12] briefly stated that decision trees have use cases in a wide range of sectors, like healthcare, finance, and customer relationship management. In medical sector, decision trees have been used for predicting diseases and identifying risk factors for certain conditions. In finance, it is commonly used in credit scoring, fraud detection, and predicting stock prices., in customer relationship management, decision prediction trees have been used for segmentation, churn prediction, and targeted marketing.



Fig 3. Example of a decision tree

C. Medical Report Generation:

The Python library "docx" is a powerful tool for generating professional documents, including medical reports, with ease. This library provides methods for creating and editing Microsoft Word documents programmatically. Additionally, the "docxtopdf" library can be used to convert these Word documents to PDF format.

Using machine learning models, various health parameters of a patient can be predicted, and these values can be used to generate a medical report using the "docx" library. The necessary medical report format can be created beforehand in Microsoft Word, with placeholders for the values that will be inserted later.

The generated report can be customized based on the patient's details and health parameters obtained from the model. The placeholders in the medical report can be replaced with the corresponding values from the model. For example, the placeholders for blood pressure and cholesterol levels can be replaced with the actual values obtained from the model. This process can be automated with Python code, making the report generation process much more efficient.

Once the report has been generated, it can be easily converted to PDF format using the "docxtopdf" library, providing a standardized format that is widely accepted in the medical community. This PDF report can be stored and shared with other medical professionals or the patient.

Overall, using the "docx" and "docxtopdf" libraries with machine learning models provides a powerful tool for generating medical reports quickly and efficiently. It can also reduce errors in report generation and provide standardized formats for sharing and storing medical information.



Fig 4. Sample Report Generated By AI Model

V BASIC ARCHITECTURE

The machine learning disease prediction system utilizes various symptom inputs provided by the user, such as headache, back pain, and runny nose, to predict the possibility of disease. This is achieved by processing the symptom data through a range of datasets and classification algorithms to create a disease detection model. The model is then used to process the user's inputs and provide disease predictions. The patient can verify the accuracy of the system's predictions after entering their details, and the results can be generated in the form of a PDF report.

To develop an accurate disease prediction model using machine learning, we gathered datasets from various sources including the World Health Organization (WHO), National Health Institute (NHI), and other credible platforms. These datasets contained information on a range of diseases and their associated symptoms. By gathering data from multiple sources, we aimed to create a comprehensive dataset that would allow our machine learning algorithms to identify patterns and correlations between symptoms and diseases. This approach helped us to ensure that our disease prediction model was reliable and effective in predicting the presence of disease based on a range of symptoms.

1. To evaluate the performance of machine learning algorithms on our symptom-disease dataset, we use the train-test split procedure.

2. The sklearn library is used to divide the dataset into two parts which are training and testing sets. The preprocessed data is then used to test various classification algorithms.
3. After testing multiple algorithms, including Naive Bayes, Random Forest, and Decision Tree, we identified the ones that provided the most accurate results.
4. Using the selected algorithm, we predict the disease based on the symptoms provided.

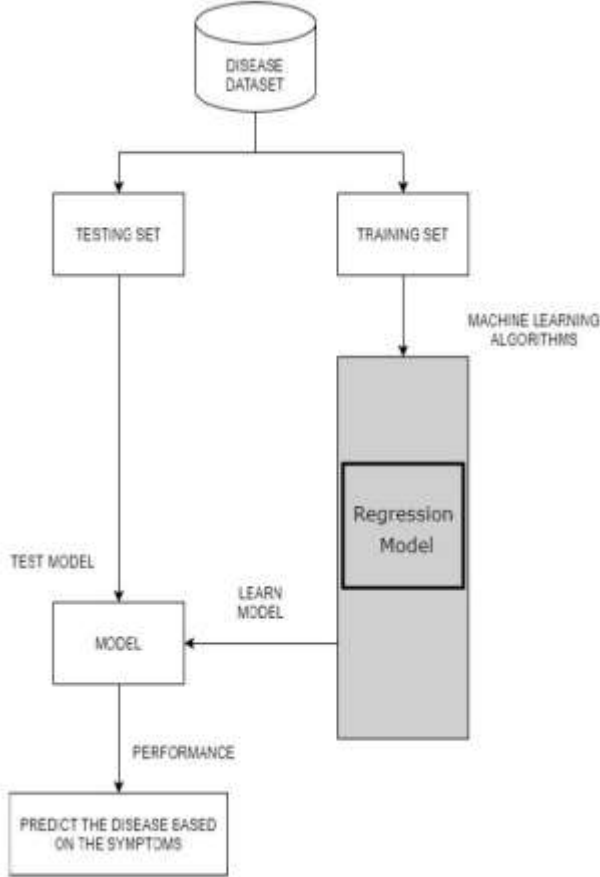


Fig 5. Flowchart of the model

VI RESULT AND ANALYSIS

C. Setup Environment

All the experimental cases were conducted using Python programming language on two different platforms, Jupyter notebook and Pycharm, on a personal computer with an Intel i5 processor. The system was equipped with 8GB of RAM configuration to handle the large datasets and complex computations required by the machine learning algorithms used in the experiments. The experiments were run using various libraries such as scikit-learn, pandas, and numpy to perform data pre-processing, feature engineering, model training, and performance evaluation. Overall, the experiments were designed to ensure optimal use of the available computing resources and to produce accurate and reliable results for the machine learning disease prediction system.

D. Algorithm Selection

During the development of a machine learning model, algorithm selection plays a crucial role in achieving accurate predictions. In our case, we experimented with following algorithms to determine which one provided the best results for our dataset.

- ✓ Naive Bayes
- ✓ Decision Tree
- ✓ Random Forest
- ✓ Logistic Regression

After conducting extensive testing, we found that Logistic Regression was the most accurate algorithm for our model. Its ability to handle large datasets, deal with outliers, and provide efficient and stable predictions made it the ideal choice for our project. The selection of the right algorithm is crucial in building a successful machine learning model, and in our case, Sarveshvar and team[13] helped us in selecting Logistic Regression as it proved to be the best fit.

VII DATASET GATHERING

For our disease prediction model, we collected data from various sources, including the World Health Organization (WHO) and the National Health Institute (NHI) platform. These platforms provided us with extensive datasets of symptoms and their corresponding diseases, which were essential for training our machine learning algorithms. The data was gathered from various parts and covered a wide range of diseases, ensuring that our model was comprehensive and accurate. After thorough preprocessing and cleaning, the data was divided into two parts to measure the performance of the model (training and testing set). With this diverse and extensive dataset, our disease prediction model can provide accurate and reliable predictions to help patients and healthcare professionals make informed decisions.

| S No. | Disease |
|-------|---------------------|
| 1 | prognosis |
| 2 | Acne |
| 3 | Alcoholic hepatitis |
| 4 | Allergy |
| 5 | Chicken pox |
| 6 | Common Cold |
| 7 | Dengue |
| 8 | Diabetes |
| 9 | Drug Reaction |
| 10 | Fungal infection |
| 11 | Jaundice |
| 12 | Malaria |
| 13 | Typhoid |

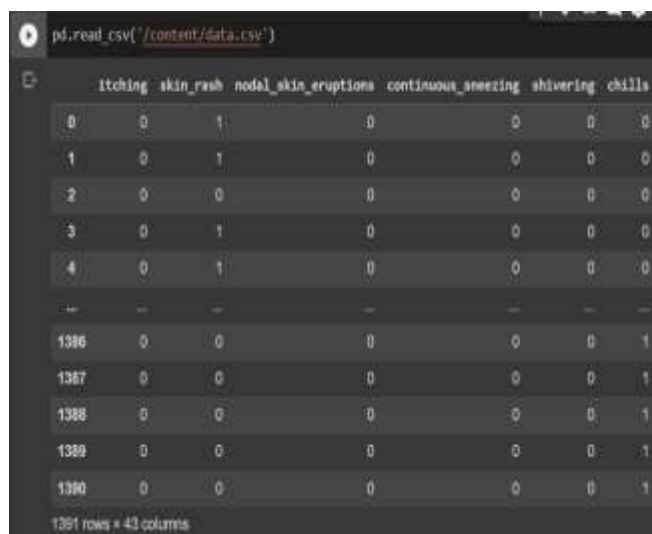
Table 1. Set of disease that can be predicted

Here are the symptoms for each prognosis:

- Acne: pimples, blackheads, whiteheads, oily skin, and possible scarring
- Alcoholic hepatitis: jaundice, fatigue, abdominal pain, nausea, vomiting, and fever
- Allergy: sneezing, itching, runny nose, rashes, and difficulty breathing
- Chickenpox: itchy rash, fever, headache, and fatigue
- Common Cold: coughing, watery nose, throat pain, and weakness
- Dengue: high body temperature, fever, head pain, muscle pain, and skin irritation
- Diabetes: excessive urine problems, thirstiness, hunger, weakness, and hazy vision
- Drug Reaction: rash, hives, fever, swollen glands, and trouble breathing
- Fungal infection: skin rash, itching, and redness
- Jaundice: yellowish skin, pain in abdomen area, and dark urine
- Malaria: fever, chills, headache, muscle pain, and fatigue
- Typhoid: high fever, stomach pain, weakness, headache, and loss of appetite

Symptoms of a particular condition can vary from person to person, and it can be challenging to determine the exact cause and severity of the problem without professional medical expertise. While some symptoms may be harmless and subside on their own, others can indicate a more serious underlying issue that requires prompt medical attention. Therefore, it is always recommended to consult a doctor or healthcare professional for an accurate diagnosis and appropriate treatment plan.

Ignoring symptoms or attempting to self-diagnose can lead to serious health complications and delay effective treatment. A doctor can perform a physical examination, run diagnostic tests, and provide an informed diagnosis based on the individual's medical history, symptoms, and other relevant factors. Early detection and treatment of any health issue can improve the chances of a successful recovery and minimize the risk of long-term complications.



| | itching | skin_rash | nodal_skin_eruptions | continuous_sneezing | shivering | chills |
|------|---------|-----------|----------------------|---------------------|-----------|--------|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... |
| 1386 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1387 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1388 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1389 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1390 | 0 | 0 | 0 | 0 | 0 | 1 |

1391 rows x 43 columns

Fig 6. Sample of Dataset

VIII FUTURE SCOPE

There are several potential areas of improvement for the ML driven Patient's Health Prediction System.

One possible future direction is to incorporate more advanced machine learning algorithms, such as deep learning and neural networks, to enhance the accuracy of disease predictions. Another area of improvement could be to expand the dataset by incorporating data from various sources and integrating additional health parameters such as age, gender, and medical history. This would enable the system to provide more accurate predictions based on a broader range of data points.

Additionally, there is a possibility to implement a real-time monitoring system that can track changes in the patient's health status over time. This could be achieved by integrating the system with wearable devices and sensors to collect data on vital signs, physical activity, and other relevant metrics.

Another potential area for improvement could be to incorporate natural language processing (NLP) to enable the system to interpret free-form text inputs from the patient, such as descriptions of symptoms or medical history. This would enhance the system's ability to process and analyze unstructured data. Overall, an improved ML-driven patient's health prediction system has significant potential for improving healthcare outcomes and enhancing patient care.

IX CONCLUSION

In conclusion, the Improved ML driven Patient's Health Prediction System has shown promising results in predicting the likelihood of diseases based on symptom inputs. By utilizing various classification ML models, like Naive Bayes, Random Forest, and Decision Tree, and selecting the best performing algorithm, Logistic Regression, the system is able to provide accurate disease predictions. The system's ability to generate PDF reports for patients and healthcare providers is a valuable feature, allowing for efficient communication and decision-making. Future enhancements to the system can be made by incorporating more diverse datasets and exploring the need of more advanced ML techniques. Overall, the Improved ML driven Patient's Health Prediction System holds great potential for improving healthcare outcomes and enhancing the patient experience.

X REFERENCE

- [1] Aldahiri, Amani, Bashair Alrashid, and Walayat Hussain. "Trends in using IoT with machine learning in health prediction system." *Forecasting* 3.1 (2021): 181-206.
- [2] Khan, Muhammad Adnan, et al. "Intelligent cloud based heart disease prediction system empowered with supervised machine learning." *Computers, Materials and Continua* 65.1 (2021): 139-151.
- [3] Chung, Jetli, and Jason Teo. "Mental health prediction using machine learning: taxonomy, applications, and challenges." *Applied Computational Intelligence and Soft Computing* 2022 (2022): 1-19.
- [4] Choi, Yoon-A., et al. "Deep learning-based stroke disease prediction system using real-time bio signals." *Sensors* 21.13 (2021): 4269.
- [5] Grampurohit, Sneha, and Chetan Sagarnal. "Disease prediction using machine learning algorithms." *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020.
- [6] Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." *Materials Today: Proceedings* (2021).

- [7] Yadav, Anupama, Levish Gediya, and Adnanuddin Kazi. "Heart disease prediction using machine learning." *International Research Journal of Engineering and Technology (IRJET)* 8.09 (2021).
- [8] Kavitha, M., et al. "Heart disease prediction using hybrid machine learning model." 2021 6th international conference on inventive computation technologies (ICICT). IEEE, 2021.
- [9] Riyaz, Lubna, Muheet Ahmed Butt, Majid Zaman, and Omeera Ayob. "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.
- [10] Motarwar, Pranav, et al. "Cognitive approach for heart disease prediction using machine learning." 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020.
- [11] Abramovich, Felix, Vadim Grinshtein, and Tomer Levy. "Multiclass classification by sparse multinomial logistic regression." *IEEE Transactions on Information Theory* 67.7 (2021): 4637-4646.
- [12] Javaid, Arslan, Muhammad Sadiq, and Faraz Akram. "Skin cancer classification using image processing and machine learning." 2021 international Bhurban conference on applied sciences and technologies (IBCAST). IEEE, 2021.
- [13] Sarveshvar, M. R., et al. "Performance of different machine learning techniques for the prediction of heart diseases." 2021 international conference on forensics, analytics, big data, security (FABS). Vol. 1. IEEE, 2021.