International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-13, Issue-3 (June 2023)
https://ijemr.vandanapublications.com https://doi.org/10.31033/ijemr.13.3.12

# Multiple Disease Prediction System Using ML

Ahsan Ahmad Beg[1], Fazla Maqsood[2] and Dr. Sifatullah Siddiqi[3]
[1]Department of Computer Science and Engineering, Integral University, Lucknow, INDIA
[2]Department of Computer Science and Engineering, Integral University, Lucknow, INDIA
[3]Department of Computer Science and Engineering, Integral University, Lucknow, INDIA

[1]Corresponding Author: sardarmirzaahsan@gmail.com

**Received:** 06-05-2023 **Revised:** 22-05-2023 **Accepted:** 05-06-2023

**ABSTRACT**

Machine learning and ai are critical to many industries. We are everywhere, from driverless cars to healthcare. In the medical industry, the abundance of patient data presents an opportunity for leveraging machine learning techniques to enhance disease detection and diagnosis. In this project, we present a comprehensive Prediction System capable of detecting multiple diseases simultaneously, addressing the limitations of existing systems that often offer lower accuracy and focus on individual diseases. Our system currently focuses on five major diseases: Heart, Liver, Diabetes, Lung Cancer, and Parkinson's disease, with the potential for expansion to include more diseases in the future.

By incorporating various parameters specific to each disease, users can input their data and receive reliable predictions regarding disease presence. The implications of this project are significant, as it enables individuals to monitor their health conditions and take proactive measures, ultimately leading to improved life expectancy. Using the power of machine learning, we aim to contribute to the well-being of countless individuals, providing accurate disease predictions that can potentially save lives.

***Keywords***-- Supervised Learning, Hypothesis Generation, Exploratory Data Analysis, Feature Engineering, Pre-processing Data, Modelling, Logistic Regression, Predictive System, Support Vector Machine (SVM), k-Nearest Neighbours Algorithm (KNN), Deployment, Streamlit Cloud

## I. INTRODUCTION

The Multiple Disease Prediction System is an end-to-end machine learning. The system was developed to analyse various medical conditions and predict the patient's probability of illness[1,2,3]. By using the power of machine learning to accurately predict and diagnose diseases, we aim to transform therapy. Our goal is to improve health by using the power of technology to create a predictive model that can predict a person's ability to produce different types of pain. By analysing extensive medical data and using advanced machine-learning algorithms. We can provide timely and accurate estimates. Our research is focused on developing a good model of that can analyse many patient factors and to predict the probability of certain diseases. Through this work, we want to improve health outcomes, support physicians, and improve the overall health of people worldwide.

### *1.1 Description*

Many analyzes of existing systems in the medical industry consider one disease at a time. For example, one system is used to measure diabetes, another system to diagnose diabetic retinopathy, and another system to predict heart disease. The largest systems focus on specific diseases. Organizations should use several standards when they want to analyze patient health information. [4].

Methods in existing systems can only be used to identify certain diseases. In multi virus prediction, users can identify multiple viruses on a single website. The user doesn't have to go around many places to guess whether he is infected or not. In many disease prediction programs, users have to select the name of a particular disease, enter its parameters, and click submit. The corresponding machine learning model will be called, predict the output and display it on the screen [4,5,6].

### *1.2 Problem System*

The current landscape of machine learning models in healthcare analysis predominantly focuses on individual diseases, necessitating separate analyses for each condition. Liver analysis, cancer analysis, and lung disease analysis are typically treated as isolated entities[5,6,7].

This fragmented approach poses a challenge for users seeking to predict multiple diseases, as they are forced to navigate through various platforms. Regrettably, there is no unified system capable of conducting comprehensive disease predictions across multiple conditions. Moreover, some existing models exhibit suboptimal accuracy, thereby compromising patient well-being. Organizational efforts to analyse patient health reports require the deployment of numerous models, resulting in increased costs and time

consumption. Furthermore, several prevailing systems rely on limited parameters, leading to potentially erroneous outcomes[8].

***1.3 Proposed System***

We present an innovative solution that revolutionizes disease prediction in healthcare analysis. Our proposed system transcends the conventional approach by enabling the simultaneous prediction of multiple diseases. By consolidating diverse analyses into a single unified platform, users can efficiently access accurate predictions for various conditions. With a focus on enhancing both accuracy and efficiency, our model considers a comprehensive set of parameters, ensuring reliable results. By eliminating the need for multiple models and streamlining the prediction process, our system holds the potential to significantly improve healthcare outcomes while optimizing resource allocation[9].

To implement multiple disease analyses, we will utilize machine learning algorithms and the Streamlit framework. When accessing the web application, users can select the specific disease they wish to predict and input the corresponding parameters. Streamlit will then invoke the appropriate model and provide the patient's status as the output. This research contributes to the advancement of healthcare analytics, providing a unique and holistic approach to disease prediction that has the capacity to transform patient care on a global scale[8,9,10].

## II. BACKGROUND

The field of healthcare has witnessed significant advancements in recent years, thanks to the emergence of machine learning techniques. With the growing availability of health data and the increasing computing power, machine learning has become a powerful tool in predicting and diagnosing various diseases[11,12,13].

The benefits of employing machine learning for multiple disease prediction are numerous. Firstly, it enables healthcare professionals to identify individuals who are at a higher risk of developing multiple diseases, facilitating early intervention and preventive measures. Secondly, it aids in optimizing healthcare resource allocation by prioritizing high-risk patients and ensuring timely interventions. Furthermore, machine learning algorithms can assist in the identification of disease patterns and risk factors, contributing to the development of targeted public health strategies[14].

While significant progress has been made in the field of multiple disease prediction using machine learning, there are still challenges that need to be addressed. These include the availability and quality of health data, ensuring patient privacy and data security, and the interpretability and explain ability of the predictive models. Additionally, the integration of machine learning algorithms into existing healthcare systems requires careful consideration of regulatory frameworks, ethical guidelines, and healthcare workflows[15].

In conclusion, the application of machine learning in multiple disease prediction holds immense potential for revolutionizing healthcare. By harnessing the power of these algorithms, healthcare providers can proactively identify individuals at risk, enhance diagnosis accuracy, and optimize treatment strategies. However, it is crucial to address the challenges associated with data quality, privacy, interpretability, and regulatory compliance to ensure the successful implementation of machine learning-based predictive models in healthcare settings.

There are several tools and technologies used which have been used to develop this project[16].

***Tools Used*:**

1: Kaggle - Kaggle is a platform that provides access to diverse datasets

2: Google Colaboratory - Colaboratory is a data analysis and machine learning tool

3: Anaconda – Anaconda aims to simplify package management and deployment.

4: Spyder IDE - An open-source cross-platform integrated development environment

6: Streamlit Cloud - Deploy, manage, share your apps with the world, directly from Streamlit

***Technologies Used*:**

1: Python - Python is dynamically typed, high-level, general-purpose programming language.

2: NumPy - A library for the Python. adding support for large, multi-dimensional arrays & matrices

3: Pandas – A software library written for the Python programming language for data manipulation and analysis.

4: Sklearn - A free software machine learning library for the Python programming language.

5: Machine Learning Algorithms - Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output.

6: Pickle - Python pickle module is used for serializing and de-serializing a Python object structure.

7: Stream Lit - A free, Open-source framework to rapidly build and share beautiful machine learning web apps.

## III. SYSTEM ANALYSIS

***3.1 Functional Requirement***

- The system allows the patient to predict the different diseases.
- The user adds disease-specific strategies and training models based on the user's strategy are published.

***3.2 Non-Functional Requirements***

- The website will provide many benefits during the forecast period.
- The Website must be reliable and consistent.

## IV. SYSTEM MODEL

For developing this project, we have used Supervised Machine Learning Model. Supervised Learning is the simplest machine learning model to understand in which input data is called training data and has a known label or result as an output. So, it works on the principle of input-output pairs. It requires creating a function that can be trained using a training data set, and then it is applied to unknown data and makes some predictive performance. Supervised learning is task-based and tested on labelled data sets[17].

We can implement a supervised learning model on simple real-life problems. We have employed machine learning models to predict the likelihood of different diseases based on user-input symptoms. To ensure accurate predictions, we selected different machine learning algorithms for each disease, considering their accuracy performance[18].

For each disease, we carefully chose a specific machine learning algorithm that best captures the patterns and relationships between symptoms and diseases. These algorithms include logistic regression, support vector machines (SVM) & K-Nearest Neighbours (KNN). The selection was based on their ability to effectively analyse the dataset and provide accurate predictions[19].

By leveraging different machine learning algorithms for different diseases and selecting the most accurate models, our project aims to provide reliable disease predictions, supporting healthcare professionals and users in early detection and intervention.

## V. EXPERIMENT

***5.1 Hypothesis Generation***

The hypothesis for the Multiple Disease Prediction System is that by analysing general medical data and advanced machine learning algorithms, it is possible to accurately predict the likelihood of individuals acquiring specific diseases. The hypothesis assumes that there are underlying patterns and relationships within the medical data that can be leveraged to develop a robust predictive model.

***5.2 Collection of Data***

To initiate this project, we began by collecting data from various sources. We utilized Kaggle as a platform to import relevant datasets, which serve as valuable resources for practice, research, and as a foundation for constructing machine learning models. These curated datasets provide a solid starting point, offering a diverse range of information that can be leveraged to train and validate our prediction system accurately.

***5.3 Data Pre- Processing / Removal of Unwanted Data***

The collected data serves various purposes, and as it is sourced from diverse platforms, it can contain a substantial amount of information. However, this imported data may also include unwanted or noisy elements that require pre-processing. The primary objective of data pre-processing is to refine the dataset by removing irrelevant or redundant information, addressing missing values, and handling outliers or noise. This step ensures that only the necessary and high-quality data is retained for further analysis[20].

***5.4 Feature Selection***

Feature selection is a critical step in the data analysis process, aimed at identifying and selecting the most relevant and informative features from a dataset. With the abundance of available data, feature selection plays a crucial role in enhancing the performance and efficiency of machine learning models. We have used statistical measures and correlation analysis techniques to assess the importance and relevance of each feature[21].

By performing feature selection, we aim to streamline the input data and retain only the most valuable features for our predictive models. This process helps us focus on the most influential factors and ensures that our model's predictions are based on the most meaningful and impactful variables. Ultimately, feature selection enables us to improve the accuracy and efficiency of our multiple disease prediction system, contributing to better healthcare outcomes and empowering medical professionals with valuable insights[22].

***5.5 Model Building***

For our multiple disease prediction system, we have utilized supervised machine learning algorithms, namely Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN). These algorithms have been chosen for their effectiveness in classification tasks and their ability to handle multi-class prediction scenarios[23].

By leveraging these supervised machine learning algorithms, we aim to develop robust and accurate models for multiple disease prediction. Each algorithm brings its unique strengths and considerations, allowing us to explore different approaches and select the most suitable model for each disease category. Through this model-building process, we aim to provide reliable and timely predictions[24].

***5.6 Deployment***

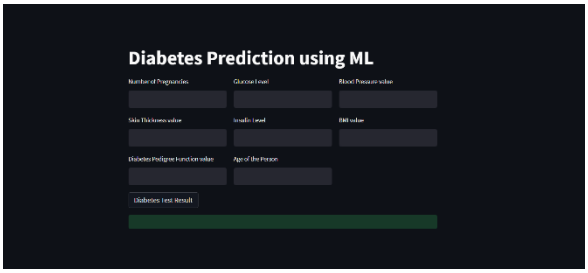
Our multiple disease prediction system has been deployed using the Streamlit Cloud server, providing a user-friendly web interface for easy access and interaction. With Streamlit Cloud, users can input disease parameters and receive accurate predictions for multiple diseases simultaneously. This deployment ensures accessibility and scalability, empowering healthcare professionals and individuals to make informed decisions about their health.

# VI. DESIGN

### *6.1 Architecture Design*

In Figure no 6.1 we have experimented on five different diseases that is Heart, Diabetes and Lung Cancer, Parkinson's and Thyroid as these are correlated to each other. The first step is to extract the dataset. we have imported the dataset from Kaggle respectively. When we import the dataset, all the input data will be visible. After previewing the data, we check the results and missing results, evaluate the data over the new dataset, and separate the data as training and testing. Next, we used various machine learning algorithms on the training map and used the test dataset to use the information for the classification algorithm, we applied different machine learning algorithms and applied knowledge on the classified algorithm using the testing dataset. After applying knowledge, we have chosen the Logistic Regression, SVM, and KNN algorithm with the best accuracy for each of the diseases. we will create a pickle file for each disease predictive model and then combine it with the Streamlit framework for the production of web templates.

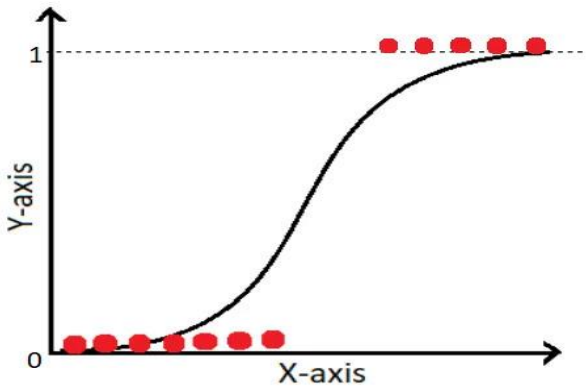
### *6.2 Architectural Design Interface*




# VII. PRELIMINARIES

### *7.1 Machine Learning Algorithm*

### *7.1.1 Logistic Regression Algorithm*




A statistical model is typically used to model a binary dependent variable with the help of a logistic function. Another name for the logistic function is a sigmoid function and is given by:

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

This function assists the logistic regression model to squeeze the values from (-k,k) to (0,1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

Logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task.

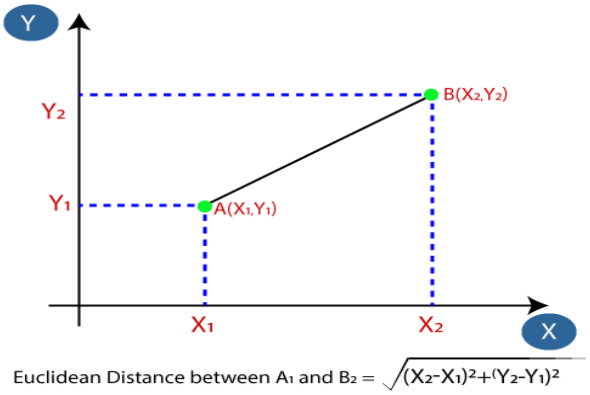
### *7.1.2 Support Vector Machine Algortihm*

The Support Vector Machine (SVM) is a supervised machine learning algorithm widely used to solve the binary classification problem. It can also be used for the multiset classification problems and regression problems.

Let's say we have n data points, each observation i has a feature p (for example x_i has length p) and y_i = -1 or y_i = 1 inch two classes. Suppose we have two separate classes of observations.

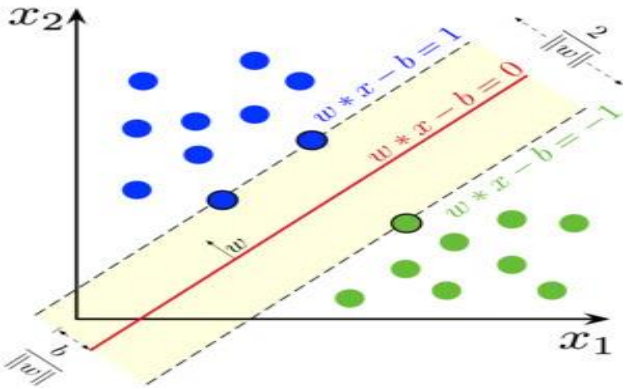
This means that we can draw a general plane from our private space, with all instances of one class on one side of the plane and all instances of the other class on the other. (the hyperplane of p dimensions is a p-1 dimensional subspace. In the 2d example below, the hyperplane is just a line.)

We define the hyperplane as follows:




$$x \cdot \bar{w} + \bar{b} = 0$$

w is a vector p and b is a real number. For simplicity we need ~w = 1, so the quantity x * ~w + ~b is the distance from the point x to the plane.




So we can write our class as y = +1/-1 and the general plane's class division requirement would be::

$$yi(x_i \cdot w' + b') \geq 0$$

### *7.1.3 K-NN Algorithm*

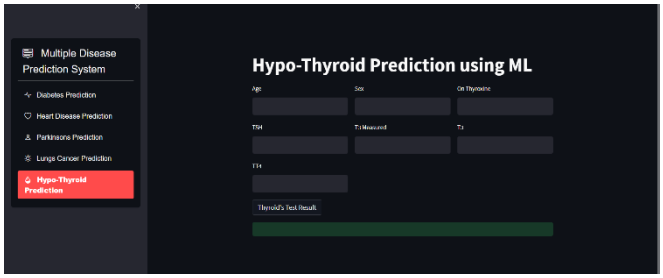
The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbour
- **Step-2:** Calculate the Euclidean distance of K number of neighbours
- **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance
- **Step-4:** Among these k neighbours, count the number of the data points in each category
- **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum
- **Step-6:** Our model is ready.

- Firstly, we will choose the number of neighbours, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry.
- By calculating the Euclidean distance we got the nearest neighbours, as three nearest neighbours in category A and two nearest neighbours in category B.
- As we can see the 3 nearest neighbours are from category A, hence this new data point must belong to category A.

## VIII. RESULT

In our system, The Diabetes Disease prediction and Parkinson's Disease prediction model utilize the Support Vector Machine (SVM) algorithm, while the heart disease prediction and Lung Cancer Prediction model employs the Logistic Regression algorithm. For the Thyroid disease prediction model, we have utilized the K-NN algorithm, as these algorithms have demonstrated the best accuracy for their respective diseases.




When a patient enters the relevant parameters based on the selected disease, the system will determine whether the patient is likely to have the disease or not. The system guides by indicating the expected range of values for each parameter. If a value is outside the specified range, invalid, or left empty, the system will display a warning sign, prompting the patient to input a correct and valid value.

By utilizing these specific algorithms and providing clear parameter requirements, we aim to enhance the accuracy and reliability of our disease prediction system. This approach empowers users to receive timely and accurate predictions while ensuring that the input data meets the necessary criteria for analysis.

## IX. CONCLUSION

The primary aim of this project was to develop a system capable of accurately predicting multiple diseases. By achieving this objective, we have eliminated the need for users to visit multiple websites, saving them valuable time. Timely disease prediction can significantly increase life expectancy and prevent financial burdens. To accomplish this, we utilized several machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN), in order to achieve the highest possible accuracy.

By harnessing the power of these algorithms, our system can provide accurate disease predictions, empowering individuals to take proactive measures for their health. Early detection and intervention are crucial in managing and treating various diseases effectively. Through this project, we have created a valuable tool that can contribute to improved healthcare outcomes and overall well-being.

In conclusion, our project represents a significant step towards revolutionizing healthcare by leveraging machine learning techniques to predict multiple diseases accurately. The system's ability to provide efficient and accurate predictions has the potential to enhance healthcare access, improve patient outcomes, and ultimately save lives.

## FUTURE SCOPE

There are several avenues for future development and expansion of our multiple disease prediction system:

- Addition of More Diseases: In the future, we can expand the system by incorporating additional diseases into the existing web application. This would enable users to predict a broader range of diseases and further enhance the system's usefulness in healthcare.
- Accuracy Improvement: As part of ongoing research and development, we can strive to improve the accuracy of disease predictions. By refining the machine learning algorithms, optimizing feature selection, and incorporating more comprehensive datasets, we can reduce false predictions and increase the overall accuracy of the system. This would ultimately contribute to lowering the mortality rate by enabling timely interventions and treatments.
- Integrating our multiple disease prediction system with electronic health records can provide a more comprehensive and personalized healthcare experience. By leveraging patient data from EHR systems, we can enhance the accuracy of predictions and enable healthcare professionals to make informed decisions based on the patient's medical history.
- Mobile Application Development: Developing a mobile application version of the multiple disease prediction system would enhance accessibility and convenience for users. It would allow individuals to access the system on their smartphones, providing real-time disease predictions and empowering them to take proactive measures for their health anytime, anywhere.

By focusing on these future directions, we can further advance the field of disease prediction, improve healthcare outcomes, and make a positive impact on individuals' lives.

## REFERENCES

[1] Gopisetti, L. D., Kummera, S. K. L., Pattamsetti, S. R., Kuna, S., Parsi, N. & Kodali, H. P. (2023, January). Multiple disease prediction system using machine learning and streamlit. In: *5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 923-931. IEEE.

[2] Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale, N. (2020). *Disease prediction from various symptoms using machine learning*. Available at: SSRN 3661426.

[3] Srivastava, S., Haroon, M. & Bajaj, A. (2013, September). Web document information extraction using class attribute approach. In: *4th International Conference on Computer and Communication Technology (ICCCT),* pp. 17-22. IEEE.

[4] Raja, M. S., Anurag, M., Reddy, C. P. & Sirisala, N. R. (2021, January). Machine learning based heart disease prediction system. In *2021 International Conference on Computer Communication and Informatics (ICCCI),* pp. 1-5. IEEE.

[5] Khan, W. & Haroon, M. (2022). An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks. *International Journal of Cognitive Computing in Engineering*, *3*, 153-160.

[6] Khan, R., Haroon, M. & Husain, M. S. (2015, April). Different technique of load balancing in distributed system: A review paper. In *2015 Global Conference on Communication Technologies (GCCT),* pp. 371-375. IEEE.

[7] Haroon, M. & Husain, M. (2015, March). Interest Attentive Dynamic Load Balancing in distributed systems. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1116-1120. IEEE.

[8] Husain, M. S. & Haroon, D. M. (2020). An enriched information security framework from various attacks in the IoT. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*.

[9] Haroon, M. & Husain, M. (2013). Analysis of a dynamic load balancing in multiprocessor system. *International Journal of Computer Science engineering and Information Technology Research*, *3*(1).

[10] Khan, W. (2021). An exhaustive review on state-of-the-art techniques for anomaly detection on attributed networks. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 6707-6722.

[11] Haroon, M. & Husain, M. (2013). Different types of systems model for dynamic load balancing. *IJERT*, *2*(3).

[12] Siddiqui, Z. A. & Haroon, M. (2023). Research on significant factors affecting adoption of blockchain technology for enterprise distributed applications based on integrated MCDM FCEM-MULTIMOORA-FG method. *Engineering Applications of Artificial Intelligence*, *118*, 105699.

[13] Khan, N. & Haroon, M. (2022). *Comparative study of various crowd detection and classification methods for safety control system*. Available at: SSRN 4146666.

[14] Siddiqui, Z. A. & Haroon, M. (2022). Application of artificial intelligence and machine learning in blockchain technology. In *Artificial Intelligence and Machine Learning for EDGE Computing,* pp. 169-185. Academic Press.

[15] Tripathi, M. M., Haroon, M., Khan, Z. & Husain, M. S. (2022). Security in digital

healthcare system. *pervasive healthcare: a compendium of critical factors for success*, 217-231.

[16] Shrestha, R. & Chatterjee, J. M. (2019). Heart disease prediction system using machine learning. *LBEF Research Journal of Science, Technology and Management*, *1*(2).

[17] Phasinam, K., Mondal, T., Novaliendry, D., Yang, C. H., Dutta, C. & Shabaz, M. (2022). Analyzing the performance of machine learning techniques in disease prediction. *Journal of Food Quality*.

[18] Lanjewar, M. G. & Panchbhai, K. G. (2023). Convolutional neural network based tea leaf disease prediction system on smart phone using paas cloud. *Neural Computing and Applications*, *35*(3), 2755-2771.

[19] Ampavathi, A. & Saradhi, T. V. (2021). Multi disease-prediction framework using hybrid deep learning: an optimal prediction model. *Computer Methods in Biomechanics and Biomedical Engineering*, *24*(10), 1146-1168.

[20] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C. & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*.

[21] Vijayalaxmi, A., Sridevi, S., Sridhar, N. & Ambesange, S. (2020, May). Multi-disease prediction with artificial intelligence from core health parameters measured through non-invasive technique. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS),* pp. 1252-1258. IEEE.

[22] Aldahiri, A., Alrashed, B. & Hussain, W. (2021). Trends in using IoT with machine learning in health prediction system. *Forecasting*, *3*(1), 181-206.

[23] Haroon, M., Tripathi, M. M. & Ahmad, F. (2020). Application of machine learning in forensic science. In: *Critical Concepts, Standards, and Techniques in Cyber Forensics*, pp. 228-239. IGI Global.

[24] Priyanka Sonar & Prof. K. JayaMalini. (2019). Diabetes prediction using different machine learning approaches. *IEEE 3rd International Conference on Computing Methodologies and Communication (ICCMC)*.