



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Symptosense Prognosis

Veena Potdar*, Lavanya Santhosh**, Chaitra M***, Chaithra N***,

Kousalya G M*** & Ankitha V*** Students, Department of CSE, Dr Ambedkar Institute of Technology

***Associate Professor, Department of CSE, Dr Ambedkar Institute of Technology

Abstract— There is a need to study and make a system which will make it easy for an end-user to predict the permanent diseases without visiting a physician or doctor for a diagnosis. To detect the various diseases through the examining symptoms of patient's using various methods of machine learning models. To manage text data and structured data is no proper method. The recommended system will examine both structure and unstructured data. The predictions accuracy will improve using machine learning.

Keywords—Diseases, Prediction, Machine learning, Decision tree, Random Forest, K-Nearest neighbor algorithm.

I. INTRODUCTION

With the rise in number of patient and disease every year medical system is overloaded and with time have become overpriced in many countries. Most of the disease involves a consultation with doctors to get treated. With sufficient data prediction of disease by an algorithm can be very easy and cheap. Prediction of disease by looking at the symptoms is an integral part of treatment. Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Healthcare is the most prime example of how machine learning is used in the medical field. We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyzed data and deliver results faster, with the use of machine learning technology doctors can make a good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare service.

II EXISTING SYSTEM

The existing system predicts the chronic diseases which are for a particular region and for the particular community. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithm is used for Disease risk prediction. The system is using Machine Learning algorithm i.e. K Nearest Neighbours, Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%. In the existing paper, they streamline machine learning algorithms for the effective prediction of chronic disease outbreak in disease-frequent communities. They experiment with the modified prediction models over real-life hospital data

collected from central China. They propose a conventional neural network-based multi modal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from the hospital.

III PROPOSED SYSTEM

This system is used to predict disease according to symptoms. This system uses decision tree classifier, Random Forest algorithm and KNN algorithm for evaluating the model. This system is used by end-users. The system will predict disease based on symptoms. This system uses Machine Learning Technology. We have named this system as "Symptoms Based Disease Prediction". This system is for those people who are always fretting about their health, for this reason, we provide some features which acknowledge them and enhance their mood too. So, there is a feature for the awareness of health, which recognize disease according to symptoms.

III SCOPE OF THE PROJECT

In this work, our goal is to provide a tool to assist professionals and consumers in finding and choosing disease. To achieve this goal, we develop an approach that allows a user to query for disease that satisfy a set of conditions based on symptoms properties, such as disease indications.

IV PROBLEM STATEMENT

The Symptoms Based Disease prediction is focused for optimally reducing the healthcare costs. There are several functionalities remain untouched into health prediction system. So, by living in the edge of technology and still if we are not able to utilize it in efficient and proper manner then there is no use of it. To tackle this, research is carried out in health prediction system. There are several applications which use any one of the technologies. This project shows the merging of both technologies to achieve efficient result.

V LITERATURE SURVEY

Emile, Sameh Hany, Waleed Ghareeb, Hossam Elfeki, Et.al [1], May 2022, Prediction of the onset of de novo gastroesophageal reflux disease (GERD) after sleeve gastrectomy (SG) would be helpful in decision-making and selection of the optimal bariatric procedure for every patient. The present study aimed to develop an artificial intelligence (AI)-based model to predict the onset of GERD after SG to help clinicians and surgeons in decision-making.

yu zhou, Tong Mu, Xiaochuan Kong, Le Zhang, Et.al[2],2021,. Current diagnostic methods, such as X-ray, computed tomography, and magnetic resonance imaging, suffer from precision errors and low sensitivity. Machine learning (ML) holds promise as a clinical diagnostic tool by utilizing probabilistic algorithms to make predictions. This systematic review aims to provide a comprehensive synthesis of machine learning applications for knee osteoarthritis diagnosis, offering insights into accuracy and performance metrics.

Marion Miguères, Sébastien Lhomme, and Jacques Izopet , Et.al[3] ,September 2021,The hepatitis A virus (HAV) belongs to the Hepatoviral genus within the Picornaviridae family. There are two types of infectious HAV particles: naked and quasi-enveloped virions. Quasi-enveloped virions have a lipid membrane and are found in the blood and culture supernatants Extra-hepatic manifestations of acute hepatitis A are rare but can include neurological symptoms such as Guillain–Barre syndrome, rash, pancreatitis, arthritis, myocarditis, acute kidney injury, and haematological disorders such as haemolysis and cryoglobulinemia. Algorithms are KNN, Random Forest.

Muhammad Bilal, Majed Alfayad, Shazia Saqib, Et.al [4], December 2021, the study focuses on diagnosing the stage of Hepatitis C using an Artificial Back-Propagation Neural Network. Multiple sensors continuously gather data from the environment, converting physical quantities into measurements. These sensors are connected to a sensor board through various topologies. The research reveals that Random Forest (54.56%, using Python) and KNN (51.06%, using R) demonstrated the highest precision in both binary class and multiclass.

George Obaido , Blessing Ogbuokiri ,Theo G. Swart , Et.al[5], November 2022.Hepatitis B is a potentially life-threatening liver infection caused by the hepatitis B virus. Various algorithms, such as decision tree, logistic regression, support vector machines, random forest, adaptive boosting (AdaBoost), and extreme gradient

boosting (XGBoost), were employed in constructing the models. These models achieved balanced accuracies of 75%, 82%, 75%, 86%, 92%, and 90%, respectively. This information underscores the importance of diverse modelling approaches in understanding and predicting Hepatitis B, while also emphasizing the need for improved interpretability in existing research.

Pietro Lampertico, Elisabetta Degaspero, Lisa Sandmann, Heiner Wedemeyer, Et.al [6], June 2023, Hepatitis D virus infection: Pathophysiology, epidemiology, and treatment. Chronic infection with the hepatitis delta virus (HDV) affects between 12-20 million people globally and represents the most severe form of viral hepatitis. This highlights the ongoing need for research to deepen our understanding of HDV and its impact on individuals, emphasizing areas where further investigation is required.

Aversano, Lerina, Mario Luca Bernardi, Et.al[7],2021., they presented a machine learning decision support system for endocrinologists treating thyroid disease. The model predicts treatment progress based on various patient-related parameters. Among the tested classifiers, the Extra-Tree Classifier stood out with an impressive 84% accuracy, showcasing the potential utility of machine learning in optimizing thyroid disease management.

KM Jyoti Rani's, Et.al [8],2020., project focuses on enhancing the accuracy of early diabetes prediction by combining results from different machine learning techniques. The study evaluates algorithms like K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree, ultimately selecting the model with the highest accuracy for predicting diabetes. This research highlights the potential of diverse machine learning approaches in improving diabetes prediction.

Wenjie Fang, Junqi Wu, Mingrong Cheng, XinlinZhu , Et.al[9] 2023 The prevalence of invasive fungal infections (IFIs), notably *Histoplasma capsulatum*, has risen due to an increased number of immunocompromised patients. This review sheds light on advancements in diagnostic tools, emphasizing nonculture methods, to adapt to evolving diagnostic landscapes in diverse settings. Artificial intelligence (AI), encompassing learning, reasoning, and self-correction, is employed in this context, reflecting the integration of machine-based processes to simulate human intelligence.

Patrizia Ferroni, Fabio M. Zanzotto, Et.al [10], 2020, Machine learning (ML) is extensively applied to create automated predictors for migraine classification, but the development of such predictors for medication overuse (MO) in migraine is still in its early stages. Support Vector Machines and Random Optimization (RO-MO) are utilized in this context. Logistic regression analysis has affirmed the efficacy of the derived RO-MO system. Migraine represents a primary type of headache marked by recurring episodes of severe headache, often preceded by transient neurological symptoms known as aura.

Leon Tsung-Ju Lee1, Hsuan-Chia Yang1,et.al.[11],March 2023, along with their colleagues, aimed to develop and validate a prediction model for Psoriatic Arthritis (PsA) using large-scale and multidimensional electronic medical records. The study utilized a machine learning algorithm and

focused on 2.5 years of diagnostic and medical records (both inpatient and outpatient). This temporal-sequential information was employed to predict the risk of PsA for a given patient within the next 6 months. The model underwent development and cross-validation using training data, and testing was performed using hold out data. An occlusion sensitivity analysis helped identify the important features of the model.

Marion R. Sills, Mustafa Ozkaynak, Hoon Jang, Et.al [12], July 2021, this study analyzed electronic health record data from five EDs within a single healthcare system, which included an academic urban children's hospital ED. The study employed an automated machine learning algorithm (autoML), employing Random Forest and logistic regression models.

Marie Binignat, Valentina Padoia, Atul J Butte, Karine Louati, Et.al [13] 2021, The objective of this systematic literature review was to offer a thorough overview of the application of machine learning (ML) in the clinical management of osteoarthritis (OA). The review, conducted in July 2021 using MEDLINE PubMed, involved the analysis of articles using key words and MeSH terms. Information such as the number of patients, ML algorithms utilized, types of analyzed data, validation methods, and data availability was collected for each selected article.

A. Saran Kumar, Dr. R. Rekha, et.al. [14], September 2021, Adverse drug reactions (ADR) are the major source of medical issues and adverse drug reactions are determined to be the major cause of death in many countries ahead of pulmonary disease, diabetes, AIDS and pneumonia. Estimates of the number of patients suffered due to drug-drug interactions vary from 5-8% of all complication errors within healthcare centers. Machine learning algorithms provide a set of methods which can increase the accuracy and success rate for well-defined issues with abundant data

Amaratunga, Dhammika, Javier Cabrera, Et.al [15], June 2020, Artificial intelligence (AI) shows promise in providing valuable insights to hypertension specialists, with significant applications already developed using large, validated datasets. Deep learning and support vector machine (SVM) models have achieved 56% to 57% accuracy in predicting cardiovascular events. A neural network algorithm successfully predicted cardiovascular events over a 10-year follow-up period, achieving a sensitivity of 68% and specificity of 71%.

Nathalie Mentha, Sophie Clément, Francesco Negro, Dulce Alfaiate, et.al. [16], May 2019, Hepatitis delta virus (HDV) is a defective virus that requires the hepatitis B virus (HBV) to complete its life cycle in human hepatocytes. HDV virions contain an envelope incorporating HBV surface antigen protein and a ribonucleoprotein containing the viral circular single-stranded RNA genome associated with both forms of hepatitis delta antigen, the only viral encoded protein. Replication is mediated by the host cell DNA-dependent RNA polymerases.

Marion Miguères, Sébastien Lhomme, and Jacques Izopet, Et.al [17], September 2021, The hepatitis A virus (HAV) belongs to the Hepatovirus genus within the Picornaviridae family. There are two types of infectious HAV particles: naked and quasi-enveloped virions. Quasi-enveloped virions have a lipid membrane and are found in the blood and culture supernatants. Extra-hepatic manifestations of acute hepatitis A are rare but can include neurological

symptoms such as Guillain-Barré syndrome, rash, pancreatitis, arthritis, myocarditis, acute kidney injury, and haematological disorders such as haemolysis and cryoglobulinemia. Algorithms are KNN, Random Forest.

Muhammad Bilal, Majed Alfayad, Shazia Saqib, Et.al [18], December 2021, the study focuses on diagnosing the stage of Hepatitis C using an Artificial Back-Propagation Neural Network. Multiple sensors continuously gather data from the environment, converting physical quantities into measurements. These sensors are connected to a sensor board through various topologies. The research reveals that Random Forest (54.56%, using Python) and KNN (51.06%, using R) demonstrated the highest precision in both binary class and multiclass.

Uddin, Shahadat, and Arif Khan's, et.al. [19], 2019., study explored machine learning algorithms for disease prediction using medical data. Among the commonly used algorithms, Support Vector Machine and Naïve Bayes were prevalent, but Random Forest exhibited superior accuracy. This research provides insights into the effectiveness of different machine learning approaches in medical data analysis for disease prediction.

Rakesh Gupta a, R.K. Sanjeev b, A. Agarwal c, et.al. [20], July 2019, Hepatitis A virus seropositivity among children aged between 1 and 5 years of age. This indicates causes an enterically transmitted viral disease mainly affecting children and endemic in many developing countries, including India. This study was conducted to assess the seroprevalence of HAV among young children aged below 5 years and the need for universal immunization. The seroprevalence of HAV was lower among younger children aged 12–23 months (odds ratio [OR] = 0.73, 95% confidence interval [CI] = 0.52–0.87, $p = 0.03$), which was statistically significant.

VI SYSTEM REQUIREMENT SPECIFICATION

A Requirement Specification, often referred to as a software requirements specification (SRS) or system requirements specification, is a comprehensive document that outlines the functional and non-functional requirements of a system or software application. It serves as a blueprint for the development team, guiding them in designing, implementing, and testing the system. The requirement specification document captures the client's or stakeholders' needs and expectations, providing a clear understanding of what the system is supposed to accomplish.

Functional Requirements

Functional requirements describe the specific functionalities or features that a system or software must provide. They are concerned with what the system should do, the actions it should perform, and the services it should offer to users.

1. User name:

User should be able to fill their name.

2. Symptom Input:

- Users should be able to input their symptoms.
- The system will support input with Drop Down format.

3. Symptom Analysis:

- The system should analyse the input symptoms using algorithms.
- The analysis should consider the correlation between symptoms and potential diseases.

4. Disease Prediction

- The system will provide a list of potential diseases based on the symptom analysis.
- And with help of their background details, we can confirm that they got the such type of diseases

5. Data Visualization:

- Display visual representations like bar graphs.
- Visual aids can help users better understand the information.

6. Notification System:

- Implement a notification system to alert users if they did not fill anything field.

7. Suggestion box:

- Giving remedies to a particular disease.

Non-Functional Requirements

Non-functional requirements, on the other hand, define the quality attributes, constraints, and characteristics that the system should exhibit. They are concerned with how well the system performs its functions rather than what functions it performs.

1. Performance:

- The system should provide accurate prediction within a reasonable time frame.

2. Scalability:

- Scalability will be considered for the computational resources.

3. User Interface Design:

- It plays a main role, where are system will be user-friendly.

4. Interoperability:

- Where our system can be used in any platform.

VII METHODOLOGY AND SYSTEM DESIGN

The whole work of this dissertation is categorized into three major parts or stages as given below:

1. Dataset Preparation
2. Data preprocessing
3. Algorithms: Decision Trees, Random Forest, K Nearest Neighbour
4. Input: from the user (Symptoms)
5. Output: prediction of disease

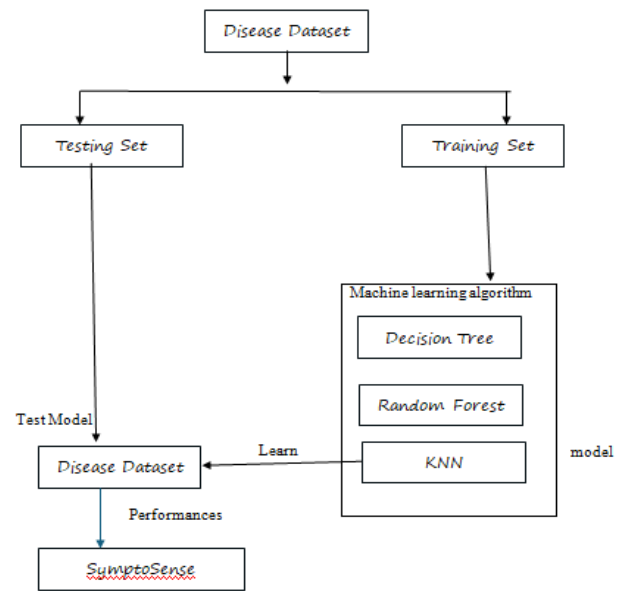


Figure 1 : System architecture for SymptoSense Prognosis

In this phase, load the dataset into program and extract the data from the .csv file. This data can be analyzed and extract the best features to preprocess the data. For the given data set, there are quite a few 'NA' values which are filtered in python. Furthermore, as the data set consists of numeric data, we used robust scaling, which is quite similar to normalization, but it instead uses the interquartile range whereas normalization is something which normalization shrinks the data in terms of 0 to 1. And different machine learning algorithms are been used to predict the disease.

VIII WORKING FLOW DIAGRAM

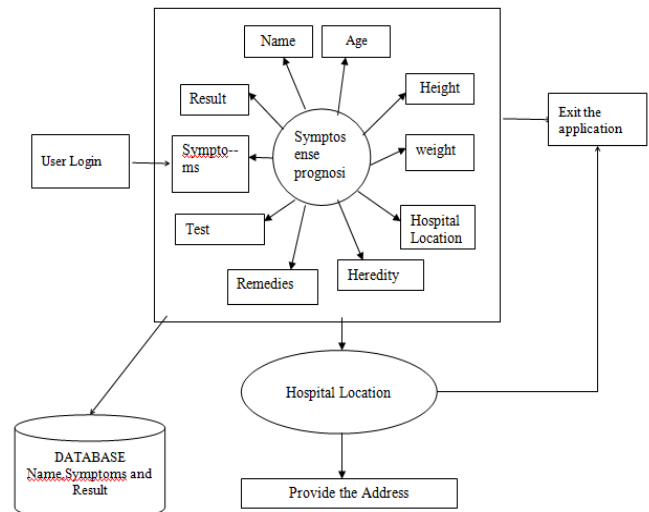


Figure 2: Complete working Flow of the system

This is the initial screen where the user enters their credentials to access the login. Once logged in, the user is directed to the personal information input section, where they provide their name, age, height, and weight. Next, the app allows the user to input their hospital location, heredity information, remedies, and test results. Users can also track their symptoms, and the app will display the corresponding results. Additionally, the app provides the hospital's address based on the location entered. It integrates all this information into a database, including symptoms and results. Finally, users can choose to exit the application.

CONCLUSION

In our research, we systematically reviewed 40 papers aligned with the objectives of our project. Each paper employed distinct machine learning algorithms, including decision tree, random forest, and KNN, to forecast diseases using specific datasets tailored to particular health conditions. The diverse methodologies and algorithms utilized in these studies offered us valuable insights into the landscape of disease prediction. Inspired by this comprehensive review, we initiated the implementation of our project, SymptoSense Prognosis.

In SymptoSense Prognosis, we leverage the collective knowledge gained from the research papers to predict the initial stages of various diseases. Our approach involves integrating user-provided symptoms into the predictive model. By doing so, we aim to create a robust system capable of offering early-stage predictions for multiple diseases. The amalgamation of machine learning techniques and user-generated symptom data forms the foundation of our innovative project, which seeks to enhance early detection and prognosis in the medical domain.

REFERENCES

[1] Emile, Sameh Hany, Waleed Ghareeb, Hossam Elfeki, Mohamed El Sorogy, Amgad Fouad, and Mohamed Elrefai. "Development and validation of an artificial intelligence-based model to predict gastroesophageal reflux disease after sleeve gastrectomy." *Obesity Surgery* 32, no. 8 (2022): 2537-2547. DOI: <https://link.springer.com/article/10.1007/s11695-022-06112-x>

[2] Mu, Tong, Xiaochuan Kong, and Le Zhang. "Diagnose and Prediction of Different Machine Learning methods for Knee Osteoarthritis: a protocol for systematic review and meta-Analysis." (2021). DOI: <https://doi.org/10.21203/rs.3.rs-903246/v1>

[3] Miguères, Marion, Sébastien Lhomme, and Jacques Izopet. "Hepatitis A: epidemiology, high-risk groups, prevention and research on antiviral treatment." *Viruses* 13, no. 10 (2021): 1900. DOI: <https://doi.org/10.3390/v13101900>

[4] Butt, Muhammad Bilal, Majed Alfayad, Shazia Saqib, M. A. Khan, Munir Ahmad, Muhammad Adnan Khan, and Nouh Sabri Elmitwally. "Diagnosing the stage of hepatitis C using machine learning." *Journal of Healthcare Engineering* 2021 (2021). DOI: <https://doi.org/10.1155/2021/8062410>

[5] Obaïdo, George, Blessing Ogbuokiri, Theo G. Swart, Nimibofa Ayawei, Sydney Mambwe Kasongo, Kehinde Aruleba, Ibomoiye Domor Mienye et al. "An interpretable machine learning approach for hepatitis b diagnosis." *Applied sciences* 12, no. 21 (2022): 11127. DOI: <https://doi.org/10.3390/app12211127>

[6] Lampertico, Pietro, Elisabetta Degasper, Lisa Sandmann, Heiner Wedemeyer, Cihan Yurdaydin, Dominique Roulot, Fabien Zoulim et al. "Hepatitis D virus infection: Pathophysiology, epidemiology and treatment. Report from the first international delta cure meeting 2022." *JHEP Reports* (2023): 100818. DOI: <https://doi.org/10.1016/j.jhepr.2023.100818>

[7] Aversano, Lerina, Mario Luca Bernardi, Marta Cimitile, Martina Iammarino, Paolo Emidio Macchia,

Immacolata Cristina Nettore, and Chiara Verdone. "Thyroid disease treatment prediction with machine learning approaches." *Procedia Computer Science* 192 (2021): 1031-1040. DOI: <https://doi.org/10.1016/j.procs.2021.08.106>

[8] KM Jyoti Ranii. "Diabetes Prediction Using Machine Learning". No. 294-305., 2020. DOI: <https://doi.org/10.32628/CSEIT206463>

[9] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." *Procedia Computer Science* 167 (2020): 706-716. DOI: <https://doi.org/10.1016/j.procs.2020.03.336>

[10] Fang, Wenjie, Junqi Wu, Mingrong Cheng, Xinlin Zhu, Mingwei Du, Chang Chen, Wanqing Liao, Kangkang Zhi, and Weihua Pan. "Diagnosis of invasive fungal infections: challenges and recent developments." *Journal of Biomedical Science* 30, no. 1 (2023): 42. DOI: <https://doi.org/10.1186/s12929-023-00926-2>

[11] Leon Tsung-Ju Lee, Hsuan-Chia Yang, Phung Anh Nguyen, Muhammad Solihuddin Muhtar, BA; Yu-Chuan Jack Li "Machine Learning Approaches for Predicting Psoriatic Arthritis Risk Using Electronic Medical Records: Population-Based Study", 2023, 1-8, DOI: 10.2196/39972

[12] Fang, Wenjie, Junqi Wu, Mingrong Cheng, Xinlin Zhu, Mingwei Du, Chang Chen, Wanqing Liao, Kangkang Zhi, and Weihua Pan. "Diagnosis of invasive fungal infections: challenges and recent developments." *Journal of Biomedical Science* 30, no. 1 (2023): 42. DOI: <https://doi.org/10.1186/s12929-023-00926-2>

[13] Sills, Marion R., Mustafa Ozkaynak, and Hoon Jang. "Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning." *International Journal of Medical Informatics* 151 (2021): 104468. DOI: <https://doi.org/10.1016/j.ijmedinf.2021.104468>

[14] A. Saran Kumar, Dr. R. Rekha "A Critical Review On Predicting Drug-Drug Reactions Using Machine Learning Techniques", v-7, 2021, 1-3. DOI: 10.19070/2572-7389-2100015

[15] Amaratunga, Dhammika, Javier Cabrera, Davit Sargsyan, John B. Kostis, Stavros Zinonos, and William J. Kostis. "Uses and opportunities for machine learning in hypertension research." *International Journal of Cardiology Hypertension* 5 (2020): 100027 DOI: <https://doi.org/10.1016/j.ijchy.2020.100027>

[16] Mentha, Nathalie, Sophie Clément, Francesco Negro, and Dulce Alfaiate. "A review on hepatitis D: From virology to new therapies." *Journal of advanced research* 17 (2019): 3-15. DOI: <https://doi.org/10.1016/j.jare.2019.03.009>

[17] Miguères, Marion, Sébastien Lhomme, and Jacques Izopet. "Hepatitis A: epidemiology, high-risk groups, prevention and research on antiviral treatment." *Viruses* 13, no. 10 (2021): 1900. DOI: <https://doi.org/10.3390/v13101900>

[18] Butt, Muhammad Bilal, Majed Alfayad, Shazia Saqib, M. A. Khan, Munir Ahmad, Muhammad Adnan Khan, and Nouh Sabri Elmitwally. "Diagnosing the stage of hepatitis C using machine learning." *Journal of Healthcare Engineering* 2021 (2021). DOI: <https://doi.org/10.1155/2021/8062410>

[19] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." *BMC medical informatics and decision making* 19, no. 1 (2019): 1-16. DOI: <https://doi.org/10.1186/s12911-019-1004-8>

[20] Ullah, Rahat, Saranjam Khan, Hina Ali, Iqra Ishtiaq Chaudhary, Muhammad Bilal, and Iftikhar Ahmad. "A comparative study of machine learning classifiers for risk prediction of asthma disease." Photodiagnosis and Photodynamic Therapy 28 (2019): 292-296. Doi : <https://doi.org/10.1016/j.pdpdt.2019.10.011>

