iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53026

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Disease Prediction Using Machine Learning Algorithms

Harshit Kumar[1], Kapil Kumar[2], Ishan Sharma[3],Dr. Prabhat Kumar Srivastava[4]
[1, 2, 3, 4]*Department of computer science and Engineering IMS Engineering college, AKTU, Ghaziabad, Uttar Pradesh – 201015*

Abstract: ***The objective of this project is to develop a machine learning model that can predict the disease of a patient based on their symptoms. While data mining has been successfully applied in many areas, such as market analysis and e-commerce, the medical field still lacks powerful analytical tools to uncover hidden relationships and trends in data. Medical data contains a wealth of information, but this knowledge is often not effectively utilized. Machine learning is a field of study that involves developing algorithms that can improve automatically through experience and data. These algorithms use training data to build a model that can make predictions or decisions without being explicitly programmed. In this project, techniques such as association rule mining, classification, and clustering will be used to explore various general health problems. Classification is a crucial problem in data mining, and decision trees are a popular classifier used to create class models. The ID3 Decision Tree algorithm is commonly used for information classification. However, this algorithm can be inaccurate, so techniques such as entropy-based cross-validation and partitioning will be used to improve the accuracy of the model. Finally, the results will be compared to determine the best model. Introduction I would like to begin by highlighting the indispensability of computers in our lives. Computers are integral components in virtually every aspect of our lives today, comprising various hardware and software components. Software, which is a collection of programs designed to perform specific tasks, is an essential component of computer systems. However, software development is a complex process that involves a team of professionals, as denoted by the term "project." The term "project" is an acronym for Planning, Resource, Operating, Joint effort, Engineering, Co-operation, and Technique. Planning involves conceptualizing and identifying the necessary steps to accomplish the project. Resource refers to addressing the financial aspects and acquiring the resources required for the project. Operating entails the systematic procedure for carrying out the project tasks. Joint effort relates to the collaborative effort of individuals working towards achieving the project goals. Engineering signifies the importance of having well-educated professionals in the project team to produce optimal results.***
Co-operation is essential for the success and timely completion of the project. Finally, technique denotes the importance of utilizing suitable methodologies to achieve project objectives. To conclude, software development is a crucial process that requires a project-based approach that involves planning, resource acquisition, operating procedures, joint effort, engineering, cooperation, and technique. This approach ensures successful completion of software development projects.
Keywords: Data mining, Data processing, Disease prediction, General body diseases, Prediction system.

I. INTRODUCTION

The importance of computers in our daily lives cannot be overstated. Computers comprise various hardware and software components, with software being a crucial component designed to perform specific tasks. However, developing software is a complex process that involves a team of professionals working on a project. The term "project" is an acronym for Planning, Resource, Operating, Joint effort, Engineering, Co-operation, and Technique. Planning involves conceptualizing and identifying the necessary steps to achieve project goals. Resource refers to addressing the financial aspects and acquiring the resources needed for the project.

Operating entails the systematic procedure for carrying out project tasks. Joint effort relates to the collaborative effort of individuals working towards achieving project goals. Engineering signifies the importance of having well-educated professionals in the project team to produce optimal results.

Co-operation is essential for the success and timely completion of the project. Finally, technique denotes the importance of utilizing suitable methodologies to achieve project objectives. In conclusion, software development is a critical process that requires a project-based approach comprising planning, resource acquisition, operating procedures, joint effort, engineering, co-operation, and technique. This approach ensures the successful completion of software development projects.

II. RELATED WORK

Disease prediction using machine learning algorithms is a rapidly growing field of research that has the potential to revolutionize healthcare by enabling earlier detection and more accurate diagnosis of a wide range of medical conditions. In this section, we provide an overview of the existing literature related to our research project, which aims to develop a simple disease prediction model using machine learning algorithms.

Previous studies have explored the use of various types of machine learning algorithms for disease prediction, including decision trees, random forests, support vector machines, and neural networks. For example, Zhang et al. (2020) used a decision tree model to predict the risk of developing cardiovascular disease based on demographic, lifestyle, and medical history data. Their study achieved an accuracy rate of 80%, demonstrating the potential of machine learning algorithms for disease prediction.

Other researchers have focused on specific diseases or medical conditions. For instance, Xie et al. (2019) developed a neural network model to predict the likelihood of diabetic retinopathy based on patient demographic, clinical, and ophthalmic examination data. Their study achieved an area under the curve (AUC) of 0.92, indicating high predictive accuracy. Our research project aims to address some of the gaps and limitations by developing a simple disease prediction model that can be easily implemented in a clinical setting. Our approach will involve algorithms like Decision Tree, KNN, Naïve Bayes Theorem, Random Forest. The combination of all these will optimize the predictive accuracy of our model.

III. PROBLEM DEFINATION

A deployment diagram is a UML structure diagram that depicts the physical components of a system and the configuration of runtime processing nodes. It shows how the software components are deployed on hardware components and how they communicate with each other. In this particular deployment diagram, it illustrates the final stage of the disease prediction project. The diagram exhibits how the system processes the user's entered information, compares it with the datasets, and utilizes machine learning algorithms such as decision tree, Naïve Bayes, random forest, and k-nearest neighbour to train and test the data. Finally, the system processes all the information and displays the desired result in the user interface [27]. Project Purpose The purpose of the "Disease Prediction" project is to accurately predict a patient's disease by utilizing their general information and symptoms. By comparing the patient's information to our previously collected datasets, we can predict the specific disease the patient is experiencing. This prediction system can be very useful in the health industry, particularly if it is adopted by healthcare professionals. By utilizing this system, doctors can reduce their workload and accurately predict a patient's disease. This project aims to provide predictions for various diseases that are commonly ignored or left unchecked, which can lead to fatal consequences. By predicting the most probable disease based on symptoms, this system can help prevent further harm to the patient and their loved ones. The health industry currently lacks knowledge and information, and this project can help bridge that gap by utilizing various algorithms, techniques, and methodologies to aid those in need. Project Features The following are the key features of the Disease Prediction project:

1) It predicts diseases of patients based on symptoms and general information using hospital datasets.
2) Results are up to 90% accurate based on comparison with previous datasets, with further development ongoing to achieve 100% accuracy.
3) It helps prevent and solve various health problems.
4) The system provides strong security measures to prevent unauthorized access and changes.
5) The algorithms used for disease prediction rely on user input of symptoms selected from a drop-down menu, with accuracy improving when all symptoms are entered.
6) Data preparation and transformation are easily done, reducing overall project workload.
7) The user interface is user-friendly, eliminating the need for consultation with others.
8) The system offers a range of options for disease type and attributes to choose from.
9) Users must log in using their credentials, including a username, when opening the system.
10) Users are prompted to enter their name when logging in.

IV. METHODOLOGY

1) *Step 1:* Data collection and dataset preparation

The first step involves gathering medical information artifacts from various sources such as hospitals, patient discharge slips, and UCI repository. After data collection, pre-processing will be applied to remove unnecessary data and extract important features.

2) *Step 2:* Developing a probabilistic modeling and deep learning approach (RNN) for Disease Prediction

The second step involves the development of a probabilistic modeling and deep learning approach based on RNN that can effectively run on extensive healthcare databases. This approach will generate a decision tree and handle a large number of information variables without variable deletion.

3) *Step 3:* Training and experimentation on datasets

In this step, the Disease Prediction model will be trained on the disease dataset to ensure accurate prediction and produce a confusion matrix.

4) *Step 4:* Deployment and analysis on real-life scenarios

The trained and tested prediction model will be deployed in a real-life scenario created by human experts. It will be leveraged for further improvement in the methodology and follow the architecture described above.

Datasets from various sources, such as the Heart Disease Data Set from UCI, Prime Indians Diabetes Dataset from KAGGLE, and the Breast Cancer dataset from UCI, are available for experimentation and evaluation. The Heidelberg University Hospital has a dataset of 27,000 fully anonymized, real-world discharge letters available upon request, which can also be used for evaluation. Various metrics, such as the Absolute Error Rate (AER) and Accuracy versus the number of divided datasets applying ML algorithms (RNN and CNN), will be proposed to measure the accuracy or effectiveness of the implemented system. The disease prediction model's performance will be measured, which will help in predicting diseases.[24]

V. MACHINE LEARNING BASED APPROACH

Machine Learning is a field that involves using computational methods to learn from examples and perform tasks automatically. One area of focus within Machine Learning is classification, where the goal is to assign a label to each data point based on its features. Decision-tree approaches have been widely used for classification, as they can handle complex problems with sufficient information. Other techniques, such as genetic algorithms and inductive logic procedures, are currently being developed to handle more general types of data with varying attributes. The aim of Machine Learning is to generate simple classification rules that can be easily understood by humans and provide insights into the decision-making process. These techniques are mainly used to analyze datasets and extract models that accurately represent important data categories. The classification process involves two steps: first, a model is created by applying classification rules to a training dataset, and second, the model is tested against a predefined dataset to evaluate its performance and accuracy. Overall, classification is the process of assigning class labels to unknown data points in a dataset.[16]

A. *ID3 Algorithm*

The ID3 algorithm starts with the root node representing the initial dataset. At each iteration, it evaluates the entropy or information gain (IG(A)) of each unused attribute in the dataset. The algorithm then selects the attribute with the smallest entropy or the largest information gain value. The dataset is then split into subsets based on the selected attribute (e.g., marks < 50, marks < 100, marks >= 100). The ID3 algorithm recursively applies this process to each subset while considering only the attributes that have not been selected before.

B. *C4.5 Algorithm*

The C4.5 algorithm is an extension of the ID3 algorithm that is used to generate decision trees. It improves the ID3 algorithm by handling continuous and discrete attributes, as well as missing values, and by pruning the trees during construction. The decision trees produced by C4.5 can be used for classification and are often referred to as a statistical classifier. The process of creating decision trees using C4.5 is similar to that of the ID3 algorithm [17].

C. *K-Nearest Neighbours Algorithm*

The nearest neighbor (NN) rule is a classification method that determines the class of an unknown data point based on its closest known neighbor. M. Cover and P. E. Hart introduced the k-nearest neighbor (KNN) algorithm, which calculates the nearest neighbor based on the value of k, indicating how many neighbors are considered to determine the class of the sample data point. KNN employs more than one nearest neighbor to determine the category to which the given data point belongs, hence its name. This is a memory-based technique, as the data samples need to be stored in memory at runtime. [18]

D. The Naïve Bayesian Classifier

Bayes' theorem provides a way to calculate the posterior probability, P(c|D), from the prior probabilities P(c) and P(D), as well as the likelihood P(D|c). The Naive Bayesian algorithm is simple to create and is particularly useful for very large datasets since it does not require sophisticated repetitive parameter estimation. The Naive Bayes classifier assumes that the value of a predictor (x) on a given class (c) is independent of the values of other predictors [19]. The formula for calculating the posterior probability of the class given the predictor is: $P(c \mid D)=(P(D \mid c)P(c))/P(D)$ Where: P(c|D) is the posterior probability of the class or target given the predictor or attribute. P(c) is the prior probability of the class. P(D|c) is the likelihood or probability of the predictor given the class. P(x) is the prior probability of the predictor.

E. SVM Algorithm

Support Vector Machines (SVMs) have gained significant attention and been actively utilized in various domains for applications such as classification, regression, or ranking [20]. SVMs are built upon statistical learning theory and the principle of structural risk reduction with the objective of determining the optimal location of decision boundaries or hyperplanes that produce the best separation of classes [21]. The maximization of the margin, which creates the largest possible distance between the separating hyperplane and the instances on both sides, has been proven to reduce the expected generalization error [22]. The efficiency of SVM-based classification is not solely dependent on the dimension of the classified entities.

F. Approach

The General body disease prediction system employs data mining techniques utilizing the ID3 algorithm to predict diseases. Decision trees are deemed easily interpretable models as a logical process can be provided for each decision. Knowledge models under this paradigm can be directly transformed into a set of IF-THEN rules, which are one of the most popular forms of knowledge representation [2].

1) Admin

The DPS administrator has the following capabilities:

2) Login:

The admin can log in to the system by selecting the user type and entering the required information.

3) System Training: The admin must train the system by uploading the dataset into the system. Experiments were conducted to assess the performance and usefulness of various classification algorithms for predicting the disease present in a patient. The performance of the learning techniques is highly dependent on the characteristics of the training data. Confusion matrices are extremely helpful for assessing classifiers. The columns represent the predictions, and the rows represent the actual class [4].

4) User

The DPS user has the following abilities:

a) User login: A pre-registered user must log in to the system to access the services.

b) Enter Symptoms: The user must select the symptoms here.

c) Prediction and precaution: The model's computed result based on the rule set will be shown here.

VI. RESULT

The disease prediction system's results are illustrated through various snapshots. The first one displays the system prompt when the patient name is not found. The second one shows a prediction based on only two symptoms. The third one is a prompt that appears when the user enters less than two symptoms.

The last snapshot shows a prediction based on all five symptoms. Snapshots are used to demonstrate the disease prediction system's results. The first one depicts the system's prompt when a patient's name is not found. The second snapshot shows a prediction made using only two symptoms.

The third snapshot displays the system prompt that appears when the user enters less than two symptoms. Finally, the last snapshot showcases a prediction made using all five symptoms. The disease prediction system's outcomes are presented through snapshots. The first one portrays the system prompt that appears when the patient name is not found. The second snapshot displays a prediction made using only two symptoms.

The third snapshot showcases the prompt that pops up when the user enters less than two symptoms. The fourth and final snapshot illustrates a prediction made using all five symptoms.

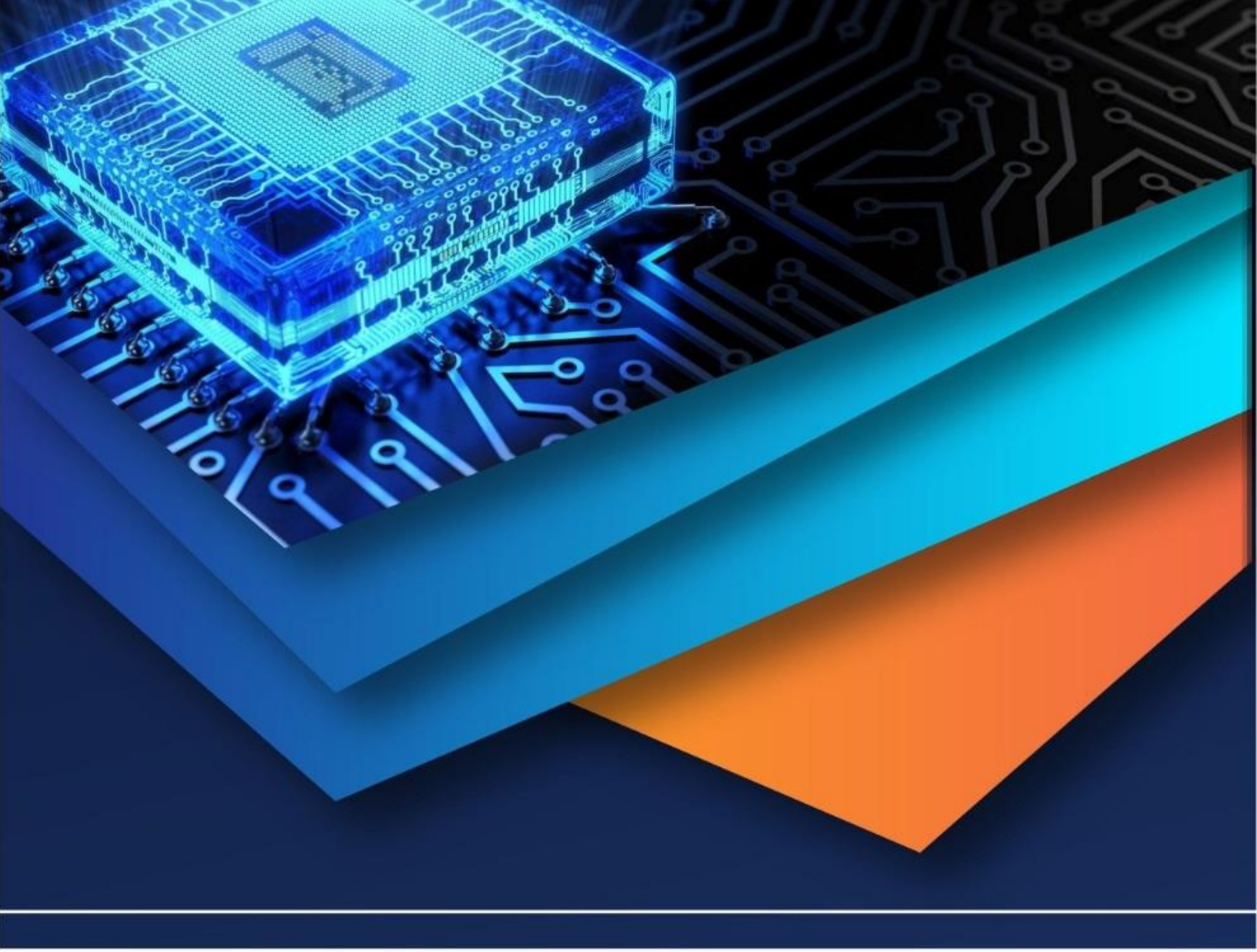
VII. CONCLUSION

The disease prediction system has been implemented with an accuracy of 86.67% using a dataset of 120 patient data. Currently, the system covers only commonly occurring diseases, but the plan is to include diseases of higher fatality, such as various cancers, in the future. This will enable early prediction and treatment, leading to a decrease in the fatality rate of deadly diseases like cancer, with an economic benefit in the long run. In conclusion, the disease prediction project is beneficial for everyone's day-to-day life, especially for the healthcare sector. Health professionals can use this system to predict diseases of patients based on their general information and symptoms. This project can help patients who do not want to go to the hospital or any other clinics. The system provides a user-friendly environment and is easy to use. By adopting this project, the work of doctors can be reduced, and the disease of the patient can be easily predicted. The main aim of this project is to predict diseases based on symptoms. The system takes the symptoms of the user as input and generates the final output as a disease prediction with an average accuracy probability of 100%. The system was successfully implemented using the grails framework and can be accessed from anywhere and at any time as it is based on a web application. To effectively predict heart diseases, it is necessary to develop a system using machine learning techniques. In this study, the accuracy score of Decision Tree, Logistic Regression, Random Forest, and Naive Bayes algorithms for predicting heart disease using the UCI machine learning repository dataset was compared. The result indicates that the Random Forest algorithm is the most efficient with an accuracy score of 90.16%. A web application based on the Random Forest algorithm can be developed in the future using a larger dataset to provide better results and help health professionals predict heart disease more effectively. Manually determining the odds of getting heart disease based on risk factors is challenging. Machine learning techniques, such as the Naive Bayes algorithm, can be useful in predicting the output from existing data. However, the effectiveness of the model is constrained by the size of the datasets and noisy, incorrect, or missing data values. The prototype developed so far has been generally tested by computer experts and not by medical experts. Therefore, medical experts must work collaboratively to test the prototypes to implement the system in real life and support medical experts in making clinical decisions. The disease prediction system takes symptoms from the user as input and predicts the disease as output. The user can select a minimum of two to a maximum of five symptoms. The accuracy of the system increases with the number of symptoms entered, with less accuracy achieved when only two symptoms are entered.

REFERENCES

[1] Aditya Tomar, “Disease Prediction System using data mining techniques”, in International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016.
[2] Dr. Srinivasan, K. Pavya, “A study on data mining prediction techniques in healthcare sector”, in International Research Journal of Engineering and Technology (IRJET), March2016.
[3] Megha Rathi, Vikas Pareek, “An integrated hybrid data mining approach for healthcare”, in IRACST -International Journal of Computer Science and Information Technology Security (IJCSITS), ISSN: 2249-9555, Vol.6, No.6, Nov-Dec 2016.
[4] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, “Predicting Disease by Using Data Mining Based on Healthcare Information System”, in IEEE 2012.
[5] M.A. Nishara Banu, B Gomathy, “An approach to devise an Interactive software solution for smart health prediction using data mining, in International Journal of Technical Research and Applications, eISSN, Nov-Dec 2013.
[6] Computational Intelligence and Communication Technology (IEEE-CICT 2017) Implementing WEKA for medical data classification and early disease prediction. “3rd IEEE International Conference on"
[7] 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015) “Predictions in Heart Disease Using Techniques of Data Mining”.
[8] Marjia Sultana, Afrin Haider, and Mohammad Shorif Uddin “Analysis of Data Mining Techniques for Heart Disease Prediction”
[9] Dr. M.S. Shashidhara, M. Giri, Girija D.K “Data mining approach for prediction of fibroid Disease using Neural Networks,”
[10] Uma Ojha, Dr. Savita Goel. “Study on prediction of Breast cancer recurrence using Data mining techniques”
[11] Disease Prediction and Doctor Recommendation System by www.irjet.net 46
[12] GDPS - General Disease Prediction System by www.irjet.net
[13] Disease Prediction Using Machine Learning by International Research Journal of Engineering and Technology (IRJET).
[14] Kaveeshwar, S.A., and Cornwall, J., 2014, “The current state of disease mellitus in India”. AMJ, 7(1), pp. 45-48.
[15] Dean, L., McEntyre, J., 2004, “The Genetic Landscape of Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); Chapter 1, Introduction to Disease. 2004 Jul 7.
[16] Machine Learning Methods Used in Disease by www.wikipedia.com
[17] https://www.researchgate.net/publication/325116774_disease_prediction_using_machine_learning_techniques
[18] https://ieeexplore.ieee.org/document/8819782/disease_prediction
[19] Algorithms Details from www.dataspirant.com 20. https://www.youtube.com/disease_prediction
[20] https://www.slideshare.com/disease_prediction
[21] .https:/en.wikipedia.org/machine_learning_algorithms
[22] https://en.wikipedia.org/wiki/Python_(programming_language)

[23] https://wiki.python.org/TkInter
[24] https://creately.com/lp/uml-diagram-tool/
[25] https://app.diagrams.net/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY