iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75492

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

AI-Powered Symptom-Based Disease Prediction

Dr. Sumalatha Bandari[1], Mr. Siddharth Raipure[2], Ms. Sakshi Varma[3], Ms. Sakshi Kate[4], Mr. Sameer Gahlot[5], Ms. Srushti Karlewar[6]

[1]Assistant Professor, Department of Computer Science and Engineering .G H Raisoni University, AMRAVTI, Nagpur
[2, 3, 4, 5]Research Scholar, Department of Computer Science and Engineering .G H Raisoni University, Amaravti, Nagpur

Keyword: Symptom-based diagnosis, Disease prediction, Machine learning, Healthcare analytics, Predictive modeling, Medical decision support system, Classification algorithms, Data mining in healthcare, Feature selection, Symptom–disease correlation, Clinical data analysis, Patient symptom dataset, Artificial intelligence in healthcare, Supervised learning, Diagnosiautomation, Machine learning models (SVM, Random Forest, Naïve Bayes),Healthcare prediction system, Data preprocessing, Early disease detection, Health informatics

I. INTRODUCTION

Today, healthcare faces the big challenge of diagnosing diseases fast.Because the human body is complex and diseases have many different symptoms, it is hard for doctors to find what is wrong with a patient. This is exacerbated by the large number of patients that require medical attention; it increased the waiting time for a doctor appointment and examination. Longer waiting will delay appropriate treatment and impact the patient's health.This problem affects people everywhere-whether they live in cities or small towns,or remote areas where there isn't much healthcare.

People most vulnerable, such as the elderly, children, and those with poor access to health care, bear the greatest burdens of undiagnosed or misdiagnosed diseases. Machine learning provides a promising solution to this issue.It produces predictive models by using data to assist doctors in diagnosing diseases more effectively and with greater speed. These models are able to discover trends and links within huge symptom-disease sets that could remain obscure to human beings.

This enables the prediction of the disease based on symptoms; therefore, it makes the diagnosis faster and more accurate.This approach can be helpful in many aspects. It could facilitate timely diagnosis and treatment of patients, which may result in improved health and potential financial savings on medical care. Machine learning tools may be instrumental for doctors, nurses, and other healthcare professionals in making better decisions and informed medical choices.

Correspondingly, healthcare systems can also benefit from more efficient and effective diagnostic tools to reduce pressure and improve the quality of care. The project is very important to the medical fraternity since it's about modernizing healthcare to make it better. The development and implementation of machine learning for disease prediction will revolutionize the way diseases are diagnosed around the world. These models are more efficient, more accurate, and more accessible. They are the foundational building blocks for progress in personalized medicine, with treatments to be provided on an individual level as guided by predictions from the analysis of data. Integrating these models into standard medical practice is a major step toward better health outcomes globally.

II. RELATED WORK

The primary objective of "Disease Prediction using Machine Learning" is to demonstrate how machine learning is possible to predict diseases by their symptoms. It discusses the significance of proper disease prediction in healthcare and addresses the application of Naive Bayes Classifier and other algorithms such as linear regression and decision tree in diseases such as Diabetes, Malaria, Jaundice, Dengue and Tuberculosis.

The article "Improving Disease Prediction by Machine Learning" investigates how machine learning and big data can be used to improve disease prediction. It talks about the rising prevalence rate of non-communicable diseases in India and emphasizes that there is necessity to detect them early.

This paper discusses the use of machine learning in predicting heart disease in diabetic patients, an area that is not sufficiently covered with data to make accurate predictions. It summarizes the data mining in health care and the way it can be used to identify those patterns and relationships that are impossible to see in large medical data. It uses various machine learning methods such as Naive Bayes, SVM and Decision Tree to predict disease.

The article titled Diabetes Disease Prediction Based on Symptoms Using Machine Learning Algorithms in the annals of R.S.C.B. addressed the topic of predicting diabetes with the help of machine learning based on the symptoms. The problem of diabetes is on the rise, and thus early detection is a significant aspect when dealing with diabetes.

The study applied an ensemble technique, which involves using various machine learning models such as Naive Bayes Classifier, SVM Classifier, J48 and Multilayer Perceptron to enhance the accuracy of predictions.

Diseases are classified using three machine learning algorithms that include Decision Tree, Random Forest, and LightGBM. The performance of such algorithms can be enhanced with the help of data preprocessing techniques.

Decision Tree model applies theID3 method of classification and LightGBM applies boosted foreststo.accelerate and increase precision. The study will be focused on identifying the relevant risk factors, differentiating the various methods of classification, and understanding the impact of altered risk factors in the prediction of disease using machine learning, which in this case is Naive Bayes Classifier that makes predictions given the input symptoms. It pools both structured and unstructured data of hospitals to enhance the analysis of different diseases.

The suggested system passes through training and testing processes to enhance the ability to predict the diseases without physical consultation. It does also operate both kinds of data to provide a holistic approach.

Review of medical disease symptoms prediction using data mining technique" is a review article on the application of data mining to predict severe diseases in medicine. It is concerned with the selection of the most appropriate classifiers and ensemble techniques. It talks about several methods of improving data mining, such as fuzzy logic, feature.machine learning, optimization, and machine learning. The proposed model will pick several clusters to operate upon as an ensemble, compute the average performance of each classifier on the clusters and classify data using the individual classifier with the best average performance.

The article "Computer-based Disease Prediction and Medicine Recommendation using Machine Learning Approach" proposes a tech to predict disease and recommend medicines through machine learning. It also highlights some of the difficulties of the conventional drug discovery methods and demonstrates the way artificial intelligence and, in particular, machine learning would be useful in accelerating the creation of medicine.

III. METHODOLOGY

A. Dataset

The data consists of symptoms of various illnesses, and each symptom is a binary characteristic. There are a total of 132 features. The sample size of the training dataset is 4920, and the test dataset is 42 samples. Itching, joint pain and skin rash are the top three most prevalent symptoms of the training data.

The mean frequency of various symptoms used in the training data is between the range of 0.02 and 0.16 and the standard deviation is used to indicate the level to which a symptom occurrence can differ. The same tendency is observed in the test data. The large size of the symptom list of diseases, which is explained by a variety of diseases, aids us in the construction of a machine learning model that would assist us in the prediction of diseases, given a variation in symptom pattern. The many features we have implies that we must be keen when making choices in the selection of the best features and streamlining the model so as to come up with a successful and efficient disease prediction system.

B. EDA

AIDS, Acnes, alcoholic hepatitis, allergy, arthritis, bronchial asthma, cervical spondylosis, chickenpox, chronic cholestasis, common cold, dengue, diabetes, dimorphic haemorrhoids, drug reaction, fungi infection, GERD, gastroenteritis, myocardial infarction, hepatitis A,B,C,D,E, hypertension, hyperthyroidism, hypoglycaemia, hyperthyroidism, impetigo, jaund

C. Modeling

1) KNN

The project Disease Prediction By Symptoms consists of the K-Nearest Neighbors classifier having Manhattan distance and distance-based weighting.The input data used to train the model in predicting diseases are the symptoms. Under normal circumstances, on making a prediction, the KNN model examines the K nearest set of neighbors and forecasts the commonest one in the group. This is indicated by very high train and test accuracy scores of 1.0, indicating that the model fits into training and test data perfectly. This may imply that overfitting may be a problem. The accuracy and the recall are both optimal with the model being good at getting the true positives and capturing all the positives.

Nevertheless, such high scores should be approached with a lot of care due to the chances of overfitting. Some further checks may be required like cross checks or other algorithms to make sure that the model is working in different cases.

2) Gaussian NB

This work predicts diseases by a Gaussian Naive Bayes classifier, which is trained with a set of symptoms.In this instance, the Naive Bayes is used under the assumption that features become independent when the classification is available, so it is effective when the high is great.dimensional data, such as symptoms. The model has done incredibly well with training and test data with a perfect accuracy, precision, and recall as well as F1-score of 1.0. High scores are an indication that the model has learnt the correlations among the symptoms and diseases. Nevertheless, the model may be strengthened and made more applicable through additional validation using alternative datasets and other algorithms. All in all, the Gaussian Naive Bayes model has tremendous potential of disease prediction using symptoms, indicating machine learning potential in health application.

3) SVM

The Support Vector Machine model was applied in this project to predict diseases using symptoms.SVM is a supervised learning model that assists in identifying the optimal hyperplane to use in segregating different classes. Here, the SVM model attained excellent results: 100 percent accuracy, precision, recall, and F1-score on the training and test data. It was an event that there were no misclassifications in all the instances. It is revealed by the high performance that the model was able to capture and generalize the underlying patterns in the data to the new and unknown examples. These strong results implicate that the SVM model can be highly applicable in the prediction of diseases and enables the effective disease evaluation on the basis of the symptoms. This may be a useful input to health decision making and patient management.

4) Decision Tree

The Decision Tree model is effective in prediction of diseases basing on symptoms.It splits the data into various segments and splits it according to the most practical features. This will continue until all the data points are classified under a particular category. The model achieves 100% accuracy, precision, recall and F1-score during training and testing, that is, it is neither overfitting nor underfitting. Such great outcomes demonstrate that the model has learnt the relationships between symptoms and diseases quite well. This renders it a reliable prediction tool of diseases. Nevertheless, in the context of implementing it in the actual healthcare setting, we must check data biases and ensure that it is effective in other groups. Making the model more updated with additional data will further increase its performance.

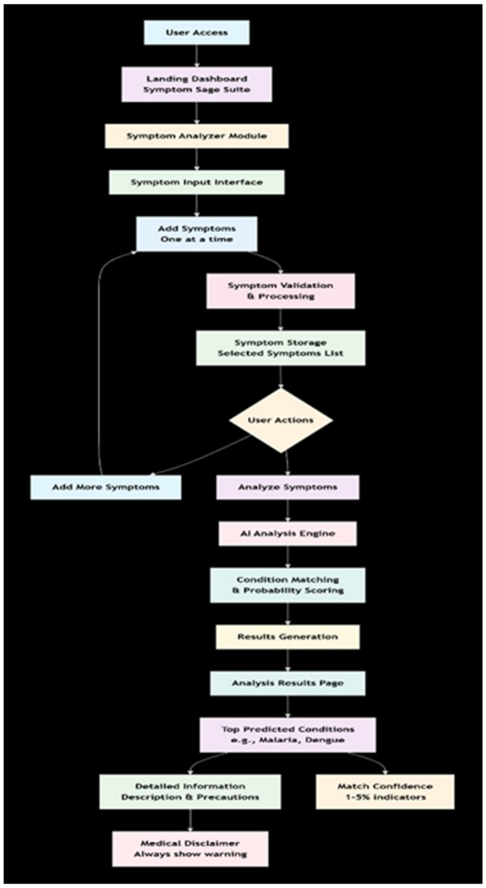
5) Random Forest

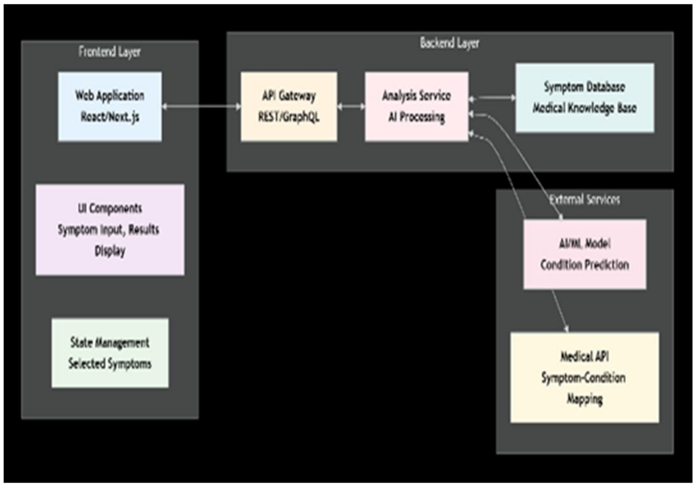
Random Forest model is superior to the others in most areas of performance.It demonstrates ideal accuracy, precision, recall, and F1-score on training and testing data that demonstrates it is robust and can be applicable to new data. Random Forest involves using several decision trees and combining their result in order to arrive at a prediction. It builds trees with random data samples, and also randomly selects features, and that is why it minimises errors with each individual tree. This renders the entire model more precise. The model, in this case, has learnt to relate symptoms with diseases in a very effective way. These are ideal scores that imply that it can be very reliable when distinguishing various diseases, depending on symptoms within a medical facility.

6) XG Boost

XGBoost model offers quite good results considering every measure of evaluation.Similar to the Random Forest, it has scored 100 on accuracy, precision, recall, and F1-score on both training and testing data, and it is very strong and able to use knowledge in new data. XGBoost is a gradient boosting algorithm, which is known to be both efficient and effective where structured data is involved. It constructs the decision trees sequentially where each tree corrects the errors of the preceding trees. XGBoost has methods of gradient descent to minimize the errors and enhance predictions. The model in this case has acquired knowledge of the intricate association that exists between diseases and symptoms. These ideal scores impose that the model can be trusted to differentiate various diseases in terms of symptoms and therefore it can be confidently used in medical diagnostics. Its advantages involve processing large and complex data and it is not overfitted easily, hence it can be highly applicable in various predictive activities, particularly in healthcare.

IV. RESULTS AND DISCUSSION





Evaluation Metrics for KNN
The evaluation metrics for KNN are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for GaussianNB
The evaluation metrics for GaussianNB are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for SVM
The metrics used for evaluating SVM are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Decision Tree Evaluation Metrics
The evaluation metrics for Decision Tree are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for Random Forest
The evaluation metrics for Random Forest are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

Evaluation Metrics for XGBoost
The evaluation metrics for XGBoost are:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1score: 1.0

In summary, the Symptom based Disease Prediction Projectis a good advance towards utilising machine learning in medicine. We have demonstrated that machine learning can be used to diagnose diseases accurately based on the symptoms, with such models as K-Nearest Neighbours, Gaussian Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, and XGBoost. The majority of these models worked very well with a score of 1.0 in all the measures, which represents that they are very effective in identifying the underlying connections between symptoms and diseases.

Such models could be applied to an ordinary web app, which allows individuals to test their symptoms and receive health recommendations his facilitates its awareness among people they take action at an early stage and ensure their health, which may positively impact health and decrease the burden on healthcare systems. Nevertheless, one should not disregard certain things.

Although these models are effective on data that we had, they should be applied and tested on more various kinds of data and on the real world scenarios so as to be certain that they are effective at all times. Moreover, the models are supposed to be maintained in terms of the changes in healthcare and the arrival of new information.

Symptom-based disease prediction can transform the healthcare working process.

It can enable individuals to discover more rapidly and properly about their health, resulting in improved health outcomes and improved healthcare systems. Machine learning is transforming the healthcare sector and this project demonstrates that it can create a tangible difference in the health outcomes across the globe.

REFERENCES

[1] Gomathy, C. K., and Mr. A. Rohith Naidu. "The prediction of disease using machine learning." International Journal of Scientific Research in Engineering and Management (IJSREM) 5.10 (2021).

[2] Singh, Smriti Mukesh, and Dinesh B. Hanchate. "Improving disease prediction by machine learning." Int. J. Res. Eng. Technol 5 (2018): 1542-1548.

[3] Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings 80 (2023): 3682-3685.

[4] Sujatha, K., et al. "Diabetes Disease Prediction Based on Symptoms Using Machine Learning Algorithms." Annals of the Romanian Society for Cell Biology 25.6 (2021): 3805-3817.

[5] Bhanuteja, Talasila, et al. "Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach." International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075.

[6] Pingale, Kedar, et al. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6.12 (2019): 831-833.

[7] Sah, Rahul Deo, and Jitendra Sheetalani. "Review of medical disease symptoms prediction using data mining technique." IOSR Journal of Computer Engineering 19.3 (2017): 59-70.

[8] Gupta, Jay Prakash, Ashutosh Singh, and Ravi Kant Kumar. "A computer-based disease prediction and medicine recommendation system using machine learning approach." Int J Adv Res Eng Technol (IJARET) 12.3 (2021): 673-683.

[9] Kanakaraddi, Suvarna G., et al. "Disease prediction using data mining and machine learning techniques." Advanced Prognostic Predictive Modelling in Healthcare Data Analytics (2021): 71-92.

[10] Jadhav, Saiesh, et al. "Disease prediction by machine learning from healthcare communities." International Journal of Scientific Research in Science and Technology 5 (2019): 8869-8869.

[11] M. Sirigineedi, T. Kumaravel, P. Natesan, V. K. Shruthi, M. Kowsalya, and M. S. Malarkodi: “Deep Learning Approaches for Autonomous Driving to Detect Traffic Signs”, 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023.

[12] M. Sirigineedi, R. N. V. J. Mohan and B. Sahu: “Improving Fisheries Management through Deep learning based Automated fish counting”, 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023.

[13] M. Srikanth, Bhanurangarao M, Manikanta Sirigineedi, Padma Bellapukonda: “Integrated Technologies for Proactive Bridge-Related Suicide Prevention”, Journal of Namibian Studies, Volume 1, Issue 33, Pages 2117-2136, Sep 2023.

[14] Srikanth Mandela, Padma Bellapukonda, Manikanta Sirigineedi: “Using Machine Learning and Neural Networks Technologies, a Bottom-Up Water Process Is Being Used To Reduce All Water Pollution Diseases”, Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), vol. 2, Oct. 2022.

[15] M. Srikanth, Padma Bellapukonda, Manikanta Sirigineedi: Protecting tribal peoples nearby patient care centers use hybrid techniques based on a distribution network, International Journal of Health Sciences, 2022.

[16] John, A., McGregor, J., Fone, D., Dunstan, F., Cornish, R., Lyons, R. A., & Lloyd, K. R. (2016). Case-finding for common mental disorders of anxiety and depression in primary care: externalvalidation of routinely collected data. BMC Medical Informatics and Decision Making, 16(1), 35

[17] Essau, C. A., Lewinsohn, P. M., Lim, J. X., Ho, M.-H. R., & Rohde, P. (2018). Incidence,recurrence, and comorbidity of anxiety disorders in four major developmental stages. Journal of Affective Disorders, 228, 248–253

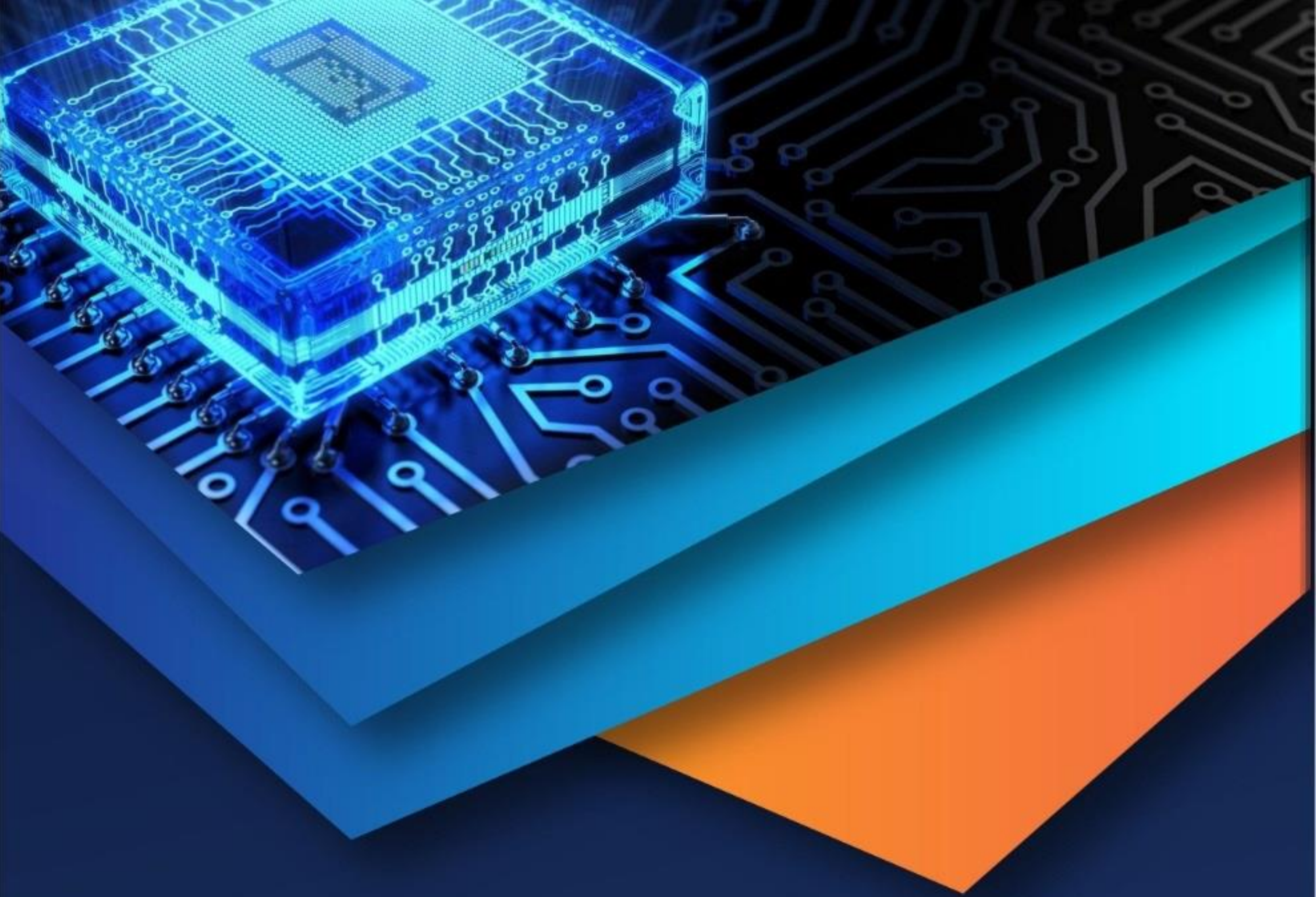
[18] Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F., Aboyans, V., Abu-Raddad, L. J., Ackerman, I. N., Adamu, A. A., Adetokunboh, O., Afarideh, M., Afshin, A., Agarwal, S. K., Aggarwal, R., … Murray,C. J. L. (2017). Global, regional, and national incidence, prevalence, and years lived withdisability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis forthe Global Burden of Disease Study 2016. Lancet, 390(10100), 1211–120

[19] Psychiatric Association - Apa, A. (Ed.). (2014). Manual diagnóstico y estadístico de los trastornosmentales DSM-5 (5a). En Madrid: Editorial Médica Panamericana.

[20] Somers, J. M., Goldner, E. M., & Waraich, P. (2006). Prevalence and incidence studies of anxiety disorders. A systematic review of the literature. Can J Psychiatry, 51, 100–113.

[21] Costa e Silva, J. A. (1998). The public health impact of anxiety disorders: a WHO perspective.Acta Psychiatrica Scandinavica. Supplementum, 393, 2–5.

[22] National Collaborating Centre for Mental Health (UK). (2013). Social anxiety disorder. British Psychological Society



doi crossref
10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

SJIFactor
TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY