iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42214

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 10 Issue V May 2022- Available at www.ijraset.com

Disease Prediction Using Machine Learning Algorithms KNN and CNN

K. Praveen Kumar[1], A. Pravalika[2], R. Pallavi Sheela[3], Y. Vishwam[4]
[1, 2, 3, 4]*Anurag Group of Institutions*

Abstract: ***Nowadays, people face various diseases due to environmental conditions and their living habits. So the prediction of disease at an earlier stage becomes an important task. But the accurate prediction based on symptoms becomes too difficult for the doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has a large amount of data growth per year. Due to the increasing amount of data growth in the medical and healthcare field the accurate analysis of medical data has been the benefit of early patient care. With the help of disease data, data mining finds hidden pattern information in a huge amount of medical data. We proposed general disease prediction based on the symptoms of the patient. For disease prediction, we use K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithms for the accurate prediction of disease. Disease prediction required a disease symptoms dataset. In this general disease prediction, the living habits of a person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using CNN is 84.5% which is more than the KNN algorithm. And the time and the memory requirement are also more in KNN than in CNN. After general disease prediction, this system can give the risk associated with the general disease which is a lower risk of general disease or higher.***

Keywords: KNN, CNN and Healthcare

I. INTRODUCTION

Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is an ongoing struggle for the past decades. Many works have applied data mining techniques to pathological data or medical profiles for the prediction of specific diseases. These approaches tried to predict the reoccurrence of the disease. Also, some approaches try to do prediction on control and progression of the disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre-processing and produce more accurate results. With the development of big data technology, more attention has been paid to disease prediction from the perspective of big data analysis; various researches have been conducted by selecting the characteristics automatically from a large amount of data to improve the accuracy of risk classification rather than the previously selected characteristics. The main focus is on using machine learning in healthcare to supplement patient care for better results. Machine learning has made it easier to identify different diseases and diagnose them correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients.

The healthcare industry produces large amounts of health- care data daily that can be used to extract information for predicting diseases that can happen to a patient in the future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision-Making for patients' health. Also, these areas need improvement by using informative data in healthcare. One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis andtreatment options can be made. Machine learning in healthcare aids humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented inhealthcare can lead to increased patient satisfaction. The k-mean algorithm is used to predict diseases using patient treatment history and health data.

II. LITERATURE SURVEY

Dahiwade et al. [9] proposed a ML based system that pre- dicts common diseases. The symptoms dataset was imported from the UCI ML repository, where it contained symptomsof many common diseases. The system used CNN and KNNas classification techniques to achieve multiple diseases pre- diction. Moreover, the proposed solution was supplemented with more information that concerned the living habits of the tested patient, which proved to be helpful in understanding thelevel of risk attached to the predicted disease. Dahiwade et al.

[9] compared the results between KNN and CNN algorithm in terms of processing time and accuracy. The accuracy and processing time of CNN were 84.5% and 11.1 seconds, respectively. The statistics proved that KNN algorithm is underperforming compared to CNN algorithm. In light of this study, the findings of Chen et al. [10] also agreed that CNN outperformed typical supervised algorithms such as KNN, NB,and DT. The authors concluded that the proposed model scoredhigher in terms of accuracy, which is explained by the capability of the model to detect complex nonlinear relationships in the feature space. Moreover, CNN
Detects features with high importance that renders better description of the disease, which enables it to accurately predict diseases with high complexity [9], [10].

This conclusion is well supported and backed with empirical observations and statistical arguments. Nonetheless, the presented models lacked details, for instance, Neural Networks parameters such as network size, architecture type, learning rate and back propagation algorithm, etc. In addition, the analysis of the performances is only evaluated in terms of accuracy, which debunks the validity of the presented findings[9]. Moreover, the authors did not take into consideration the bias problem that is faced by the tested algorithms [9], [10]. In illustration, the incorporation of more feature variables could immensely ameliorate the performance metrics of under performed algorithms [11].

Dahiwade et al. [9] proposed a ML based system that predicts common diseases. The symptoms dataset was imported from the UCI ML depository, where it contained symptoms of many common diseases. The system used CNN and KNN as classification techniques to achieve multiple diseases pre- diction. Moreover, the proposed solution was supplemented with more information that concerned the living habits of the tested patient, which proved to be helpful in understanding the level of risk attached to the predicted disease. Dahiwade et al. [9] compared the results between KNN and CNN algorithm in terms of processing time and accuracy. The accuracy and processing time of CNN were 84.5% and 11.1 seconds, respectively. The statistics proved that KNN algorithm is under performing compared to CNN algorithm. In light of this study, the findings of Chen et al. [10] also agreed that CNN outperformed typical supervised algorithms such as KNN, NB, and DT. The authors concluded that the proposed model scored higher in terms of accuracy, which is explained by the capability of the model to detect complex nonlinear relationships in the feature space. Moreov

III. EXISTING SYSTEM

Prediction using a traditional disease risk model usually involves a machine learning and supervised learning algorithm which uses training data with the labels for the training of the models. High-risk and Low-risk patient classification is done in group test sets. But these models are only valuable in clinical situations and are widely studied. A system for sustainable health monitoring using smart clothing by Chen et.al. thoroughly studied heterogeneous systems and was able to achieve the best results for cost minimization on the tree and simple path cases for heterogeneous systems.

The information of patient's statistics, test results, and disease history is recorded in EHR which enables the identification of potential data-centric solutions which reduce the cost of medical case studies. Bates et al. propose six applications of big data in the healthcare field. Existing systems can predict the diseases but not the subtype of diseases. It fails to predict the condition of people.
The predictions of diseases have been non-specific and indefinite

IV. PROPOSED SYSTEM

In this paper, we have combined the structure and unstructured data in healthcare fields that let us assess the risk of disease. The approach of the latent factor model for reconstructing the missing data in medical records which are collected from the hospital. And by using statistical knowledge, we could determine the major chronic diseases in a particular region and particular community. To handle structured data, we consult hospital experts to knowuseful features.

In the case of unstructured text data, we select the features automatically with the help of the k-mean algorithm. We propose a k-mean algorithm for both structured and unstructured data.

V. THE K-MEANS ALGORITHM

The k-means algorithm is a simple iterative method to partition a given dataset into a specified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1, . . . , N}, where xi □ Rd denotes the ith data point. The algorithm is initialized by picking k points in Rd as the initial k cluster. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data, or perturbing the global mean of the data ktimes.

VI. SYSTEM ARCHITECTURE

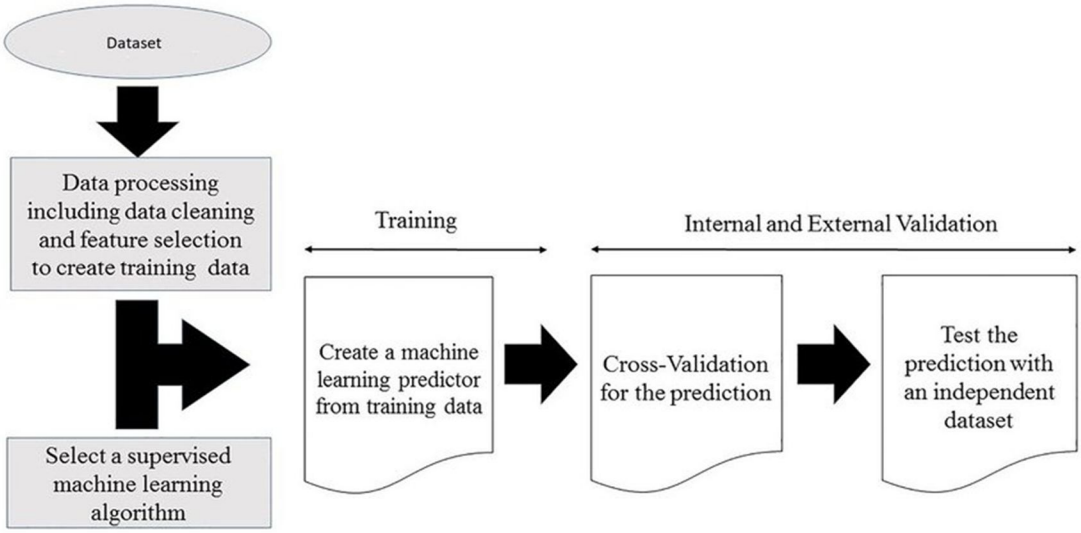


Fig .1: System Architecture

VII. RESULTS

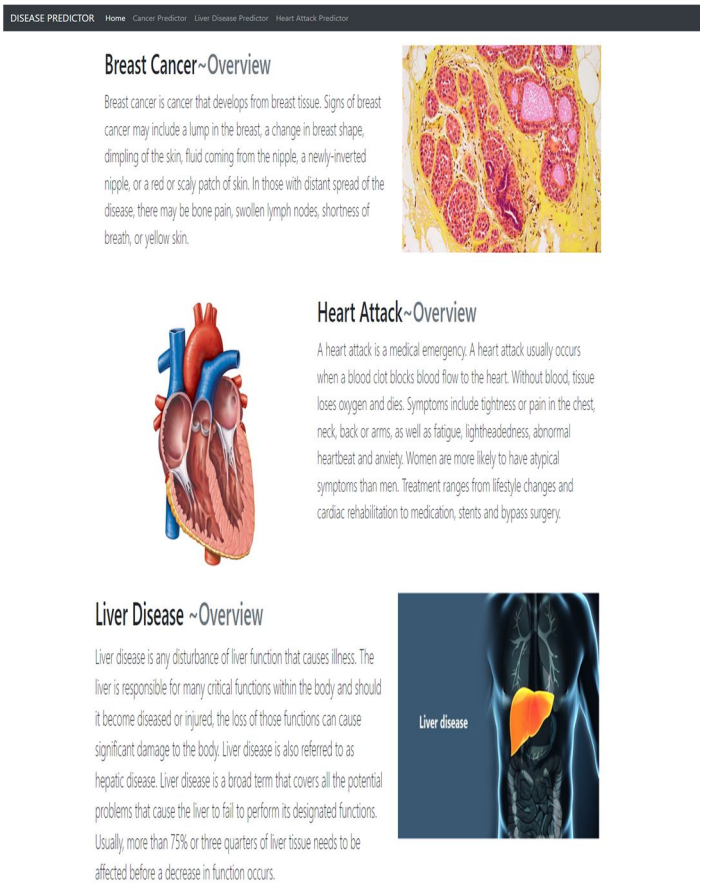


Fig.2: Home screen

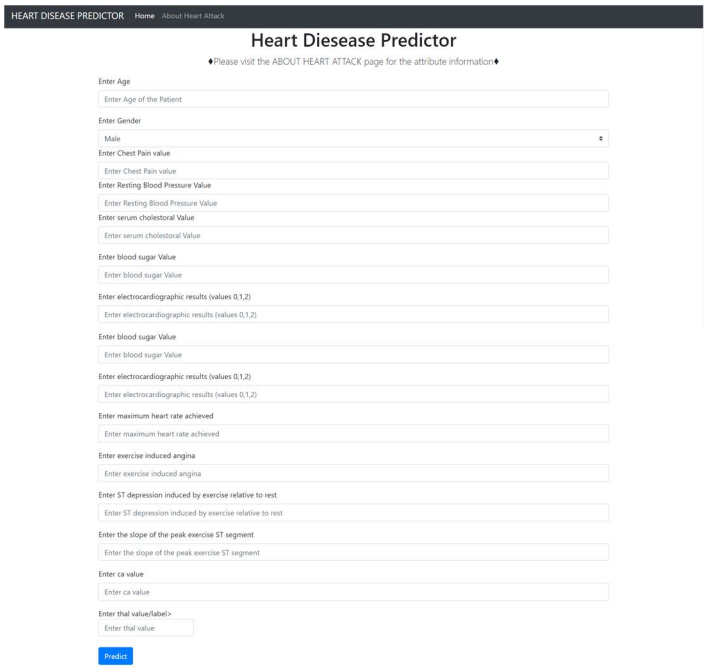


Fig.3: Heart disease prediction

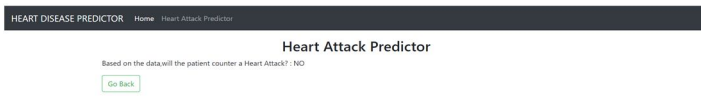


Fig.4: Results page

VIII. CONCLUSION

With the proposed system, higher accuracy can be achieved.We not only use structured data, but also the text data of thepatient based on the proposed k-mean algorithm. To find that out, we combine both data, and the accuracy rate can bereached up to 95%. None of the existing systems and work isfocused on using both the data types in the field of medicalbig data analytics. We propose a K-Mean clustering algorithm for both structured and unstructured data. The disease risk model is obtained by combining both structured and unstructured features.

IX. ACKNOWLEDGEMENT

We would like to express our sincere thanks to Mr. Dr. K.S. Reddy, Professor & Head of the Department of Information Technology Anurag University Hyderabad, whose motivation in the field of software development has made us overcome all hardships during our study and successful completion of the project.

We would like to express our profound sense of gratitude to all our faculty members for having helped us in completing this dissertation. We would like to express our deep-felt gratitude and sincere thanks to our guide K. Praveen Kumar, Assistant Professor, Department of Information Technology Anurag University Hyderabad, for his skillful guidance, and timely suggestions and encouragement in completing this project.

Finally, we would like to express our heartfelt thanks to our parents who were very supportive both financially and mentally, and for their encouragement to achieve our set goals.

REFERENCES

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in healthcare: using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

[2] K.R.Lakshmi, Y.Nagesh, and M.VeeraKrishna, "Performance comparison of three data mining techniques for predicting kidney disease survivability", International Journal of Advances in Engineering & Technology, Mar. 2014.

[3] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis of the Occurrence of Heart Disease Using Data Mining Techniques", International Journal ofPure and Applied Mathematics, 2018.

[4] Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Predictionand Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164-168.

[5] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renalfunction deterioration," Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.

[6] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra,"A proposed model for lifestyle disease prediction using support vector machine," in 2018 9th International Conference on Computing, Com- munication and Networking Technologies (ICCCNT), 2018, pp. 1–6.

[7] F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Confer- ence on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907.

[8] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning ma- chine (elm) technique for heart disease diagnosis," in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015,pp. 1–3.

[9] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.

[10] Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatolog-ical disorders through image-processing and machine learning," in 2017 International Conference on Intelligent Sustainable Systems (ICISS),2017, pp. 1047–1051.

[11] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019.

[12] R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in 2020 International Conferenceon Electronics and Sustainable Communication Systems (ICESC), 2020,pp. 302–305.

[13] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in 2018 4th International Conference on Computing Com- munication and Automation (ICCCA), 2018, pp. 1–4.

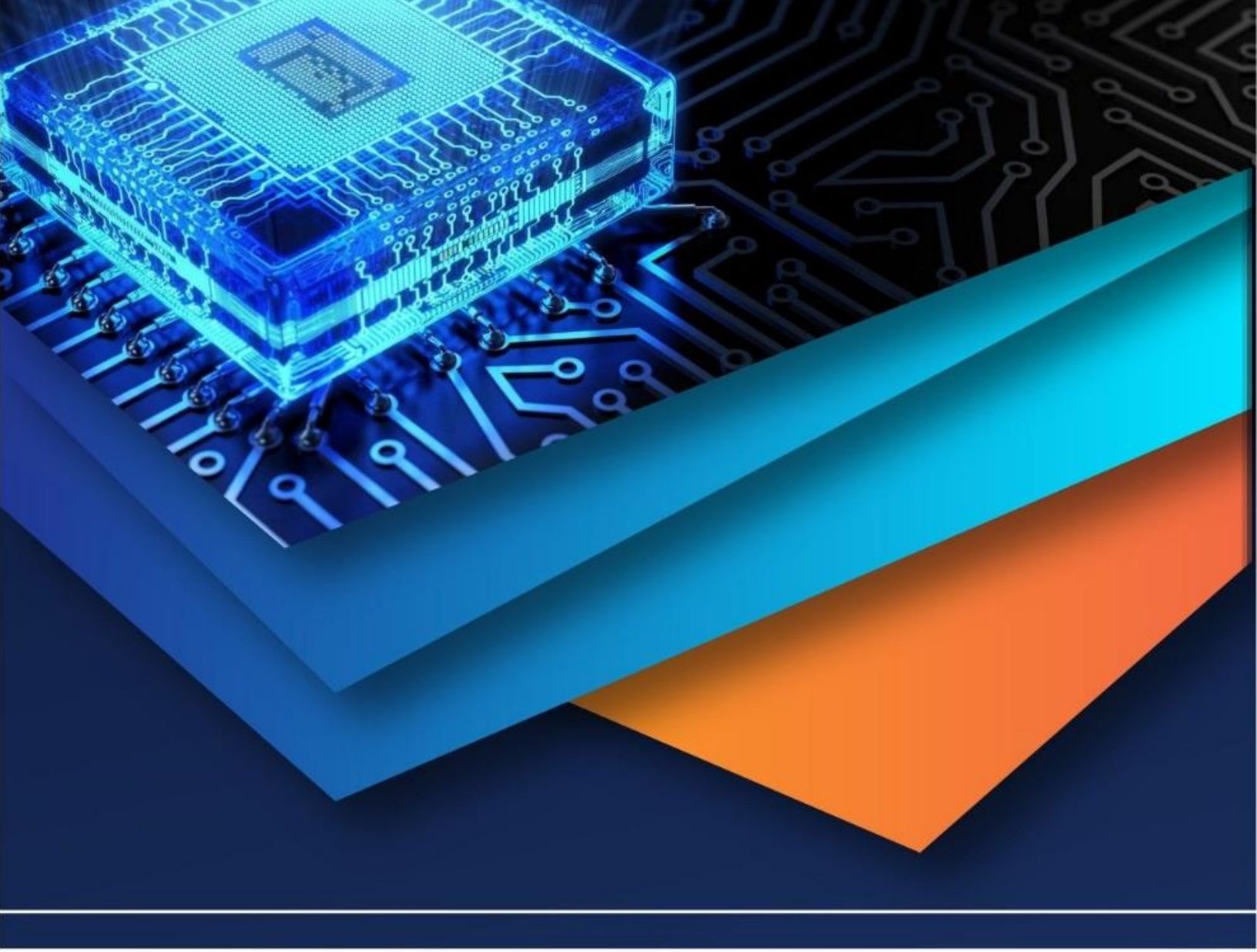
AUTHORS DETAILS

First Author – K Praveen Kumar
Assistant Professor
Department of Information Technology
Anurag University Hyderabad
Email id: praveenit@cvsr.ac.in
https://orcid.org/**0000-0002-8378-4191**

Second Author – A Pravalika
B.Tech IV / II Semester
Department of Information Technology
Anurag University Hyderabad
Email id: pravalika.amadapaku12@gmail.com

Third Author – R Pallavi Sheela
B.Tech IV / II Semester
Department of Information Technology
Anurag University Hyderabad
Email id: iampallavisheela3@gmail.com

Fourth Author – Y Vishwam
B.Tech IV / II Semester
Department of Information Technology
Anurag University Hyderabad
Email id: vishwam2022@gmail.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY