International Journal of Applied Engineering Research

Chat Bot Design Using NLP & KNN Algorithm

Mr. P.V.M Vijay Bhasker[1], Mr. K. Hari Krishna[2],

Mrs. A.Sirisha[3], Mr. V. Madhava Reddy[4],

[1,2,3,4]Assistant Professor, Department of ECE, QISCET

Email:pvm.vijaybhasker@qiscet.edu.in

Abstract:- It's becoming increasingly common for chatbots to take over tasks that were once handled exclusively by humans, such as those of online customer service agents and teachers. Since their inception with rule-based chatbots and continuing into the present day of rapid advancements in artificial intelligence (AI), the capabilities of chatbots have steadily grown. These days, chatbots may learn from their interactions with users and mimic human conversation. This project's goal is to create a conversational robot based on openai's GPT chat platform. This novel technique has the potential to redirect future studies of chatbots. In light of the findings of our investigation, we provide several suggestions for further study.

Key words:- Chat bots, ML, KNN & NLP.

1. INTRODUCTION

Conversational Agents (CAs) like as Siri, Google Assistant, and Alexa are included in the majority of smartphones and tablets. The firms that create these CAs and, more significantly, the people who use them are aware of the anthropomorphic goal: conversing with a CA must feel like conversing with another human. Depending on how they are utilized, CAs are categorized as speech-based (like Siri and Alexa) or text-messaging (like Google Assistant and Messenger bots). The term "chatbot" is often used to refer to CAs that interact through the exchange of messages. In 2016, there was a great deal of interest in the concept that consumers could "text" intelligent business agents on their smartphones, as they do with their friends and family. Technology firms have hurried to develop platforms for the development of chatbots that can comprehend genuine speech (such as Facebook Messenger, IBM Watson Conversation, and api.ai). Consequently, several chatbots have been developed lately. For example, more than one hundred thousand have been created exclusively using Facebook's Chat function. Chatbots are used for a number of purposes, including buying pizza (Domino's) and shopping (Burberry), connecting people (Chatible) and booking flights (Kayak), conversing (Pandorabots) and reading the news (NewsBot) (CNN). Developers are transitioning from an app-first design strategy, in which each programme has its own functionality and a tiny learning curve, to a chatbot-first design style, which employs a familiar messaging interface [1].

One of the major problems of artificial intelligence (AI) is equipping robots with natural language communication. Early conversational systems like as Eliza, Parry, and Alice were created to imitate human behavior in text-based conversations, enabling them to pass the Turing test in a limited capacity. Even though these systems, which predated current social chatbots, were extremely successful, the bulk of their rules were manually established. Consequently, they are only effective in certain situations. [2]

Human vision has the advantage of being taught throughout a lifetime to distinguish between objects, their distance, if they are moving, and whether an image is inaccurate. Computer vision trains

computers to execute these things more quicker by using cameras, data, and algorithms instead of retinas, optic nerves, and a visual brain. Because a system designed to check items or monitor a manufacturing asset may analyze hundreds of products or processes every minute and detect minute flaws or problems, it can soon outperform human skills. [3]

The term "artificial intelligence" was invented by John McCarthy, an American computer scientist, during The Dartmouth Conference in 1956, marking the beginning of the science. Robotics nowadays includes anything from robotic process automation to real robots. It has grown in prominence in recent years, in part because organizations are collecting more, bigger, and more diversified forms of data. AI can spot patterns in data faster and more precisely than humans, helping businesses to learn more from their data.

Machine learning (ML) is a field of study that focuses on the comprehension and development of systems that "learn," or use data, to improve their performance on a set of tasks. Assumed to be a component of artificial intelligence. Machine learning algorithms use sample data, known as "training data," to build a model that they may then use to make predictions or judgements on their own. Machine learning approaches are employed in a variety of sectors, including health, email filtering, speech recognition, and computer vision, which focuses on producing predictions using computers. However, not all machine learning is statistical learning. The methodology, theory, and application fields derived from the study of mathematical optimization may be used in machine learning. Uninstructed data analysis is the subject of the related branch of research known as data mining. Some applications for machine learning The primary objective of this project is to develop an intelligent attendance system using face recognition. Therefore, we must achieve the necessary objectives while gathering all the necessary information along the route. Using the appropriate algorithm for each stage, this approach requires step-by-step setup. First, we must get the picture's students, and then we must process the image. Second, you must train all of the captured photos. After that, you must take the pupils' attendance and record the information on a page. Using data and neural networks in a manner comparable to that of a biological brain. When used to corporate problem-solving, machine learning is also known as predictive analytics. The subfield of AI known as "Machine Learning" [4].

2. RELATED WORK & RESEARCH

A variety of papers from the aforementioned sources of literature have been reviewed to determine the most essential chatbot components. We evaluated the contents of the following reference publications. This procedure consisted merely of examining the papers and determining their primary goals. Following this, a thorough evaluation of the reference papers was undertaken to determine the primary elements of chatbots that have been the subject of previous study.

This feature includes all papers that provided a comprehensive description of the evolution of chatbots across time. This category is essential since it helped us comprehend the trends and technologies that arose or were abandoned throughout time, reflecting the chatbot's progress. It also helped us understand how and why chatbots originated, as well as how their applications and functions evolved over time.

This section contains all studies that offered a thorough account of the development of chatbots across time. This category is crucial since it helped us understand the trends and technologies that evolved or were abandoned throughout time, representing the evolution of the chatbot. It also helped

us comprehend the genesis and purpose of chatbots, as well as the evolution of their applications and functions.

Specific domain applications for chatbots, include education, banking, customer service, and psychology. The papers in this area helped us link information from other categories and obtain a better understanding of which models and attributes are used for specific applications to achieve particular goals.

The categorization of chatbots is dependent on the data set used to train machine learning algorithms for language model development. These categories were often referenced in the titles and abstracts of the papers we examined. It may be claimed that these categories are popular in the literature because they are essential to the design of chatbots. In actuality, the creation of chatbot applications necessitates the examination of several implementations and other applications (which are the result of chatbots' growth through time), in addition to a dataset and an assessment model. Every new component may be classed as a subset of one of these major categories.

3.PROPOSED METHOD:

The proposed method used to design chatbot according to our requirements. The block diagram of proposed method

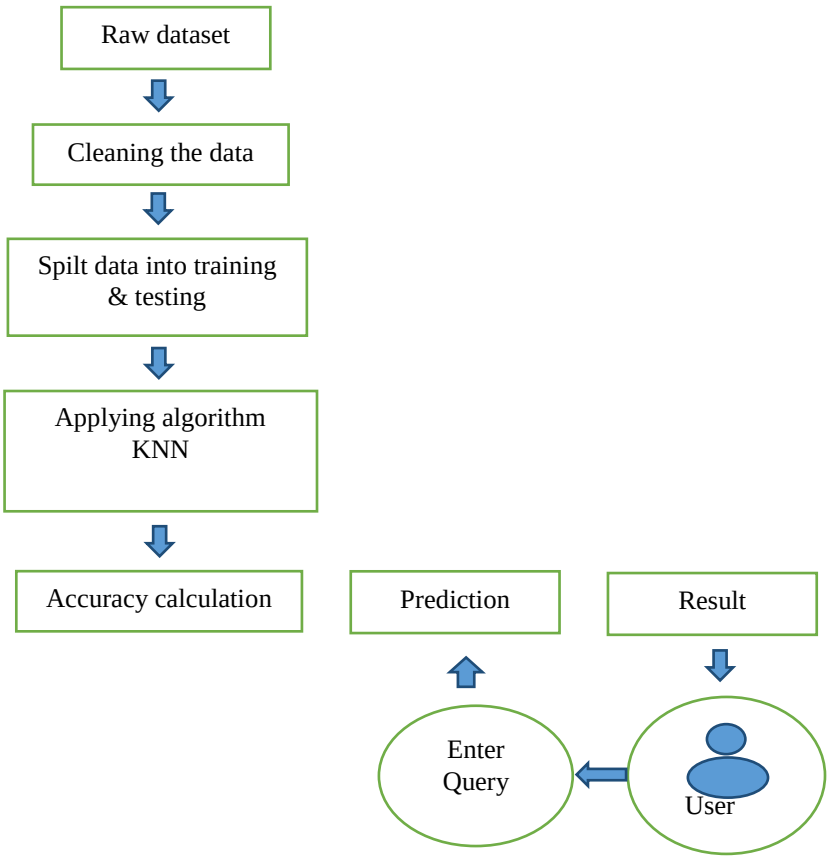


Figure 3.1 . Block Diagram of Proposed Method.

These are the steps that will be followed in this project

During the admissions process, a dataset of candidates' responses to commonly requested questions was compiled. I visited various college websites in search of commonly asked questions, from which I developed the following inquiries. Using an Excel sheet, the dataset has three columns and up to 2020 rows. There are eleven unique categories for the questions and answers. Machine learning is dependent upon data. The data will lack organization.

Pre-processing: format; consequently, we must convert to structured data in order for the data to be clearly identifiable and inserted into the database. Numerous libraries are introduced during this pre-processing; here, we used the well-known library NLTK, commonly known as a natural language tool kit. It assists the computer in understanding and interpreting written text by techniques including as tokenization, pos, stemming, and name entity recognition.

Depending on the data and model used for feature extraction, several vectorization methods are used. Text is transformed to vectors using tfidf so that machines can understand it. The TF-IDF algorithm is used to find word predictions or the similarity between words. TF- term frequency IDF- inverse document frequency. T-term refers to the term weighting of frequent terms in information retrieval, which is mostly used for document categorization (word). Document number and number of corpus (set of words). dfit) occurrence of articles linked to tin, whole corpus.

K Nearest Neighbor (KNN) is an algorithm that, when implemented, follows this supervised learning technique. The approach sets the training predictors or feature vectors in a multidimensional feature space during training. Consider it a two-dimensional space for the purpose of simplicity, a 2D graphic with X and Y axes. The picture below depicts a two-dimensional feature space, with training predictors belonging to either the A or B class. To predict the class of a new feature vector, the approach computes the distance (let's say the Euclidean distance) between the features and all training data in the dimensional space. After calculating these distances, the K closest observations must be identified. The algorithm will then assign the feature vector to the class with the highest frequency among the K closest observations.

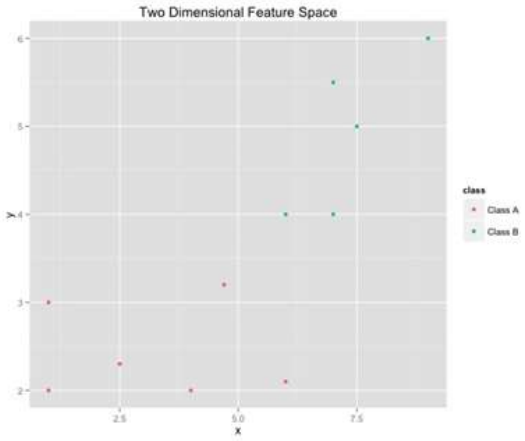


Figure 3.2.Graph of Two-Dimensional Feature Space.

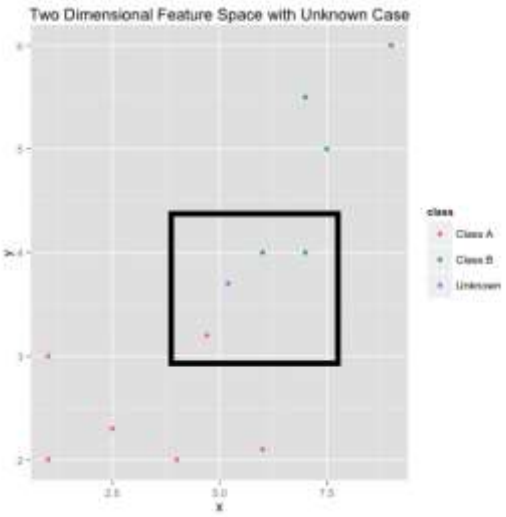


Figure 3.3. Graph of Two-Dimensional Feature Space with Unknown Case.

The preceding and previous images share the same feature space. Suppose we desire to estimate the label associated with the blue data point and K equals 3. To do this, the algorithm picks the three data points closest to the blue point — the points in the square — and finds the class with the highest frequency among these points. Given that two of the points belong to the B class and one to the A class, the class in this case is green or B.

In addition to providing training, KNN bot may also give insight into the model. Using the command status, for example, displays several pieces of information regarding the model and training, including the number of training cases inputted, the different classes observed during training and their frequency, the value of K, the status of the bot, and a 2D plot displaying feature space.

4.RESULTS

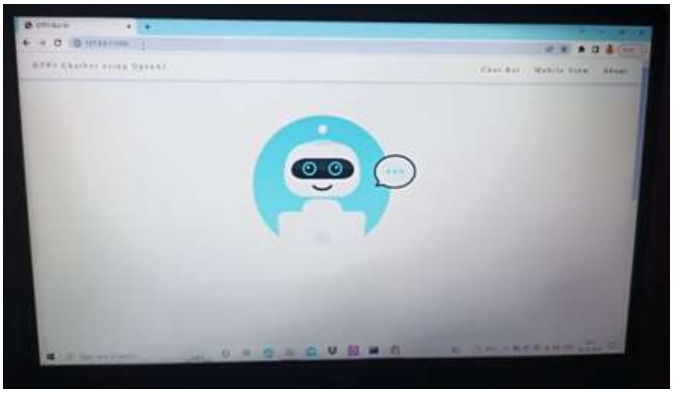


Fig 4.1. Opening URL in web browser.

Interface of the chat bot for communication. Here we communicate with chat bot by sending messages.

Figure 4.2. Interface of the chat bot for communication.

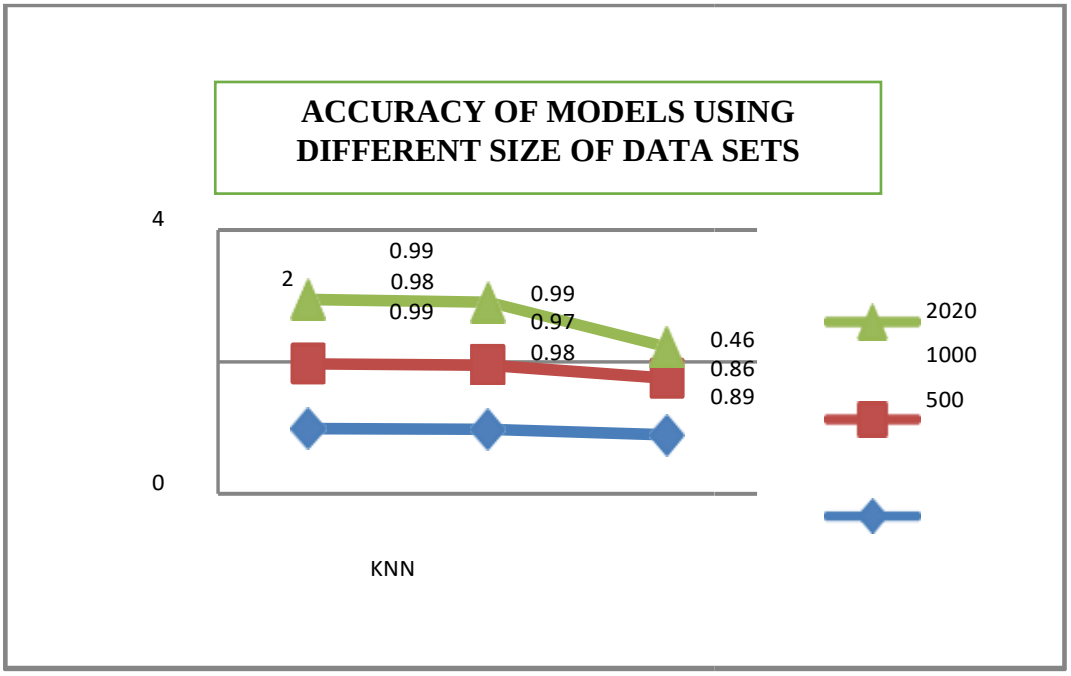
Interaction with chat bot. here we sent a message to chat bot and waiting for reply.

Figure 4.3. Interaction with chat bot.

Chat bot interaction with user. Here we can see that chat bot is giving reply to user.

Figure 4.4. Chat bot interaction with user.

Graphs:



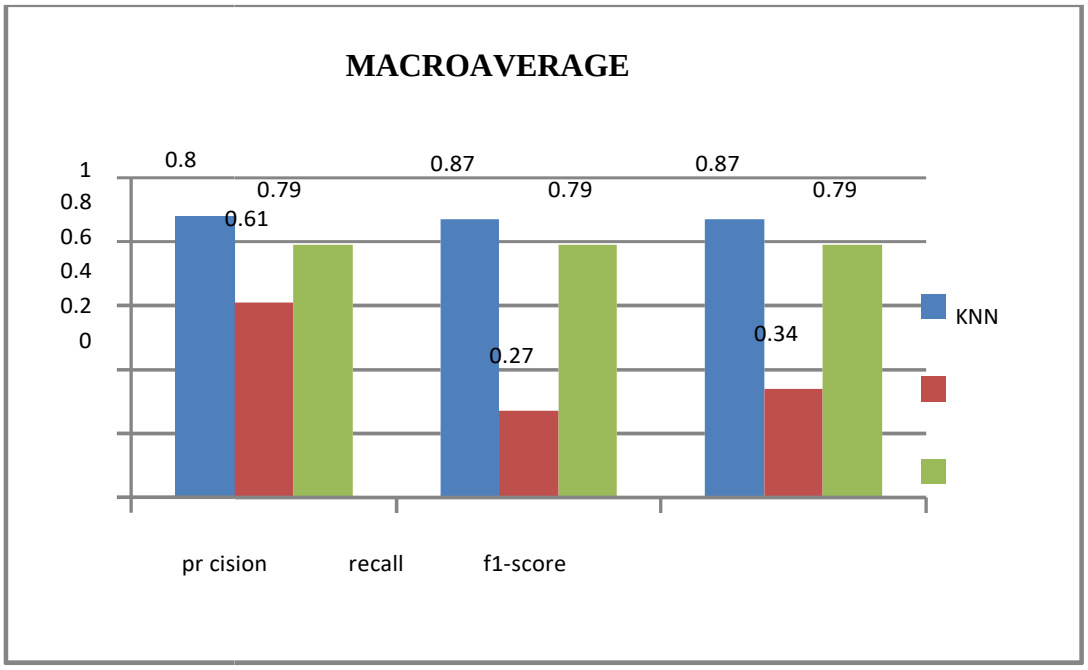


Table 4.1:Accuracy

Test	Set Accuracy
	KNN
500	0.89

1000	0.86
2020	0.46

Table 4.2 Showing testing accuracy of different size data set for KNN.

S.No	Algorithm	Accuracy
1	KNN	0.463510481

5. CONCLUSION

Chatbots have taken the place of apps. This project, as stated in prior deliverables, adds chatbot capacity to Yioop to increase its usability. Chatbots in Yioop may provide certain interactions a more human-like aspect and make them more interesting. And their primary function is to offer information and execute tasks for those with whom they engage. For all deliverables, this feature is implemented and provided in the Yioop source code. Using the aforementioned deliverables, I was able to include a simple chatbot into Yioop. I.e., configuring and creating accounts for both users with bot settings, as discussed in deliverable 2, activating a bot whenever a user requests it via a post in a thread, and implementing a simple weather chatbot that provides weather information whenever a user requests it, as discussed in deliverable 4, as shown in Fig. 3. I want to improve on the mechanism created in CS298. The next step in the evolution of chatbots is to assist people in interacting with computers using natural language or a set of rules. Future Yioop chatbots will be able to remember and learn from previous talks in order to provide more accurate responses in the future. The complexity would come from interacting with a large number of human and bot users.

ACKNOWLEDGEMENT

The author want to extend the thanks to department of Electronics & Communication Engineering and QIS Management.

REFERENCES

[1] Toth, B.P. and Czeba, B., 2016, September. Convolutional Neural Networks for Large-Scale Bird Song Classification in Noisy Environment. In CLEF (Working Notes) (pp. 560-568).

[2] Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing, 2007(1), pp.64-64.

[3] Pradelle, B., Meister, B., Baskaran, M., Springer, J. and Lethin, R., 2017, November. Polyhedral Optimization of TensorFlow Computation Graphs. In 6th Workshop on Extreme-scale Programming Tools (ESPT-2017) at The International Conference for High Performance Computing, Networking, Storage and Analysis (SC17).

[4] Cireşan, D., Meier, U. and Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. arXiv preprint arXiv:1202.2745.

[5] Andr´eia Marini, Jacques Facon and Alessandro L. Koerich Postgraduate Program in Computer Science (PPGIa) Pontifical Catholic University of Paran´a (PUCPR) Curitiba PR, Brazil 80215–901 Bird Species Classification Based on Color Features

[6] Image Recognition with Deep Learning Techniques ANDREIPETRU BĂRAR, VICTOR-EMIL NEAGOE, NICU SEBE Faculty of Electronics, Telecommunications & Information Technology Polytechnic University of Bucharest.

[7] Exception: Deep Learning with Depth wise Separable Convolutions François Chollet Google, Inc.

[8] Zagoruyko, S. and Komodakis, N., 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

[9] Inception-v4,Inception-ResNetand the Impact of Residual Connections on Learning Christian Szegedy, Sergey Effervescent Vanhoucke ,AlexandrrA.Alemi

[10] Stefan Kahl, Thomas Wilhelm-Stein, Hussein Hussein, Holger Klinck, Danny Kowerko, Marc Ritter, and Maximilian Eibl Large-Scale Bird Sound Classification using Convolutional Neural Networks.

11. **Dr**.R.Chinnaiyan , M.S.Nidhya (2018), " Reliability Evaluation of Wireless Sensor Networks using EERN Algorithm" , Lecture Notes on Data Engineering and Communications Technologies, Springer International conference on Computer Networks and Inventive CommunicationTechnologies (ICCNCT - 2018), August 2018 (Online)
12. Dr.R.Chinnaiyan , R.Divya (2018), " Reliable AI Based Smart Sensors for Managing Irrigation Resources in Agriculture" , Lecture Notes on Data Engineering and Communications Technologies, Springer International conference on ComputerNetworks and Inventive CommunicationTechnologies (ICCNCT - 2018), August 2018 (Online)
13. Dr.R.Chinnaiyan , S.Balachandar (2018) , " Reliable Digital Twin for Connected Footballer" , Lecture Notes on Data Engineering and Communications Technologies, Springer International conference on Computer Networks and Inventive Communication Technologies(ICCNCT - 2018), August 2018 (Online)
14. Dr.R.Chinnaiyan , S.Balachandar (2018) , " Centralized Reliability and Security Management of Data in Internet of Things (IoT) with Rule Builder"
, Lecture Notes on Data Engineering and Communications Technologies, Springer International conference on Computer Networks and Inventive Communication Technologies(ICCNCT - 2018), August 2018 (Online)
15. Dr.R.Chinnaiyan, Abishek Kumar (2017) " Reliability Assessment of Component Based Software Systems using Basis Path Testing" , IEEE International Conference on Intelligent Computing and Control Systems, ICICCS 2017, 512 – 517
16. Dr.R.Chinnaiyan, AbishekKumar(2017)
,"Construction of Estimated Level Based Balanced Binary Search Tree", 2017 IEEE International Conference on Electronics,Communication, and Aerospace Technology (ICECA 2017), 344 - 348,
978-1-5090-5686-6.
17. Dr.R.Chinnaiyan, AbishekKumar(2017), Estimation of Optimal Path in Wireless Sensor Networks based on Adjancy List, 2017 IEEE International Conference on Telecommunication,Power Analysis and Computing Techniques (ICTPACT2017) ,6,7,8th April 2017,IEEE 978-1-5090-3381-2.
18. Dr.R.Chinnaiyan, R.Divya (2017)," Reliability Evaluation of Wireless Sensor Networks", IEEE International Conference on Intelligent Computing and Control Systems, ICICCS 2017, 847 – 852
19. Dr.R.Chinnaiyan, Sabarmathi.G (2017)," Investigations on Big Data Features , Research Challenges and Applications", IEEE International
Conference on Intelligent Computing and Control Systems, ICICCS 2017, 782 – 786
20. G.Sabarmathi , Dr.R.Chinnaiyan (2018), "Envisagation and Analysis of Mosquito Borne Fevers – A Health Monitoring System by Envisagative Computing using Big Data Analytics" in ICCBI 2018 – Springer on 19.12.2018 to 20.12.2018 (Recommended for Scopus Indexed Publication IEEE Xplore digital library)
21. G.Sabarmathi , Dr.R.Chinnaiyan, Reliable Data Mining Tasks and Techniques for Industrial Applications, IAETSD JOURNAL FOR ADVANCED RESEARCH IN APPLIED SCIENCES, VOLUME 4, ISSUE 7, DEC/2017,PP- 138-142, ISSN NO: 2394-8442
22. Dr. M. Thangamani, Jafar Ali Ibrahim, Information Technology E-Service Management System, International Scientific Global Journal in Engineering Science and Applied Research (ISGJESAR). Vol.1. Issue 4, pp. 13-18, 2017. http://isgjesar.com/Papers/Volume1,Issue4/paper2.pdf
23. Ibrahim, Mr S. Jafar Ali, K. Singaraj, P. Jebaroopan, and S. A. Sheikfareed. "Android Based Robot for Industrial Application." International Journal of Engineering Research & Technology 3, no. 3 (2014).
24. Ibrahim, S. Jafar Ali, and M. Thangamani. "Momentous Innovations in the Prospective Method of Drug Development." In Proceedings of the 2018 International Conference on Digital Medicine and Image Processing, pp. 37-41. 2018.
25. Ibrahim, S. Jafar Ali, and M. Thangamani. "Prediction of Novel Drugs and Diseases for Hepatocellular Carcinoma Based on Multi-Source Simulated Annealing Based Random Walk." Journal of medical systems 42, no. 10 (2018): 188. https://doi.org/10.1007/s10916-018-1038-y ISSN 1311-8080, https://acadpubl.eu/hub/2018-119-16/1/94.pdf

26. Jafar Ali Ibrahim. S, Mohamed Affir. A “Effective Scheduling of Jobs Using Reallocation of Resources Along With Best Fit Strategy and Priority”, International Journal of Science Engineering and Advanced Technology(IJSEAT) – ISSN No: 2321- 6905, Vol.2, Issue.2, Feb-2014, http://www.ijseat.com/index.php/ijseat/article/view/62
27. M. Thangamani, and Jafar Ali Ibrahim. S, "Knowledge Exploration in Image Text Data using Data Hiding Scheme," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2018, 14-16 March, 2018, Hong Kong, pp352-357 http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp352-357.pdf
28. M. Thangamani, and Jafar Ali Ibrahim. S, "Knowledge Exploration in Image Text Data using Data Hiding Scheme," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2018, 14-16 March, 2018, Hong Kong, pp352-357 http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp352-357.pdf
29. S. Jafar Ali Ibrahim and M. Thangamani. 2018. Momentous Innovations in the Prospective Method of Drug Development. In Proceedings of the 2018 International Conference on Digital Medicine and Image Processing (DMIP '18). Association for Computing Machinery, New York, NY, USA, 37–41. https://doi.org/10.1145/3299852.3299854
30. S. Jafar Ali Ibrahim and Thangamani, M “Proliferators and Inhibitors Of Hepatocellular Carcinoma”, International Journal of Pure and Applied Mathematics (IJPAM) Special Issue of Mathematical Modelling of Engineering Problems Vol 119 Issue. 15. July 2018
31. Thangamani, M., and S. Jafar Ali Ibrahim. "Ensemble Based Fuzzy with Particle Swarm Optimization Based Weighted Clustering (Efpso-Wc) and Gene Ontology for Microarray Gene Expression." In Proceedings of the 2018 International Conference on Digital Medicine and Image Processing, pp. 48-55. 2018. https://dl.acm.org/doi/abs/10.1145/3299852.3299866
32. Testing” , IEEE International Conference on Intelligent Computing and Control Systems, ICICCS 2017, 512 – 517
33. Dr.R.Chinnaiyan, Abishek Kumar(2017) ,”Construction of Estimated Level Based Balanced Binary Search Tree”, 2017 IEEE International Conference on Electronics,Communication, and Aerospace Technology (ICECA 2017), 344 - 348, 978-1-5090-5686-6.
34. R.Chinnaiyan, S.Somasundaram (2012) , Reliability Estimation Model for Software Components using CEP”, International Journal of Mechanical and Industrial Engineering (IJMIE) , ISSN No.2231-6477, Volume-2, Issue-2, 2012, pp.89-93.
35. R.Chinnaiyan, S. Somasundaram (2011) ,”An SMS based Failure Maintenance and Reliability Management of Component Based Software Systems”, European Journal of Scientific Research, Vol. 59 Issue 1, 9/1/2011, pp.123 (cited in EBSCO, Impact Factor: 0.045)
36. R.Chinnaiyan, S.Somasundaram(2011), “An Experimental Study on Reliability Estimation of GNU Compiler Components - A Review”, International Journal of Computer Applications, Vol.25, No.3, July 2011, pp.13-16. (Impact Factor: 0.814)
37. R.Chinnaiyan, S.Somasundaram(2010) “Evaluating the Reliability of Component Based Software Systems “ ,International Journal of Quality and Reliability Management , Vol. 27, No. 1., pp. 78-88 (Impact Factor: 0.406)
38. Dr.R.Chinnaiyan, Abishek Kumar(2017), Estimation of Optimal Path in Wireless Sensor Networks based on Adjancy List, 2017 IEEE International Conference on Telecommunication,Power Analysis and Computing Techniques (ICTPACT2017) ,6,7,8th April 2017,IEEE 978-1-5090-3381-2.