Disease Identification using Machine Learning and NLP

[1] Akila S, [2]Prakash J, [3]Dr Uma S,
[1] Research Scholar, [2]Assistant Professor, [3]Research coordinator
[1-3]Hindusthan College of Engineering and Technology, Coimbatore.
[1] selvamakila2098@gmail.com

Abstract: Artificial Intelligence (AI) technologies are now widely used in a variety of fields to aid with knowledge acquisition and decision-making. Health information systems, in particular, can gain the most from AI advantages. Recently, symptoms-based illness prediction research and manufacturing have grown in popularity in the healthcare business. Several scholars and organisations have expressed an interest in applying contemporary computational tools to analyse and create novel approaches for rapidly and accurately predicting illnesses. In this study, we present a paradigm for assessing the efficacy of combining Machine Learning (ML) and Natural Language Processing (NLP) technologies in a disease prediction system. We scraped a disease-symptom dataset with NLP characteristics from one of the UK's most trusted National Health Service (NHS) websites as an example. In addition, we will thoroughly examine our data using symptom frequency, similarity, and clustering analysis. As a consequence, we can observe that the forecast has a high efficiency rate, but there are still some challenges to work out.

Keywords - Artificial Intelligent, Data Analysis, Machine Learning, Nature Language Processing, Health Information System, Symptoms, Disease Prediction, Symptom Frequency.

INTRODUCTION

Nowadays, there are an increasing number of medical and healthcare-related items available through mobile and online applications.

Corresponding Author: *Akila S*
Hindusthan College of Engineering and Technology.
Mail: selvamakila2098@gmail.com

Many machine learning research projects begin to leverage information gathered from internet platforms such as social media, forum conversations, and a variety of other resources to construct AI-supported healthcare suggested apps. The study's findings were highly encouraging, and these AI applications can deliver useful hints or even pre-diagnose advise based on relatively simple facts, such as condition and symptom relational databases. Moreover, the majority of the work lacks a straightforward framework for maintaining data and applying ML algorithms on trustworthy data. Predicting cancer using machine learning and clustering, forecasting dermatological disorders using a naive Bayes classifier, and predicting the occurrence of swine flu using a naive Bayes classifier have all been studied in the past, and these techniques appear to generate pretty decent results. However, the majority of these studies focus on one or a few diseases or medical problems that are critical in the healthcare profession [1-12]. However, several studies provided novel approaches for predicting more illness based on symptoms.

Machine learning (ML) is the empirical study of algorithms and statistical models that computer systems use to do a certain task efficiently without utilising explicit instructions, instead relying on models and inference [13-22]. It is thought to be a subset of artificial intelligence. Machine learning algorithms create a mathematical model using sample data, referred to as "training data," in order to make predictions or choices without being expressly programmed to do so. Machine learning methods are utilised in applications such as email filtering, network intrusion detection, and computer vision, when developing an algorithm containing explicit instructions for accomplishing the task is impractical [23-34].

PROPOSED SYSTEM

Based on the symptoms presented, our suggested framework will study, analyse current solutions, and design an efficient framework that can be utilised to anticipate general illnesses or medical situations [35-38]. To forecast 298 distinct medical diseases, the project combines NLP and the Multinomial Nave Bayes Algorithm. The framework is made up of three primary components:

1. Data extracting and NLP processing
2. Data evaluation and Analyse
3. Prediction

PROPOSED METHOD

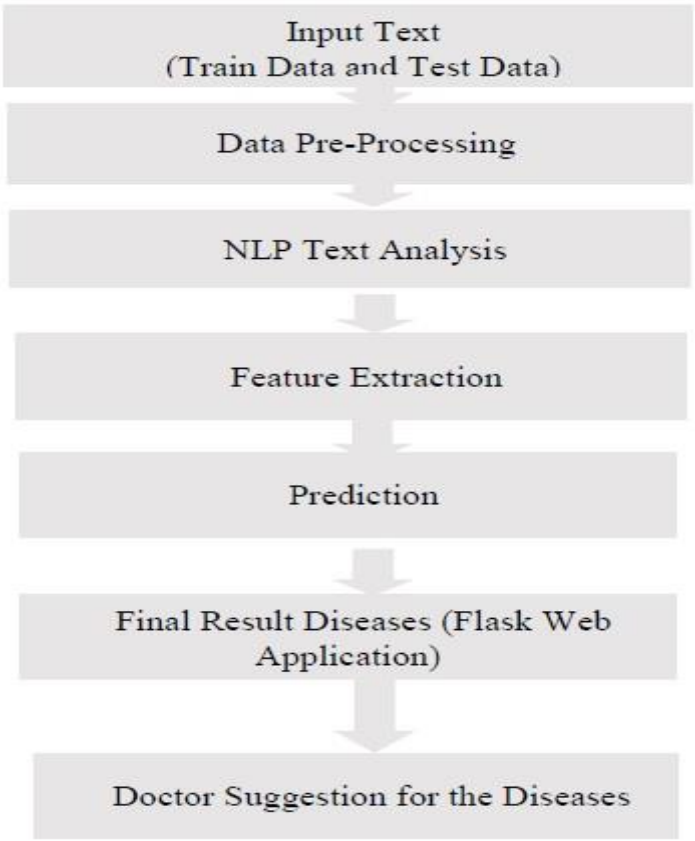

Figure 1 Overview of the Proposed Method

ALGORITHMS

Regexp Tokenizer (NLP)

Multinomial Naïve Bayes Algorithm(Classifier)

Advantage:

1. This method is fast and can save you a lot of time.

2. Naive Bayes is appropriate for multi-class prediction issues.
3. If the premise of feature independence remains true, it can outperform other models while using far less training data.
4. NLP enhanced analysis

NAIVE BAYES CLASSIFIER ALGORITHM:

1.The Nave Bayes algorithm is a supervised learning technique based on the Bayes theorem that is used to solve classification issues.

2.It is mostly utilised in text classification with a large training dataset.

3.Nave Bayes Classifier is a basic and effective Classification method that aids in the development of rapid machine learning models capable of making quick predictions.

MULTINOMIAL NAIVE BAYES

1.A supervised learning approach is required, which classifies each new document by assigning one or more class labels from a fixed or predetermined class.
2.It employs the bag of words technique, in which the individual words in the document comprise the document's characteristics and the sequence of the words is ignored.
3.This method differs from the way we speak with one another.
4.It treats language as if it were a bag full of words, with each message being a random mouthful of them.
5.Large documents include a large number of words and are often distinguished by a very high dimensionality feature space with thousands of characteristics.

TECHNOLOGIES USED

1. Machine Learning
2. Natural Language Processing (NLP)

MACHINE LEARNING

AI is the use of artificial intelligence (AI) that enables frameworks to gain and develop without being fundamentally altered. AI is concerned with the progress of computer programmes that can access information and utilise it to learn on their own. Machine Learning Methods:

ADMINISTERED AI

AI-assisted computations can apply what has been learned in the past to fresh data using designated guidance to predict future events. Beginning with an assessment of a known preparation dataset, the learning calculation provides a derived ability to anticipate the outcome values. After proper preparation, the framework may focus on any new input. The learning calculation may also compare its output to the predicted output and identify errors to adjust the model accordingly. Algorithms for solo machine learning are employed when the data used to prepare is not categorised or identified. Unaided learning focuses on how frameworks can induce the ability to depict a hidden design from unlabelled data. The circumstance does not provide the desired outcome, yet it investigates the information and can attract inductions from datasets to portray concealed structures from unlabelled information.

NATURAL LANGUAGE PROCESSING (NLP)

Natural language processing (NLP) is the capacity of a computer software to interpret spoken and written human language, often known as natural language. It's part of artificial intelligence. NLP has been around for over 50 years and has its roots in linguistics. It has a wide range of real-world applications, including medical research, search engines, and corporate intelligence. Doctor Suggestions with Flask Frame Work

Hospitals are the most common place for a sick individual to receive medical examinations, illness diagnosis, and treatment recommendations. This is something that practically everyone on the planet does. People see it as the most dependable method of determining their health state. The suggested approach is intended to provide an alternative to the traditional way of visiting a

hospital and scheduling an appointment with a doctor to obtain a diagnosis. The goal of this study is to construct a Doctor Suggestion application using natural language processing and machine learning technologies. People may engage with the Doctor suggestion exactly like they would with another human, and through a series of enquiries, the Doctor suggestion will detect the user's symptoms and thereby anticipate the outcome. predicts the disease and recommends treatment. This approach may be very useful in doing daily check-ups, making individuals aware of their health state, and encouraging them to take right precautions to stay healthy. According to the findings of this study, such a system is not commonly used, and people are unaware of it. By putting this recommended framework into action, users may skip the time-consuming process of visiting hospitals by utilising this free application from wherever they are. A wealthy society is one in which all of its citizens are healthy. If one aspires to be happy, it is critical to preserve one's health. Only a healthy body can have a healthy mind, and it improves people's performance. According to the most recent TOI [1] news, individuals place little value on their health and find it time-consuming to visit hospitals for check-ups. Health has no place in today's hectic lifestyle. The majority of people in the working class report that their hectic schedule prevents them from getting regular medical check-ups and that they ignore any discomfort displayed by their bodies until it becomes too severe. AI (ML) is the logical analysis of computations and quantifiable models that computer frameworks employ to successfully carry out a certain task without the need of explicit rules, relying on models and induction all things considered. It is regarded as a subset of artificial reasoning.

AI computations create a numerical model of test data, known as "preparing information," to make projections or decisions without being explicitly specialised to carry out the task. Machine learning calculations are used in applications such as email sorting, identifying organisational gatecrashers, and PC vision, when it is impossible to generate a calculation of clear recommendations for carrying out the task. AI is inextricably linked with computational insights, which revolve around forecasting using PCs. The study of numerical streamlining contributes tools, hypotheses, and application domains to the science of artificial intelligence. Information

mining is a branch of artificial intelligence that focuses on exploratory data analysis via unassisted learning. AI is sometimes referred to as predictive analysis in its application across business challenges. Earlier in this post, we mentioned a few applications of AI. To further understand the concept of AI, consider the following models: web query items, continuous adverts on website pages and mobile phones, email spam sifting, network interruption location, and example and picture recognition. Each of these are the outcomes of using AI to analyse massive amounts of data.

In general, information examination was shown by experimentation, a process that becomes confusing when informative collections are large and diversified. AI provides a solution to this issue by providing clever alternatives rather than studying massive amounts of data. AI may give precise conclusions and examination by developing rapid and effective computations and information-driven models for continuous information management. AI tasks are classified into a few broad categories. The computation in controlled learning assembles a numerical model of a collection of information that includes both the data sources and the desired outputs. For example, if the task was to determine whether a picture contained a specific item, the preparation information for a managed learning calculation would include pictures with and without that article (the info), and each picture would have a mark (the result) indicating whether it contained the article. In certain circumstances, the information may be only partially available or limited to unique input. Semi-directed gaining computations promote numerical models from insufficient preparation knowledge, when a portion of the example inputs suffer from the loss of the ideal outcome.

Managed learning includes grouping calculations and relapsing calculations. When the findings are limited to a certain set of data, grouping computations are used. For a message-channeling order computation, the information would be an approaching email, and the outcome would be the name of the envelope in which to record the email. For a formula that identifies spam communications, the expected outcome would be all items regarded "spam" or "not spam," as handled by the Boolean characteristics valid and misleading. Relapse computations are called

from their persistent outcomes, which means they can incorporate any value within a reach. Temperature, length, and cost of an item are examples of unending worth. Administered Machine Learning: The majority of commonsense AI makes use of controlled learning. Administered learning is a situation in which you have input factors (x) and a result variable (Y), and you use a computation to obtain planning capability from the contribution to the outcome Y = f(X). The goal is to estimate the planning capacity so accurately that when new information (x) comes in, you can predict the outcome variables (Y) for that information.

Supervised Machine Learning calculation methods include direct and strategic relapse, multi-class characterisation, Decision Trees, and backing vector machines. Regulated learning anticipates that the material needed to produce the calculation has now been marked with correct answers. A grouping computation, for example, will find out how to identify animals after being prepared on a dataset of photographs that are suitably labelled with the kind of creature and some recognising qualities. Administered learning problems can also be combined into Regression and Classification challenges. The two concerns share the goal of developing a simple model that can predict the worth of the reliant characteristic based on property parameters. The distinction between the two tasks is that the dependent quality is mathematical for relapse and straightforward for arrangement.

CLUSTERING

The K-means clustering technique is the simplest unassisted learning computation for dealing with the bunching problem. K-implies calculation divides n perceptions into k clusters, with each perception having a position in the group with the nearest mean serving as the bunch's model.

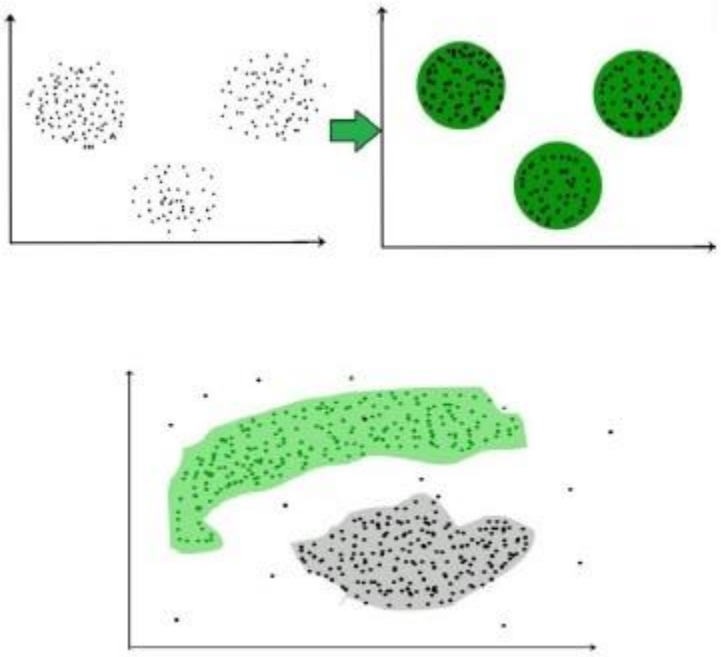

Figure 2 Clustering process

Innovative learning calculations find the best outcomes (preparation marks) for a limited set of information sources based on a budget, and enhance the selection of contributions for which it will secure preparation names. When used wisely, they can be presented to a human customer for naming. In a strong atmosphere, support learning calculations are provided criticism as sure or negative support and are used in autonomous cars or in finding out how to play a game against a human adversary. Other specific AI calculations include topic demonstrating, in which the computer software is given a set of regular language reports and searches for other archives that cover relevant themes. . In thickness evaluation concerns, AI computations may be used to monitor the unnoticeable probability thickness work. By all accounts, AI is the most obvious example. It is commonly associated with phrases referring to various logical procedures for information disclosure or expectancy (named as machine or factual learning methods). Towards Data Science provides a forum for a large number of people to exchange ideas and broaden our understanding of information science. Data science, like information mining, is an

interdisciplinary topic that use logical processes, cycles, computations, and frameworks to extract information and experiences from information in various patterns, both ordered and unstructured.

Disease identification

- Fungal infection
- Allergy
- GERD
- Chronic cholestasis
- Drug Reaction
- Peptic ulcer diseae
- AIDS
- Diabetes
- Gastroenteritis
- Bronchial Asthma
- Hypertension
- Migraine
- Cervical spondylosis
- Paralysis (brain hemorrhage)
- Jaundice
- Malaria
- Chicken pox
- Dengue
- Typhoid
- hepatitis A
- Hepatitis B
- Hepatitis C

- Hepatitis D
- Hepatitis E
- Alcoholic hepatitis
- Tuberculosis
- Common Cold
- Pneumonia
- Dimorphic hemmorhoids(piles)
- Heart attack
- Varicose veins
- Hypothyroidism
- Hyperthyroidism
- Hypoglycemia
- Osteoarthristis
- Arthritis
- (vertigo) Paroymsal Positional Vertigo
- Acne
- Urinary tract infection
- Psoriasis
- Impetigo

PRINCIPAL COMPONENT ANALYSIS

The basic idea behind principal component analysis (PCA) is to reduce the dimensionality of an informative collection composed of several aspects that are related to one another, either intensively or sensitively, while retaining the variation existing in the dataset to the greatest extent possible. The equivalent is completed by changing the factors to another arrangement of factors known as the primary parts (or essentially, the PCs) and are symmetrical, asked with the

ultimate goal that the variety existing in the initial factors lessens as we decrease in the request. As a result, the first head component contains the most variation that was accessible in the initial sections. The important components are the eigenvectors of a covariance network, and hence they are symmetrical.

Importantly, the dataset on which the PCA approach will be used should be scaled. The results are also affected by the relative scale. As a layperson, it is a method of summarising facts. Consider several wine bottles on a dining table. Each wine is defined by its attributes such as tone, strength, age, and so on. Overt repetitiveness will develop in any situation since a significant number of them will quantify comparable qualities. So, in this case, PCA will essentially sum up each wine in stock with a lower quality.

When seen from its most illuminating perspective, Principal Component Analysis can intuitively provide the client with a lower-layered image, a projection or "shadow" of this item. Information science is a "concept to unite insights, information examination, AI, and their related tactics" in order to "comprehend and examine actual oddities" using information. It employs methods and hypotheses derived from a variety of areas within the context of arithmetic, measurements, data science, and science. Information examination is the process of examining, purifying, modifying, and displaying information with the goal of discovering useful facts, highlighting aims, and providing direction. Information inquiry includes a variety of elements and methodologies, integrating multiple methods under a variety of titles and being used in a variety of business, scientific, and sociological fields.. In today's business, information evaluation plays a role in making more rational decisions and supporting the organization in achieving successful activity.

CONCLUSIONS

The development of AI in health care might encompass jobs ranging from simple to sophisticated, such as answering the phone, reviewing medical records, and analysing population health trends. We provide an experimental evaluation research approach for constructing a symptom-based

illness prediction system in this project. We gathered 298 disease information from the UK's most trusted NHS website and used powerful NLP and ML algorithms for Disease Prediction and Doctor Suggestion with the Flask framework.

REFERENCES

1. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang,"Disease prediction by machine learning over big data from healthcare communities", ," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
2. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.
3. IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.
4. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.
5. L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184– 189.
6. Disease and symptoms Dataset –www.github.com.
7. Heart disease Dataset-WWW.UCI Repository. com
8. AjinkyaKunjir, HarshalSawant, NuzhatF.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," in IEEE big data analytics and computational intelligence, Oct 2017 pp.2325.
9. ShanthiMendis, PekkaPuska, Bo Norrving, World Health Organization (2011), Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.
10. Amin, S.U.; Agarwal, K.; Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors", IEEE Conference on Information & Communication Technologies (ICT), vol., no.,pp.1227-31,11- 12 April 2013.
11. Karthick, R., Senthilselvi, A., Meenalochini, P. et al. Design and Analysis of Linear Phase Finite Impulse Response Filter Using Water Strider Optimization Algorithm in FPGA. Circuits Syst Signal Process (2022). https://doi.org/10.1007/s00034-022-02034-2
12. Meenalochini, P., and E. Sakthivel. "An efficient GBDTRSO control strategy for PV connected H-Bridge Nine Level MLI System with quasi-Z-source inverter." Applied Soft Computing 113 (2021): 108026.N. C. Mediaite. Harvard educator sounds caution on 'probable' Covid p
13. Pandemic: 40% to 70% of world could be tainted for the current year. Gotten to on 2020.02.18. [Online]. Accessible: https://www.mediaite.com/news/harvardprofessor-sounds-alert on-possible Covid pandemic-40-to-70-ofworld-could-be-contaminated for this present year/.

14. Punarselvam, E., Suresh, P., & Parthasarathy, R. (2013). Segmentation of CT scan lumbar spine image using median filter and canny edge detection algorithm. Int J Comput Sci Eng, 5, 806-814.
15. Punarselvam, E., et al. "Segmentation of Lumbar spine image using Watershed Algorithm." International Journal of Engineering Research and Applications, ISSN (2013): 2248-962
16. Punarselvam, E., et al. "Segmentation Analysis Techniques and Identifying Stress Ratio of Human Lumbar Spine Using ANSYS." Journal of Medical Imaging and Health Informatics 10.10 (2020): 2308-2315.
17. Karthick, R., et al. "Overcome the challenges in bio-medical instruments using IOT–A review."; Materials Today: Proceedings 45 (2021): 1614-1619.
18. Suresh, Helina Rajini, et al. "Suppression of four wave mixing effect in DWDM system." Materials Today: Proceedings 45 (2021): 2707-2712.
19. Soundari, D. V., et al."Enhancing network-on-chip performance by 32-bit RISC processor based on power and area efficiency." Materials Today: Proceedings 45 (2021): 2713-2720.
20. Sabarish, P., et al., Investigation on performance of solar photovoltaic fed hybrid semi impedance source converters." Materials Today: Proceedings 45 (2021): 1597-1602.
21. Karthick, R., and M. Sundararajan. "SPIDER-based out-of-order execution scheme for HtMPSOC "; International Journal of Advanced Intelligence paradigms 19.1 (2021): 28-41.
22. Karthick, R., and P. Meenalochini. "Implementation of data cache block (DCB) in shared processor using field-programmable gate array (FPGA)." Journal of the National Science Foundation of Sri Lanka 48.4 (2020).
23. Sabarish, P., et al., "An Energy Efficient Microwave Based Wireless Solar Power Transmission System" IOP Conference Series: Materials Science and Engineering. Vol. 937.No. 1. IOP Publishing, 2020
24. Vijayalakshmi, S., et al. "Implementation of a new Bi-Directional Switch multilevel Inverter for the reduction of harmonics" IOP Conference Series: Materials Science and Engineering. Vol.937. No. 1. IOP Publishing, 2020.
25. Karthick, R., and M. Sundararajan. "Design and implementation of low power testing using advanced razor based processor" International Journal of Applied Engineering Research 12.17 (2017): 6384-6390.
26. Punarselvam, E., and P. Suresh. "Non-Linear Filtering Technique Used for Testing the Human Lumbar Spine FEA Model." Journal of medical systems 43, no. 2 (2019): 1-13.
27. Punarselvam, E., Hemalatha, E., Dhivahar, J., Gowtham, V., Hari, V., & Thamaraikannan, R. PREDICTING WIRELESS CHANNELS FOR ULTRA-RELIABLE LOW-LATENCY COMMUNICATIONS.
28. Punarselvam, E., et al. "Segmentation of Lumbar spine image using Watershed Algorithm." International Journal of Engineering Research and Applications, ISSN (2013): 2248-9622.
29. Punarselvam, E., and P. Suresh. "Edge detection of CT scan spine disc image using canny edge detection algorithm based on magnitude and edge length." 3rd International Conference on Trendz in Information Sciences & Computing (TISC2011). IEEE, 2011. (2021): 4991-5004.

30. Punarselvam, Dr E., and S. Gopi. "Effective and Efficient Traffic Scrutiny in Sweet Server with Data Privacy." International Journal on Applications in Information and Communication Engineering 5.2 (2019): 1-5
31. Punarselvam, E., and P. Suresh. "Investigation on human lumbar spine MRI image using finite element method and soft computing techniques." Cluster Computing 22, no. 6 (2019): 13591-13607.
32. Karthick, R., and M. Sundararajan. “A Reconfigurable Method for Time Correlated Mimo Channels with a Decision Feedback Receiver” International Journal of Applied Engineering Research 12.15 (2017): 5234-5241.
33. Karthick, R., and M. Sundararajan. “A novel 3-D-IC test architecture-a review” International Journal of Engineering and Technology (UAE) 7.1.1 (2018): 582-586.
34. Karthick, R., and M. Sundararajan.,”PSO based out-of-order (ooo) execution scheme for HTMPSOC” Journal of Advanced Research in Dynamical and Control Systems 9 (2017): 1969.
35. Punarselvam, Dr E., and S. Gopi. "Effective and Efficient Traffic Scrutiny in Sweet Server with Data Privacy." International Journal on Applications in Information and Communication Engineering 5.2 (2019): 1-5.
36. Punarselvam, E., et al. "Different loading condition and angle measurement of human lumbar spine MRI image using ANSYS." Journal of Ambient Intelligence and Humanized Computing 12.5
37. Punarselvam, E., and P. Suresh. "Non-Linear Filtering Technique Used for Testing the Human Lumbar Spine FEA Model." Journal of medical systems 43, no. 2 (2019): 1-13.
38. Punarselvam, E., Hemalatha, E., Dhivahar, J., Gowtham, V., Hari, V., & Thamaraikannan, R. PREDICTING WIRELESS CHANNELS FOR ULTRA-RELIABLE LOW-LATENCY COMMUNICATIONS.