Original Article

Cancer Stage Detection Using Deep Learning

Aarzu Sangwan, Aakriti Singh, Aakash tyagi, Suraj Bhatnagar

Department of Computer Science and Engineering
Meerut Institute of Engineering and Technology, Meerut, U.P., India
{aarzu.sangwan.cs.2019, aakriti.singh.cs.2019, aakash.tyagi.cs.2019@miet.ac.in, suraj.bhatnagar@miet.ac.in}
@miet.ac.in
DOI: 10.47750/pnr.2022.13.S10.401

Abstract

Lung cancer is one of the common types of cancer we see around the world. It's very important to diagnose and start to treat lung cancer at an early stage, the earlier we start to diagnose it we will be able to save more and more lives dealing with it. There are different methods for diagnosing lung cancer including X rays**,** CT scans**,** PET-CT scans, bronchoscopies, and biopsies. However, staining techniques that stain tissue taken from a biopsy are often used to determine the subtypes of lung cancer based on H and E tissue, since knowing the sub type of lung cancer is very important, as it is curable if detected early. Histological examination of the suspicious tissue with regards to clinical and radiological features forms the definitive diagnosis of the disease. It is extremely important to analyse the histopathological image of lung cancer. Deep learning techniques are being used to speed up the critical process of lung cancer diagnosis and reduce the burden on pathologists and several studies have reported the importance of convolutional neural networks (CNNs) in classifying histopathological images of various types of cancer. Our objective is to provide a timely investigation to the people, a deep learning inceptionV3 model for early stage lung cancer detection is built using the dataset consisting of 5000 images of each :- Lung benign tissue, Lung adenocarcinoma , Lung squamous cell carcinoma. Model is built with an accuracy of 94%.

Keywords: Lung cancer, Histopathological images, Deep learning inception V3 model, convolutional neural networks.

1 Introduction

Lung cancer is the leading cause of cancer-related death around the globe. In 2020, 2.2 million new cancer cases were found worldwide, and 1.8 million deaths were due to lung cancer, representing 18.0% of all deaths from cancer. Like any other cancer, lung cancer too develops due to alterations in the affected one's DNA or due to epigenetic changes which leads to uncontrolled cell proliferation and reduced apoptosis. Most of the time the cells of lung tumour metastasize to more distant parts of the body like brain, bones, liver, and adrenal glands. Hence lung cancer is synonymously referred to as Lung carcinoma. There are several carcinogens that induce DNA damage further leading to lung cancer. Tobacco smoking, asbestos and exposure to other air pollutants and ionizing radiations like Gamma rays and X-rays are some of the contributors.

The main two types of cancerous tumours are small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). Many imaging methods, such as Chest X-ray, CT scan and Magnetic Resonance Imaging(MRI), are used to detect early tumours. Computed tomography(CT) imaging is often used in diagnosis of lung cancer as it also tells about the type and extent of the disease. Histological examination of the suspicious tissue with regards to clinical and radiological features forms the definitive diagnosis of the disease. Detection includes the grouping of tumours into two categories: (i) non-cancerous (benign) and (ii) cancerous tumours (malignant).

Lung cancer that spreads to the brain can cause difficulties with vision, weakness on one side of the body. Symptoms of primary lung cancers include cough, coughing up blood, chest pain, and shortness of breath. Cancer that spreads to the airways may block airflow and make breathing difficult. The obstruction may cause secretions to build up behind it, increasing the risk of pneumonia. Unfortunately, most diagnoses occur at the later stages of the disease, largely due to a lack of early-stage symptoms. Lung cancer staging is therefore one of the factors affecting both the prognosis and the potential treatment of the disease. An image processing technique is built to

detect the growing disease at an early stage so that the chances of mortality can be reduced. The accuracy and the quality of image is one of the core factors of this research.

2 Literature Review

There have been several studies done on lung cancer detection. Unfortunately, most diagnoses occur at the later stages of the disease, largely due to lack of early-stage symptoms. Prediction models have been built using machine learning and deep learning algorithms including CNN, SVM classifiers, Naïve Bayes ,genetic algorithm and more. Researchers have tried to use deep learning techniques for the detection of lung cancer using the histopathological images to detect early tumours.

[1] A computer-aided diagnosis (CAD) system for lung cancer diagnosis consisting of four main processes: lung field segmentation, detection of nodules within the lung field, segmentation of detected nodules and diagnosis of nodules as benign or malignant. But for certain lung nodules like Ground-glass nodules and cavities, the diagnosis cannot be done with current growth rate techniques, therefore to diagnose these nodules additional methods would be required.

[2] They built a system for detection and classification using the CT scan images, consisting of five steps that included: dataset collection, pre-processing and nodules extraction followed by feature extraction and classification. The model did the classification as benign or malignant with an accuracy of 93.52%.

[3] They developed a SVM classifier to detect and classify lung cancer by using the X-ray images of benign or malignant lungs. In the beginning a median filter is used at the time of noise detection phase while the pre-processing stage is being carried out. During segmentation, further K-Means and Fuzzy C-Means clustering can be used for the purpose of finding the features.

[4] The prediction model is built using the CT scan images of lungs, but the images were analysed with the support of Linear Discriminate Analysis (LDA) and Optimal Deep Neural Network (ODNN) applied to the images and than enriched using Modified Gravitational Search Algorithm (MGSA) . Extraction and reduction of dimensions of the deep features for categorization into benign or malignant is done with an accuracy of 94.56%.

[5] They did a survey and presented their views on the different methodologies used for the purpose of classification using histopathological images as its features forms the definitive diagnosis of the disease, the survey consisted of over 130 papers , that let them conclude the progress with respect to different machine learning and deep learning techniques along with the associated challenges and limitations with these approaches.

[6] The classification of lung cancer was done on the basis of CT images using artificial neural networks , taking the parameters out of the segmented CT images , the process was carried out with the help of feed forward and feed forward back propagation that categorize better than the former. They found out that the training function gives the maximum accuracy around 91.1%, they also proposed two new training functions.

[7] Since it is difficult to extract the symptoms of lung cancer due to the structure of tissues, they examined the digital images using the Principal Component Analysis(PCA) algorithm. The feature extraction, pre-processing was done using the GLCM method for the classification of normal and infected images through which the chances of survival of a patient could be estimated.

[8] A K-Nearest-Neighbor based technique was used for the estimation of likelihood of lung cancer in a person, trying to find out the stage. For efficient feature selection genetic algorithm was used, an experimental procedure was used to calculate the value of k for improving the efficiency and accuracy of the algorithm.

3 Methodology

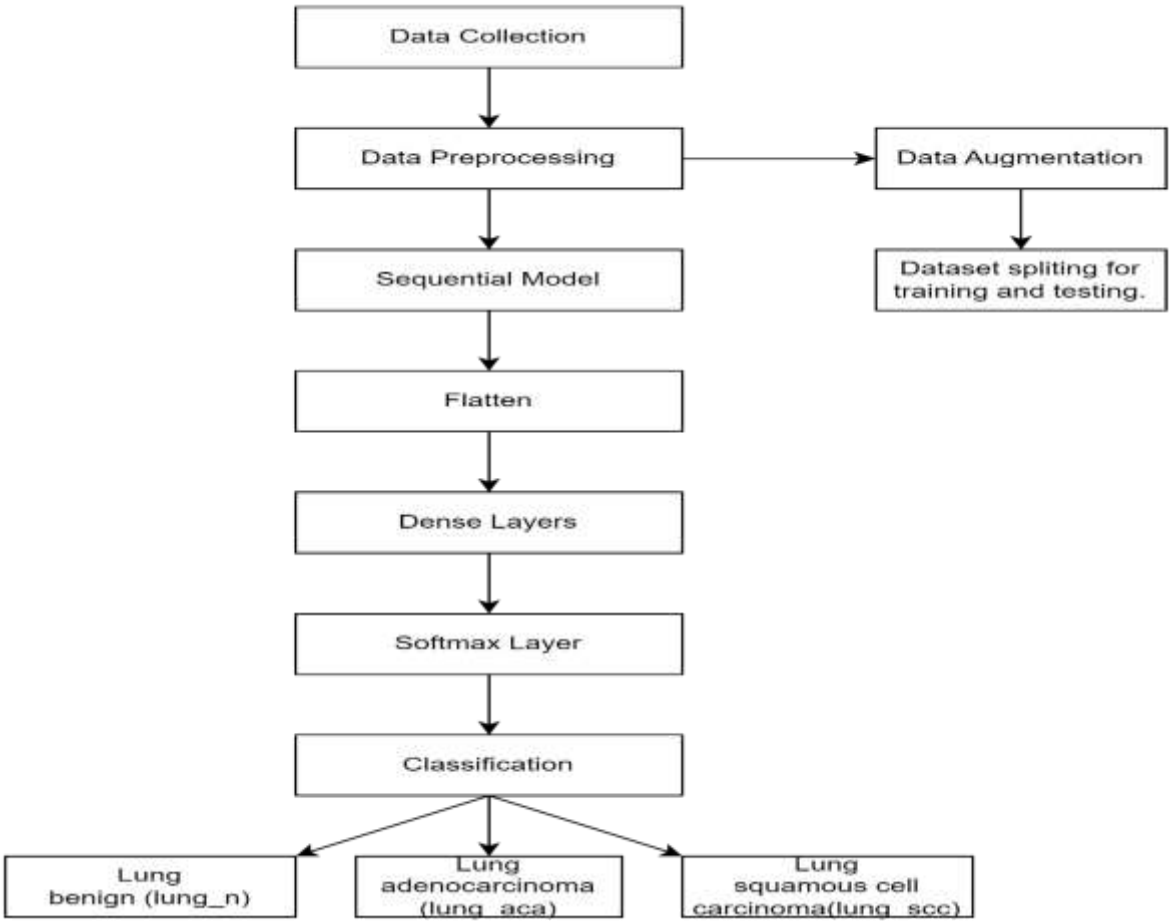


Fig.1 workflow of the model.

The system is based on transfer learning, where we use the previous learning or experience in new model so that the accuracy of current model can be increased and even lesser amount of data gives more accuracy. To achieve transfer learning this system used inception model which was developed by google to provide the arrangement of layers so that the training can be done in more efficient way [9][10].

The necessary packages used are numpy for arrays and numerical calculations, Keras which is a high level API that runs on top of tensorflow, Tensorflow is developed by google and is an open source library for deep learning, Pandas for data loading, cleaning and analysis. Pre Trained inception V3 model is used for classification [11][12][13].

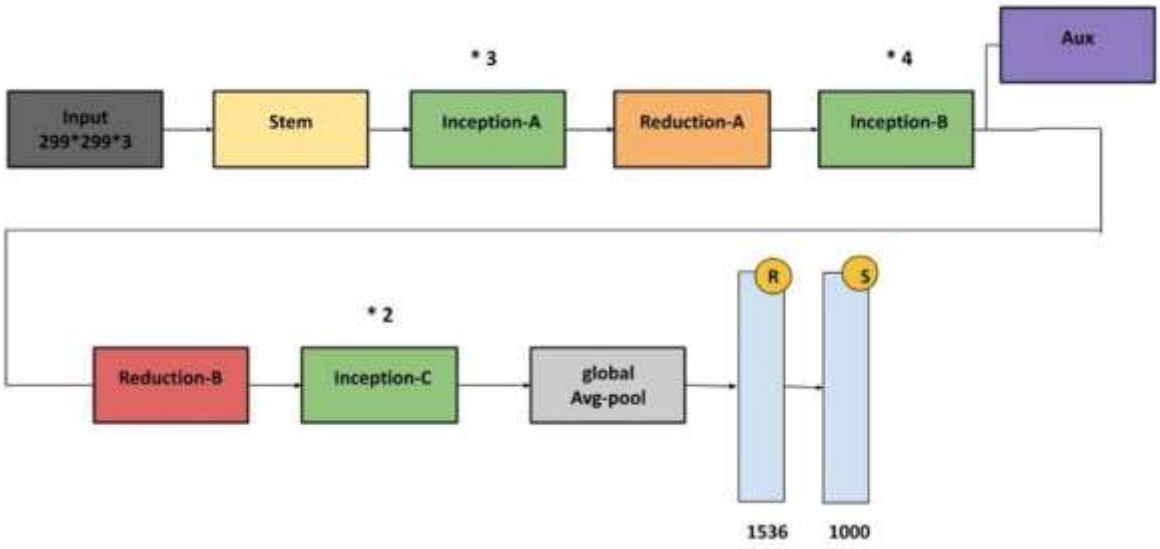


Fig.2 Inception V3 CNN architecture

First of all the images of lungs afflicted by cancer are collected. Images are then pre-processed to remove the noise from data so that model can be fed with proper dataset. After image pre-processing the data is fed to model in which data is processed in various steps. The model apply convolution step on the image so that useful information can be highlighted in the image.

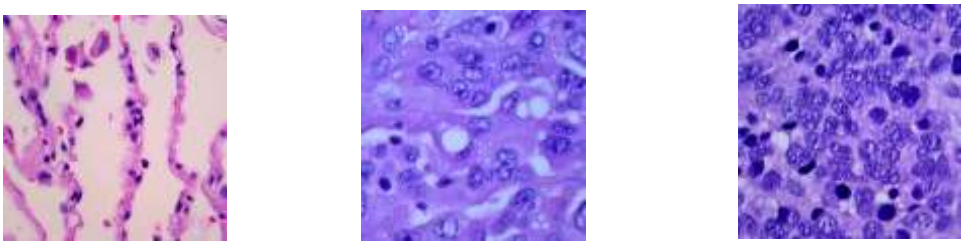
After that the image size is reduced by average pooling and max pooling respectively so that further steps can be carried out in lesser time. After pooling the dense layer is used so that each neuron can be connected with this layer which in turn takes care of every single feature.

Finally the neuron corresponding to the output generated by the model gets activated, the neurons present will be equal to the number of possible outputs. This way the model prediction is identified [14][15].

3.1 ALGORITHM

Step 1: Data Collection

The dataset (histopathological images of all three types of lung cancer) are collected at one place. The dataset consisting of 5000 images of each :- Lung benign tissue, Lung adenocarcinoma , Lung squamous cell carcinoma. The images should have same size and resolution so that the efficiency of model can be maintained [16][17].



Lung Benign | Lung Adenocarcinoma | Lung Squamous Cell Carcinoma

Fig.3 Dataset for training and testing

Step 2: Image Pre-processing

Pre-processing is done in order to make the raw data available for computers to understand. This step reduces data noise [18].

Step 3: Convolution

This layer highlights the main feature of images using filters. Filters are matrix which gets multiplied with the image matrix and make feature more clearly detectable. Some common filters are Gaussian Blur and Prewitt Filter [19].

Step 4: Average Polling

Pooling, in general, is done to reduce image size without loosing it's important information. In average pooling the image's matrix is divided into small part and averages of those parts, later on represent their parent matrix and these representatives are joined together to get the image with selected features [20][21].

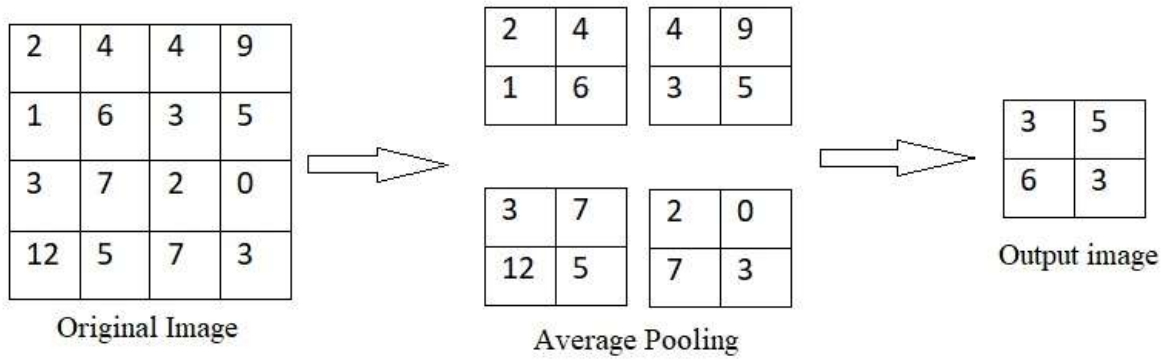


Fig.4 average Pooling

Step 5: Max Polling

Pooling, in general, is done to reduce image size without loosing it's important information. The image matrix is divided into small parts and maximum value of those parts later on represent their parent matrix and these representatives are joined together to get the image with selected features [22][23].

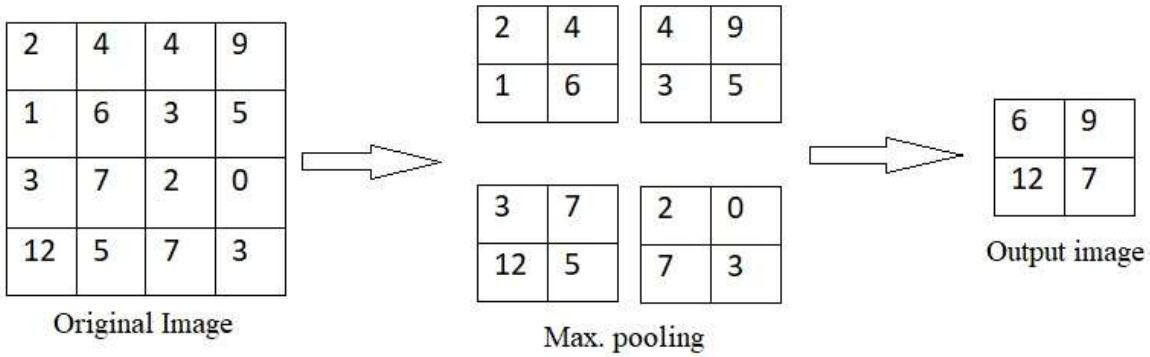


Fig.5 Max pooling

Step 6: Fully/Dense Connected Layer

This layer is used to connect all neurons from the previous layer so that every feature can be processed, even the output of a single neuron is not processed. If there is a neuron whose output does not affect the outcome of the model, its weight is assigned to zero. This layer improves the overall performance of the model.

Step 7: Activation

The activation function is responsible for either activating or not activating a particular neuron. In the process of prediction using simple mathematical operations, it is determined whether the inputs of the neurons in the network are important or not.

Step 8: Output Layer

This layer has number of neurons equal to the number of possible outcomes. The neuron corresponding to the result predicted by model gets activated so that user can understand the output of machine.

4 Experimental result discussion

Model Accuracy Table-

Total images	Correctly predicted	Wrongly predicted	Accuracy
10	8	2	80%
20	17	3	85%
30	26	4	86.6%
50	42	8	84%
100	83	17	83%

The table above shows data for correct and incorrect predictions. In inaccurate cases, model predicts different result than expected like lung adenocarcinoma(lung_aca) infected lungs were predicted as lung benign(lung_n). After careful observation of the system, we found that most of the poor results were due to poor lighting or low resolution devices used to capture the images. These problems can be overcome by paying attention to the lighting at night. To avoid overfitting and enhance the accuracy of the model data augmentation was performed and model was trained up to 10 epochs. The training loss and accuracy along with the validation loss and accuracy are depicted using the matplotlib library, the accuracy of training was around 94% and the accuracy of validation was around 95.33% [24][25].

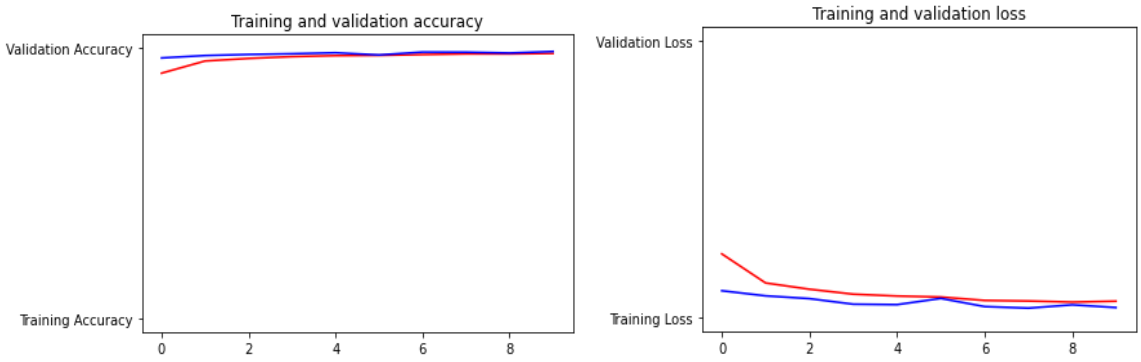


Fig.6 Comparing loss and training/validation accuracy

5 Conclusion

Lung Cancer is a wide spread problem and it is one of the leading causes of cancer-related problems and deaths. Unfortunately, most diagnoses occur at the later stages of the disease, largely due to a lack of early-stage

symptoms. Staging in lung cancer is therefore one of the factors affecting both the prognosis and the potential treatment of the disease. This problem can be reduced to some extent by detecting it earlier to save millions of lives, So this proposed model helps in easy and early identification of cancer which results in proper treatment and early diagnosis by analysing the histopathological images that provides definitive diagnosis and staging using transfer learning with CNN. The training accuracy of the proposed model reached to 94% and validation accuracy to 95.33%.

6 References

[1] Ayman El-Baz, Garth M. Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman, and Behnoush Abdollahi,"Computer-Aided Di agnosis Systems for Lung Cancer: Challenges and Methodologies" International Journal of Biomedical Imaging, Article ID 942353, 46 pages, 2013

[2] Rotem Golan, Christian Jacob, J"org Denzinger, "Lung Nodule Detection in CT Images using Deep Convolutional Neural Networks", 978-1-5090-0620- 5/16

[3] Joon P., Bajaj S. B., Jatain A. Progress in Advanced Computing and Intelligent Engineering . Singapore: Springer; 2019. Segmentation and detection of lung cancer using image processing and clustering techniques; pp. 13–23

[4] Lakshmanaprabu S. K., Mohanty S. N., Shankar K., Arunkumar N., Ramirez G. Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems . 2019;92:374–382. doi: 10.1016/j.future.2018.10.009.

[5] Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal. 2021;67:101813.

[6] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 202–209, 2014.

[7] DBhatnagar AKTiwari VVijayarajan AKrishnamoorthy ''Classification of normal and abnormal images of lung cancer IOP Conference Series: Materials Science and Engineering Vol 263 2017.

[8] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," Expert Systems with Applications, vol. 164, article 113981, 2021.

[9] Narayan, Vipul, and A. K. Daniel. "Multi-tier cluster based smart farming using wireless sensor network." 2020 5th international conference on computing, communication and security (ICCCS). IEEE, 2020.

[10] Narayan, Vipul, and A. K. Daniel. "A novel approach for cluster head selection using trust function in WSN." Scalable Computing: Practice and Experience 22.1 (2021): 1-13.

[11] Narayan, Vipul, and A. K. Daniel. "RBCHS: Region-based cluster head selection protocol in wireless sensor network." Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020. Springer Singapore, 2021.

[12] Narayan, Vipul, and A. K. Daniel. "FBCHS: Fuzzy Based Cluster Head Selection Protocol to Enhance Network Lifetime of WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.3 (2022): 285-307.

[13] Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." Journal of Scientific & Industrial Research 81.12 (2022): 1297-1309.

[14] Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.

[15]Narayan, Vipul, et al. "E-Commerce recommendation method based on collaborative filtering technology." International Journal of Current Engineering and Technology 7.3 (2017): 974-982.

[16] Vipul, Narayan, and A. K. Daniel. "A novel protocol for detection and optimization of overlapping coverage in wireless sensor network." International Journal of Engineering and Advanced Technology 8.6 (2019): 422-462.

[17]Narayan, Vipul, et al. "To Implement a Web Page using Thread in Java." (2017).

[18] Srivastava, Swapnita, and P. K. Singh. "HCIP: Hybrid Short Long History Table-based Cache Instruction Prefetcher." International Journal of Next-Generation Computing 13.3 (2022).

[19] Srivastava, Swapnita, and P. K. Singh. "Proof of Optimality based on Greedy Algorithm for Offline Cache Replacement Algorithm." International Journal of Next-Generation Computing 13.3 (2022).

[20] Srivastava, Swapnita, and P. K. Singh. "Proof of Optimality based on Greedy Algorithm for Offline Cache Replacement Algorithm." International Journal of Next-Generation Computing 13.3 (2022).

[21] Narayan, Vipul, and A. K. Daniel. "CHOP: Maximum coverage optimization and resolve hole healing problem using sleep and wake-up technique for WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.2 (2022): 159-178.

[22]Narayan, Vipul, and A. K. Daniel. "CHHP: coverage optimization and hole healing protocol using sleep and wake-up concept for wireless sensor network." International Journal of System Assurance Engineering and Management 13.Suppl 1 (2022): 546-556.

[23] Narayan, Vipul, and A. K. Daniel. "IOT based sensor monitoring system for smart complex and shopping malls." Mobile Networks and Management: 11th EAI International Conference, MONAMI 2021, Virtual Event, October 27-29, 2021, Proceedings. Cham: Springer International Publishing, 2022

[24] Smiti, Puja, Swapnita Srivastava, and Nitin Rakesh. "Video and audio streaming issues in multimedia application." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.

[25] Smriti, Puja, Swapnita Srivastava, and Saurabh Singh. "Keyboard invariant biometric authentication." 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT). IEEE, 2018.