[1]Dr. Seelam, Nagarjuna Reddy

[2]Gayatri Parasa

[3]R. Janani

[4]Srinivasulu Thiruveedula

[5]Dr. K. Sujatha

[6]Dr. B. Varaprasad Rao

[7]Dr. Sharath M N

Modernizing Cancer Diagnosis with an Intelligent Model for Lung and Colon Cancer

Abstract: - Cancer is a lethal condition resulting from a mix of genetic mutations and several metabolic irregularities. Lung and Colon Cancer (LCC) are the primary reasons for mortality and impairment in humans. Identifying these malignancies by histopathology is often crucial for deciding the most appropriate therapy. Early discovery of the disease on the side significantly reduces the chances of death. Machine Learning (ML) methods accelerate cancer diagnosis, enabling researchers to analyze more individuals in a smaller time frame and at a reduced fee. The research presented a hybrid ensemble-based feature-extracting system to detect LCC effectively. It combines advanced feature extraction and ensemble-based learning with efficient filtration for datasets of LCC images. The system is assessed using histopathology datasets for the LCC. The research indicates that the hybrid system can accurately diagnose the lungs and colon. Therefore, these models might be used in clinical settings to assist doctors in diagnosing malignancies.

Keywords: Lung cancer diagnosis, Colon cancer diagnosis, Hybrid ensemble-based model, Machine Learning in cancer detection, Histopathology image analysis, Transfer Learning in cancer classification, Ensemble Learning for cancer diagnosis, Intelligent model for cancer detection, Feature extraction in cancer diagnosis, ML-based cancer detection system

1 Introduction to lung and colon cancer diagnosis

Cancer is considered among the most severe illnesses globally, as stated by the World Health Organization [1]. Colon Cancer (CC) accounts for 10.4% of cancer-related fatalities worldwide [2]. Lung cancer accounts for 12.8% of all diagnosed cancers and 17.6% of all cancer deaths [3]. The incidence of malignant tumors has been on the

[1 1]Associate Professor,Department of Computer Science & Engineering,

Lakireddy Bali Reddy College of Engineering, Affiliated to JNTUK Mylavaram (A P), India

Mail id: nagvik@gmail.com

[2]Assistant ProfessorComputer Science and Engineering,Koneru Lakshmaiah Education Foundation,Vaddeswaram, Andhra Pradesh, India.

gayathriparasa20@gmail.com

[3]Assistant professor,Department of computer science and business systems,panimalar engineering college

Jananideivamani@gmail.com

[4]Assistant Professor,Department of IT,Aditya college of engineering and technology, Surampalem

tsrinu531@gmail.com

[5]Assistant Professor,Department of CSE,CSE, SRMIST Ramapuram

sujathak@srmist.edu.in

[6]Professor, Department of Computer Science and Engineering,RVR&JC College of Engineering,Guntur, AP.

bvpr@rvrjc.ac.in

[7]Associate Professor,Department of CSE,Rajeev Institute of Technology,

rise worldwide in the past decade. Cancer affects individuals, but it is more difficult for people from 50 to 75. If current trends continue, there is a forecasted rise of over 70% in the probability of mortality by 2035. Squamous cell tumors and adenocarcinoma are among the most common types of lung cancer.

The other histological kinds include big and small cell malignancies. Adenocarcinoma of the lung is more prevalent among current or former smokers, although it equally impacts those who do not smoke. This condition is more prevalent among young individuals and females, often originating at the outer edges of the lungs before spreading. There is a connection between previous smoking and squamous cell cancers. It spreads quickly and expands, which might complicate therapy.

The colon is the last segment of the human gastrointestinal tract. Colon cancer develops when cancerous cells enter the colon. Colon cancer is not directly correlated with age. It is more prevalent in the elderly population. Polyps, which are benign groupings of cells, often develop on the inside of the colon. Some of these tumors have the potential to develop into colon cancer as time passes.

Cancer is caused by a range of variables, including behavioral habits such as alcohol and smoking, excessive Body Mass Index (BMI), and ecological poisons. The causes vary among individuals. Common indicators of cancer include bruises, hemorrhaging muscular soreness, losing weight, difficulty breathing, coughing that persists, nausea, tiredness, and discomfort. Identifying tumours without advanced diagnostic techniques like biopsies, ultrasonography, Positron Emission Tomography (PET) scans [4], Magnetic Resonance Imaging (MRI) [5], and Computer Tomography (CT) [6] scans are challenging. Patients often exhibit little or no problems in the early stages, and it is often too late when symptoms manifest.

Machine Learning (ML) [7] is a branch of Artificial Intelligence (AI) [8] technology that originated from studying pattern recognition and cognitive learning principles. It creates systems that efficiently adjust to a wide range of data and make predictions using past facts. ML has been effectively used in several challenging areas to achieve impressive results when establishing explicit procedures with acceptable effectiveness was considered problematic and unattainable.

The following parts of the research are given below: section 2 deals with the background and related works about lung and colon cancer. The proposed hybrid ensemble-based feature-extracting model with different ML models is shown in section 3. Section 4 analyses and showcases the outcomes of varying ML models for feature extraction. The conclusion and findings of the study are listed in section 5.

2 Background and literature survey

Classifying histopathological pictures of several types of cancer has garnered significant interest. Different methods, including ML, Deep Learning (DL) [9], and Transfer Learning (TL) [10], are analyzed to identify Lung and Colon Cancer (LCC).

Shafi et al. studied a hybrid technique combining a Supporting Vector Machine (SVM) and a neural network for classifying Lung Cancer (LC) [11]. The researchers first preprocessed the database and used a boost-up resilient characteristics-based approach for obtaining features from the photos, which has been improved using a genetic algorithm. The categorization was done using a hybrid approach.

Nguyen et al. created a method using supervision to classify and locate colorectal cancer, distinguishing between cancerous and normal tissue [12]. The design was segmented into three stages. They consist of an ensemble, a cell-level, and an image-level structure. Within the picture-level model, the complete picture was examined at a reduced clarity, and the likelihood of malignancy in the surrounding cells was forecasted using the Convolutional Neural Network (CNN) method [13]. The picture was evaluated with increased clarity. The heating map generated from the tissue-level structure and the characteristics extracted from the picture-level structure were merged in the combination phase to conduct the categorization. The technique's success was assessed using standard performance criteria.

Obayya et al. suggested a Dual Horizontally Squashed CapsNet (DHS-CapsNet) for classifying colon and lung cancer [14]. Two approaches, the horizontally squashed function and encoding feature fusion, were included in

this structure. The squashed function was employed to compress matrices effectively and create sparsity to enable prejudiced capsules to obtain essential data from images with diverse backgrounds. The encoding feature fusion integrated the extracted characteristic from the 2-lane convolutional layering.

Naga Raju et al. proposed an LCC categorization system using a capsuling system, which functioned as a DL system [15]. This model consists of two stages. During the first step, the unprocessed input picture was used for categorization. Preprocessed histopathological photos were used for categorization in the second classifier. The different layers handled the raw input photos, while the blocks treated the preprocessed pictures. The results of these two levels were merged in the capsule levels to conduct the categorization.

Mehmood et al. utilized histopathology images to develop a CNN categorization algorithm to differentiate between five types of colon and lung malignancies [16]. The contrasting effect of each picture was enhanced using the standard sharpening method at first. Two picture alteration methods were used to extract attributes from histopathology cancer pictures. The techniques used are single-level discrete 2D wavelet transformation and two-dimensional independent Fourier transformation. The collected characteristics were combined to create a collection of attributes.

3 Proposed hybrid ensemble-based feature-extracting model

The primary objective of the suggested structure is to classify lung and colon tumors into five groups based on histopathological pictures. The conditions include benign, lung cancer, and colon cancer. This is achieved using a Hybrid Extreme Learning Machine (HELM). The technique consists of three main elements. The tasks include preprocessing, extracting features, and categorization. Figure 1 displays the system architecture of the suggested HELM method.

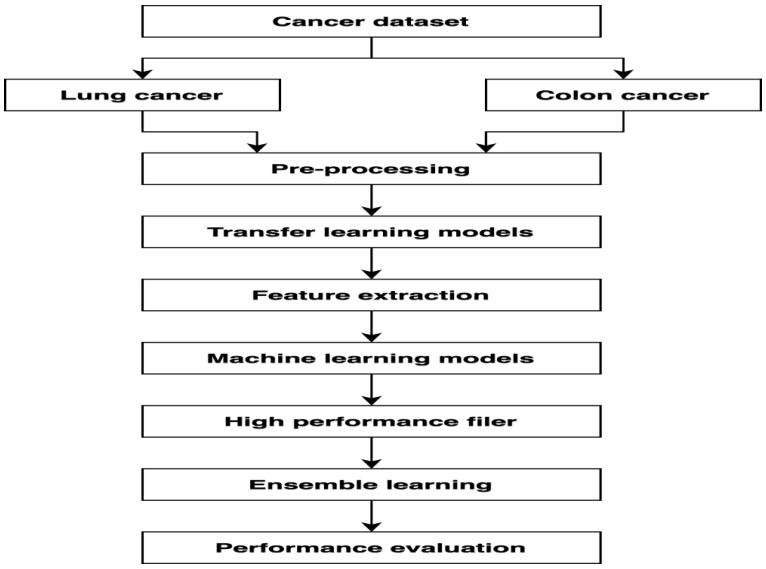

Fig. 1. Architecture of the proposed HELM system

The research created the model by extracting features and Machine Learning (ML) methods on LC25000 LCC histopathology imaging databases to predict LCC outcomes. The steps of the proposed model are listed below:

Step 1: The research merges the histopathological imaging databases for LCC to use in the investigation. The databases include three kinds of lung malignancies and two categories of CC.

Step 2: During pre-processing, the picture is resized to 128 × 128, converted from BGR to RGB, and then transformed into a NumPy array. The research conducts the scaling process by dividing the pixel values of the picture array by 255, the highest brightness value of a picture.

Step 3: Features are extracted from the picture databases using multiple TL models, VGG-16, VGG-19, DenseNet-169, and DenseNet-201, in the deep feature mining stage. These features are gathered for the following.

Step 4: The gathered characteristics are used for six popular ML methods: Random Forests (RF), SVM, Logical Regression (LR), Multiple-Layered Perceptron (MLP), Extreme Gradient Boosting (XGB), and Lighter Gradient Boosting (LGB) to assess their outcomes.

Step 5: In the Higher-Performance Filtration stage, the researchers refine the ML methods according to their accuracy. The research uses this filtering process to choose the best three methods for the subsequent round.

Step 6: In this combining phase, the best three methods selected from high-performance filtering are used to create majority and weighted average voting classifications for every TL. The most efficient ensemble polling classification is chosen for every transfer learning method.

Step 7: The efficacy is evaluated by assessing the outcomes of every transfer learning and choosing the optimal transfer learning method. The chosen model in the study is evaluated using performance measures such as Accuracy (%), Recalls (%), Precisions (%), F1-Scores (%), Region Under Curve (ROC) (%), Mean Absolute Errors (MAE) (%), Mean Squared Errors (MSE) (%), and Root Mean Squared Errors (RMSE) (%).

3.1 Ensemble learning

EL, often called multiple classification or committee-based training, involves effectively training and combining numerous classifiers to address a problem. Significant interest has been in recent years, especially in AI and ML. The system is reasonable, efficient, adaptive in several vital areas, and adds considerably, making the attention it receives appropriate. The principle of majority voting involves using anticipated category labeling by hard employees. Soft calculates every class label by selecting the arg max of the predicted likelihoods suitable for a set of well-calibrated trainees.

Soft voting involves predicting category labels based on the classifier's predicted possibilities by summing the class possibilities. By standard, the system estimates soft voting using an identical value. The weight employed in the weighted average method is consistent. The classifiers are well-calibrated, indicating that their results align well with calculated probabilities. Equation (1) can observe the soft voting mechanism.

$$\hat{y} = arg\ arg\ \{\sum_{i=0}^{N-1} \quad w_i p_i\} \qquad (1)$$

where w_i represents the weights of classifications and p_i represents the predictions of classifications.

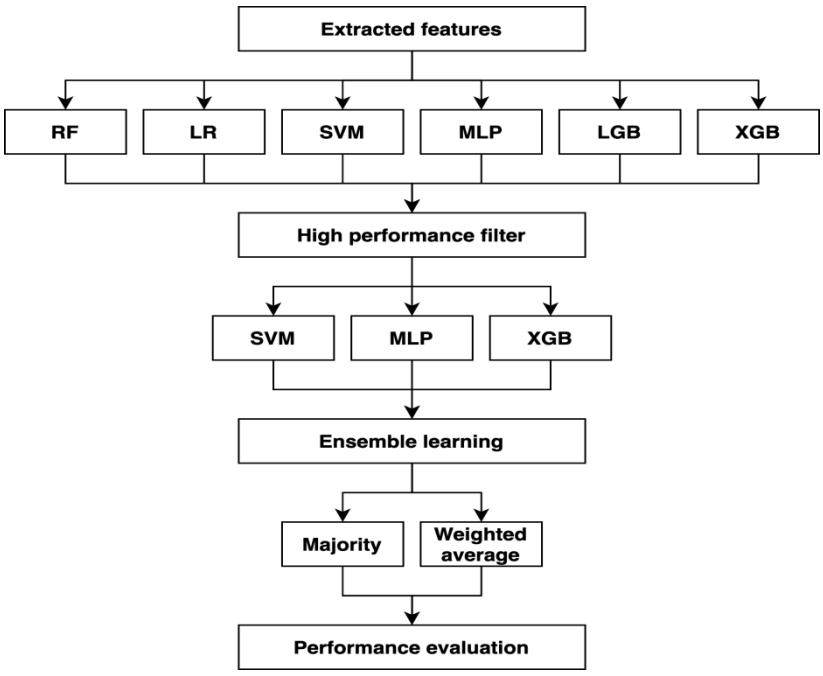

Fig. 2. Ensemble learning-based high-performance filtering design

Figure 7 demonstrates the process of ensemble learning using high-performance filtering. Deeply retrieved features are inputted into several ML computations, resulting in three methods via high-performance filtering. Ensemble learning is applied to these filtered methods. EL uses soft and hard voting techniques to assess effectiveness. They use ensemble learning with the estimations obtained from the high-performance filtering method. The research utilized estimation methods for classifications accordingly.

4 Simulation outcomes

The tests are conducted on a personal computer equipped with an Intel processor operating at 2.40 GHz, four cores, and eight logical processes. The system also has 512 GB of internal storage, 1 TB of external hard disk, 16 GB of RAM, and 32 GB of cloud storage. The test is conducted in the Anaconda Navigation model via a Jupyter system. The proposed ensemble learning method is implemented using Python and standard records.

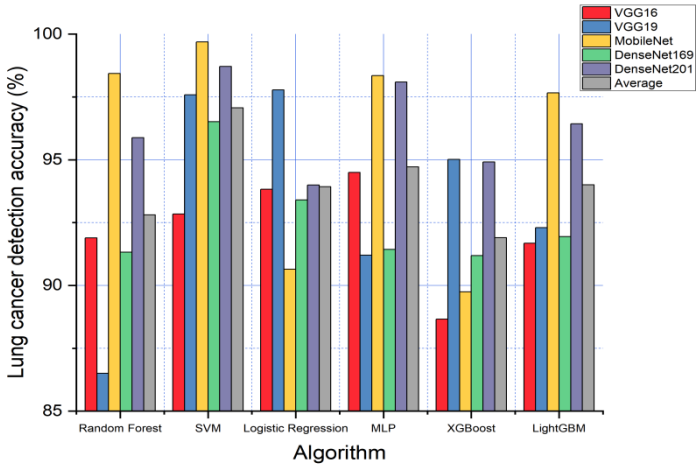

Fig. 3. Lung cancer detection accuracy analysis

Figure 3 illustrates the lung cancer detection outcomes, showing that SVM is the most efficient method with an impressive accuracy of 99.689%. MLP and DenseNet-201 provide strong performance with accuracies of 94.716% and 95.882%, correspondingly. Logistic regression, random forest, and LightGBM regularly achieve accuracy levels of over 90% despite some fluctuations. SVM is the most effective algorithm in this study, demonstrating outstanding performance in detecting lung cancer.

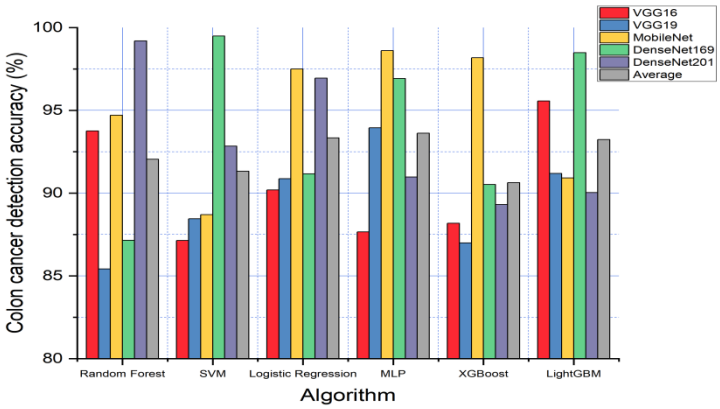

Fig. 4. Colon cancer detection accuracy analysis

Figure 4 displays the diverse accuracies of several ML methods in detecting colon cancer. DenseNet-201 obtains the best accuracy of 99.198%, followed by Logistic Regression at 96.945%. SVM has a strong performance,

getting an accuracy of 92.839%. Random Forest, XGBoost, and LightGBM have accuracies over 90%, while MLP's accuracy is slightly lower at 90.972%. DenseNet-201 is the most effective method for detecting colon cancer in this research, showing better accuracy in categorizing occurrences of colon cancer.

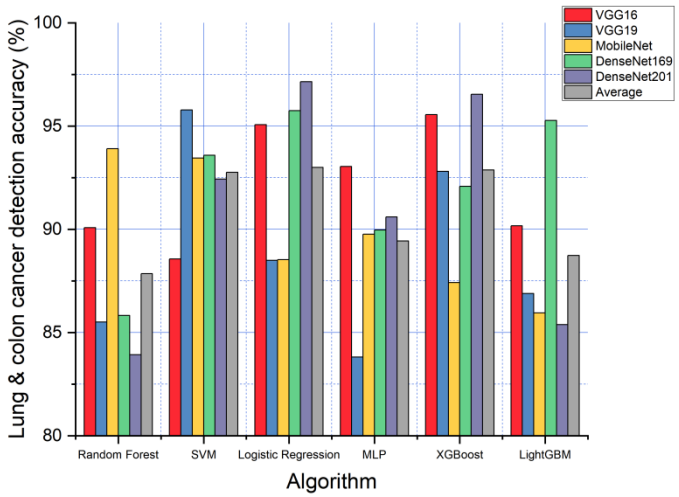

Fig. 5. LCC detection accuracy analysis

Figure 5 displays separate algorithmic outcomes for LCC detection accuracy. Logistic Regression has exceptional accuracy, averaging 92.999%, and excels in managing both forms of cancer. SVM demonstrates a high level of accuracy, averaging 92.764%, showing consistent performance in both lung and colon datasets. MLP and XGBoost show balanced results with average accuracies of 89.44% and 92.881%, respectively. Logistic Regression is the favored method for detecting lung and colon cancers in this research because of its consistent and dependable performance for both types of cancer.

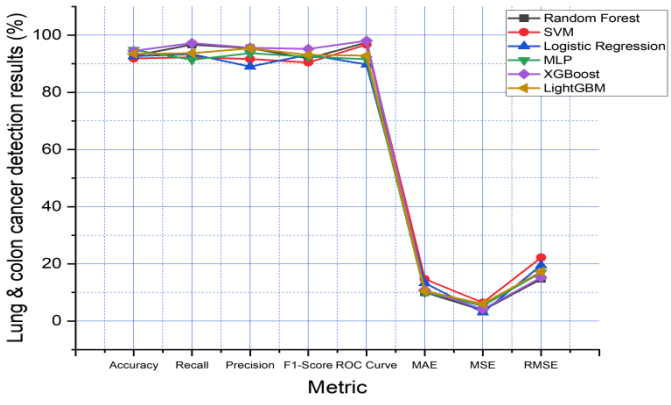

Fig. 6. Lung and colon cancer detection result analysis

Figure 6 presents a detailed summary of the outcomes for detecting lung and colon cancer, highlighting the varied algorithmic performances based on different criteria. MLP achieves the most fantastic accuracy of 94.87%, with XGBoost following behind at 94.5%. Random Forest and LightGBM exhibit high levels of accuracy, surpassing 92%. XGBoost demonstrates its ability to appropriately detect positive situations with a recall rate of 97.2%. Logistic Regression has a precision of 95.59%, indicating a high accuracy in predicting positive situations. XGBoost is a strong option, doing very well in accuracy, recall, and precision, making it a preferred algorithm for detecting LCC together in this research.

5 Conclusion and future scope

The research introduced a hybrid system for detecting LCC. It combines pre-processing, extracting features using transfer learning methods, applying ML computations, selecting the best methods using hyperparameter tuning, and including ensemble learning methods. Next, choose the optimal ensemble voting classification method and the most appropriate feature extraction TL method. The research used five transfer learning models for feature extraction: VGG-16, VGG-19, MobileNet, DenseNet-169, and DenseNet-201. The study evaluated performance using six standard ML methods. The comprehensive progress has been assessed using several performance measures, including accuracy, recalls, precisions, F1-scores, ROC, MAEs, MSEs, and RMSEs. The research has selected the MobileNet model for extracting features from several transfer learning models.

The research will use an ensemble classification created from two ensemble algorithms to assess the outcomes. The test utilizes the LC25000 LCC imaging databases to determine that the proposed approach is more effective for identifying LCC. The proposed model excelled in performance measures and surpassed previous strategies. The model might be utilized in medical facilities to detect LCC automatically. While the structure provides increased accuracy, more advanced enhancements to picture pre-processing and the suggested approach should improve the task. This study needs to gain the ability to gather characteristics effectively from clean and sharp images, which is crucial for achieving superior performance outcomes.

There are many issues that subsequent studies must address.

- The research will now investigate the discovery of novel variables in colon malignancies.
- It analyzes how different factors affect the categorization effectiveness of various methods.
- ML techniques for analyzing complicated genomic information in colon cancer cases.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no conflict of interest.

Human and Animal Rights

This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed Consent

Informed consent does not apply as this was a retrospective review with no identifying patient information.

Funding: Not applicable

Conflicts of interest Statement: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and material:

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Code availability: Not applicable

References

1. R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal. *Cancer statistics, 2022.* CA Cancer J Clin., vol.72, no. 1, pp. 7-33, 2022

2. S.N. Jia, Y.B. Han, R. Yang, Z.C. Yang. *Chemokines in colon cancer progression.* Semin Cancer Biol., Academic Press, vol.86, pp.400-407, 2022

3. A. Leiter, R.R. Veluswamy, J.P. Wisnivesky. *The global burden of lung cancer: current status and future trends.* Nat. Rev. Clin. Oncol., vol.20, no. 9, pp. 624-639, 2023

4. I.D. Apostolopoulos, N.D. Papathanasiou, D.J. Apostolopoulos, G.S. Panayiotakis. *Applications of generative adversarial networks (GANs) in positron emission tomography (PET) imaging: A review.* Eur J Nucl Med Mol Imaging, vol.49, no.11, pp. 3717-3739, 2022

5. F. Bruno, V. Granata, F. Cobianchi Bellisari, F. Sgalambro, E. Tommasino, P. Palumbo, A. Barile. *Advanced Magnetic Resonance Imaging (MRI) Techniques: Technical Principles and Applications in Nanomedicine.* Cancers, vol.14, no. 7, pp. 1-17, 2022

6. Y. Inoue. *Radiation dose modulation of computed tomography component in positron emission tomography/computed tomography.* Semin. Nucl. Med., WB Saunders vol.52, no. 2, pp.157-166, 2022

7. T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, R. Buyya. *Machine learning (ML)-centric resource management in cloud computing: A review and future directions.* J Netw Comput Appl., vol.204, pp. 1-51, 2022

8. A. Haleem, M. Javaid, M.A. Qadri, R.P. Singh, R. Suman. *Artificial intelligence (AI) applications for marketing: A literature-based study.* Int J Intell Netw, 2022

9. K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. Wolverton. *Recent advances and applications of deep learning methods in materials science.* Npj Comput. Mater., vol.8, no.1, pp. 1-26, 2022

10. A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, Azim, M.A. *Transfer learning: a friendly introduction.* J. Big Data, vol.9, no.1, pp.1-19, 2022

11. I. Shafi, S. Din, A. Khan, I.D.L.T. Díez, R.D.J.P. Casanova, K.T. Pifarre, I. Ashraf. *An effective method for lung cancer diagnosis from CT scan using a deep learning-based support vector network.* Cancers, vol.14, no.21, pp.1-18, 2022

12. Mohan, A., Prabha, G. and V., A. 2023. Multi Sensor System and Automatic Shutters for Bridge- An Approach. International Journal of Intelligent Systems and Applications in Engineering. 11, 4s (Feb. 2023), 278–281.

13. Prabha , G. , Mohan, A. , Kumar, R.D. and Velrajkumar, G. 2023. Computational Analogies of Polyvinyl Alcohol Fibres Processed Intellgent Systems with Ferrocement Slabs. International Journal of Intelligent Systems and Applications in Engineering. 11, 4s (Feb. 2023), 313–321.

14. Study On Structural Behaviour Of Ductile High-Performance Concrete Under Impact And Penetration Loads, Lavanayaprabha, S. Mohan, A. Velrajkumar, G., Mohammedharoonzubair, A. Journal of Environmental Protection and Ecology., 2022, 23(6), pp. 2380–2388.`

15. Mohan, A., & K, S. . (2023). Computational Technologies in Geopolymer Concrete by Partial Replacement of C&D Waste. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 282–292.

16. Mohan, A., Dinesh Kumar, R. and J., S. 2023. Simulation for Modified Bitumen Incorporated with Crumb Rubber Waste for Flexible Pavement. International Journal of Intelligent Systems and Applications in Engineering. 11, 4s (Feb. 2023), 56–60.

17. R.Gopalakrishnan, Mohan, "Characterisation on Toughness Property of Self-Compacting Fibre Reinforced Concrete", Journal of Environmental Protection and Ecology 21, No 6, 2153–2163 (2020)

18. H.G. Nguyen, O. Lundström, A. Blank, H. Dawson, A. Lugli, M. Anisimova, I. Zlobec. *Image-based assessment of extracellular mucin-to-tumor area predicts consensus molecular subtypes (CMS) in colorectal cancer*. Mod Pathol., vol.35, no. 2, pp.240-248, 2022

19. A. Shah, M. Shah, A. Pandya, R. Sushra, R. Sushra, M. Mehta, K. Patel. *A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN).* Clinical eHealth, 2023

20. M. Obayya, M.A. Arasi, N. Alruwais, R. Alsini, A. Mohamed, I. Yaseen. *Biomedical image analysis for colon and lung cancer detection using tuna swarm algorithm with deep learning model.* IEEE Access, vol.11, pp.94705-94712, 2023

21. M.S. Naga Raju, B. Srinivasa Rao. *Lung and colon cancer classification using hybrid principle component analysis network-extreme learning machine.* Concurr. Comput. Pract. Exp., vol.35, no. 1, 2023

22. S. Mehmood, T.M. Ghazal, M.A. Khan, M. Zubair, M.T. Naseem, T. Faiz, M. Ahmad. *Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing.* IEEE Access, vol.10, pp.25657-25668, 2022