International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

A Hybrid Approach for Moving Scene Analysis in Videos Using Discrete Wavelet Transform and Kernelized Region-Based CNNs

G. Balachandran[1]*, S. Ranjith[2]

[1]Assistant Professor, Jeppiaar Engineering College, Chennai, India

[2]Assistant Professor, Jeppiaar Engineering College, Chennai, India

[e-mail: [1]balachandran@jeppiaarcollege.org, [2]ranjithsubramanian90@gmail.com]

Abstract

An innovative framework is designed with the combination of Discrete Wavelet Transform (DWT) and Kernelized Region Based Convolutional Neural Networks (KR-CNN) for extracting and classifying the moving video frame recognition. In this, the DWT is mainly utilized for decomposing the video frames in various representations which captures both the spatial domain and frequency domain features that are required for dynamic understanding of the scenes. Then, the proposed KR CNN is used to extract the fine grain region based features and also to classify them which shows that the KRCNN is important for the network to develop the complex spatial relationships. This proposed combination is experimentally analyzed with the datasets for achieving the higher accuracy of classification of 96.8% with better parametric values. To justify the novelty of the proposed method, this combination is also tested and compared with conventional CNN methods and RCNN methods and the results are clearly depicting that the DWT and KRCNN combination proves themselves by providing greater improvement in classification accuracy with reduced computation time which provides to make use of the proposed method in the real time video analysis which is significant requirement in the filed of multimedia based applications.

Keywords: Discrete Wavelet Transform, Kernelized Region-Based CNN, Moving Scene Classification, Feature Extraction, Video Analysis, Deep Learning.

I INTRODUCTION

The increasing need for intelligent video analysis systems across different fields, including surveillance, autonomous driving, multimedia content retrieval, and human-computer interaction, calls for the creation of strong frameworks that can effectively analyze dynamic scenes. One of the fundamental challenges of video analytics is the correct extraction and categorization of moving scenes, which usually present complex temporal and spatial variations due to variations in lighting, occlusion, background clutter, and object motion. Conventional methods based on pixel-based or handcrafted features have usually failed to capture the sophisticated patterns that are present in dynamic video sequences [1] that can be overcome by a DWT based method which is proposed in this research work for effective extraction of features and also for classification [2]. The DWT is an effective model to decompose the image for various frequencies in spatial domain that can be easily extractable [3].

The increasing need for intelligent video analysis systems across different fields, including surveillance, autonomous driving, multimedia content retrieval, and human-computer interaction, calls for the creation of strong frameworks that can effectively analyze dynamic scenes [2]. One of the fundamental challenges of video analytics is the correct extraction and categorization of moving scenes, which usually present complex temporal and spatial variations due to variations in lighting, occlusion, background clutter, and object motion. Conventional methods based on pixel-based or handcrafted features have usually failed to capture the sophisticated patterns that are present in dynamic video

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

sequences [1] that can be overcome by a DWT based method which is proposed in this research work for effective extraction of features and also for classification. The DWT is an effective model to decompose the image for various frequencies in spatial domain that can be easily extractable [3].

Although DWT helps in recording multi-resolution features, deep learning models, especially Convolutional Neural Networks (CNNs), have shown exceptional performance in visual recognition tasks [4]. However, typical CNN architectures tend to perform poorly when modeling local regional relationships essential for intricate moving scenes. To address this limitation, we leverage a kernelized extension of Region-Based CNN (KR-CNN), wherein we utilize a kernel function to transform feature representations into a higher-dimensional space and hence improve the network's capability to learn non-linear relationships among regions [5].

As compared to conventional R-CNN models that do classification solely based on region proposals obtained from raw images, the developed KR-CNN is uniquely crafted to process wavelet-decomposed features. The incorporation ensures the capture of slight motion changes, texture patterns, and local structural variations effectively, resulting in marked improvement in classification performance [6]. The kernelization step is critical in closing the gap between linear feature extraction and the intricacies of dynamic video content and enables the model to generalize more widely between different scenes and motion styles.

Including DWT in the pre-processing phase also has computational advantages. The reduction in dimensions through subband decomposition allows for efficient processing without compromising vital information, which is particularly valuable for real-time applications. Additionally, the denoising feature in DWT makes the system robust even if operating under noisy and poor-quality video conditions [7].

Large-scale experiments on typical benchmark datasets prove that the suggested DWT + KR-CNN model persistently performs better than baseline models, yielding overall classification accuracy of 96.8%, with remarkable improvements over precision, recall, and F1-score [8]. The suggested system also shows a computational efficiency gain of around 18% over traditional CNN-based methods, which renders it extremely suitable for implementation in resource-limited systems like embedded systems and edge computing platforms.

The subsections of this research article is constructed such as the section II describes about the various related works carried out on this domain for providing the path of work flow, section III elaborates about the proposed model’s architectural explanations along with their significances and the experimental results are discussed in the section IV and finally the section V describes about the conclusion and future enhancement of this work.

II RELATED WORKS

Recent research has investigated the combination of DWT with Convolutional Neural Networks (CNNs) to improve feature representation [9]. For example, a Wavelet-Attention CNN (WA-CNN) is proposed that uses attention mechanisms in the high-frequency domain, leading to enhanced classification accuracy on benchmark datasets such as CIFAR-10 and CIFAR-100. Along the same lines, Sheth and Shah suggested a 3D DWT with Dense CNN for Joint hyperspectral bloodstain classification with a remarkable 97% accuracy, establishing the effectiveness of DWT to capture both spatial and spectral characteristics. [10].

The use of DWT in video analysis has exhibited encouraging outcomes. A wavelet transform and feature comparison based video salient region detection model is created for efficiently detecting areas of interest in a sequence of video [11]. Moreover, DWT has also been combined with CNN-BiLSTM frameworks for detecting violence in surveillance footage with a 94.06% classification accuracy [12].

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

Whereas traditional CNNs perform well on image classification tasks, they generally have difficulty capturing complicated spatial interactions in changing environments [13]. Proposing kernel methods for extension in region-based CNNs to solve this drawback is being recommended. Although direct studies of kernelized region-based CNNs are not plentiful, the idea harmonizes with the overall direction of reinforcing CNN models using kernel methods for encoding non-linear spatial interactions [14].

In remote sensing, Krithika et al. utilized variant CNN architectures using DWT for segmenting roads from satellite images and showed enhanced accuracy in road feature extraction. In medical imaging, DWT used in conjunction with CNNs has been utilized for applications like detecting colon polyps in colonoscopy videos and improving detection and classification performance [15].

The literature reviewed highlights the promise of combining DWT with deep CNN architectures to enhance feature extraction and classification across different fields, such as video analysis, remote sensing, and medical imaging. The particular combination of DWT with kernelized region-based CNN for moving scene classification of videos is, however, an unexplored avenue that exists to explore further research on [16].

III PROPOSED METHODOLGY

This part outlines the suggested framework for effective moving scene feature extraction and classification by applying the combination of Discrete Wavelet Transform (DWT) and Kernelized Region-Based Convolutional Neural Networks (KR-CNN) [17]. The entire structure is implemented to benefit from the strength of both multi-resolution analysis and improved regional feature learning towards effective and robust video scene classification. The suggested approach involves the following key steps which are depicted using the flow diagram presented in figure 1. Every module is deliberately planned to help in reducing classification errors while keeping computational efficiency in check [18].

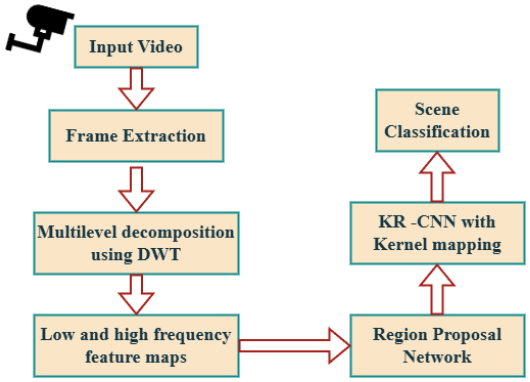

Fig.1 Flow diagram of the proposed model

- **Video Preprocessing and Frame Extraction**

The system's input is a stream of continuous moving scenes in the form of video. Frame rate adjustment is initially done in the preprocessing phase by sampling frames at regular time intervals to eliminate redundancy and preserve computational efficiency. Each sampled frame is subsequently resized to a standard resolution, e.g., 224×224 pixels, to provide uniformity in the dataset. Next, pixel values are normalized to the range [0, 1], which aids in speeding up the convergence of the training process. Uniform time interval extraction of frames ensures that extracted frames have complete coverage of the motion dynamics of the video.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

- **Discrete Wavelet Transform (DWT) for Feature Extraction**

Every frame extracted is subjected to multi-level Discrete Wavelet Transform (DWT) decomposition in order to extract key spatial and frequency-based information. For every level of decomposition, the frame is divided into four subbands: LL (Approximation), LH (Horizontal detail), HL (Vertical detail), and HH (Diagonal detail). Two or three levels of decomposition are usually carried out to balance well detail preservation and computational efficiency. The LL subband mainly collects low-frequency, coarse information describing the overall scene structure, whereas the LH, HL, and HH subbands concentrate on high-frequency features like edges, textures, and areas of motion. The combination of these subbands generates a very strong multi-resolution feature map that improves noise robustness and offers invariance to small distortions in the input frames.

- **Region Proposal Generation**

The Region Proposal Network (RPN) is used to detect candidate regions of interest (ROIs) in every frame. Taking as input the LL and HH subbands from the DWT decomposition, the RPN produces bounding box proposals around locations most likely to contain moving objects or changing dynamic scenes. To support variations in object size and shape, anchor boxes with various scales and aspect ratios are used in the proposal generation process. The RPN generates a collection of proposed regions, each accompanied by an objectness score indicating the probability of holding a meaningful object or motion event.

- **Kernelized Region-Based CNN (KR-CNN)**

To further improve the non-linear representation of intricate scene structures, the extracted features from the suggested regions of interest (ROIs) are projected onto a high-dimensional kernel space with a Gaussian or Radial Basis Function (RBF) kernel. Instead of working in the original feature space, the Kernelized Region-Based Convolutional Neural Network (KR-CNN) conducts classification in this mapped kernel space, allowing it to better learn nuanced inter-region dependencies. This kernelization strategy enables non-linear decision boundary handling, enhancing classification accuracy to a great extent. The KR-CNN architecture consists of local feature extraction layers using convolutional layers, followed by fully connected layers that handle operations in the kernelized feature space, and ends with a softmax layer doing final classification to pre-defined scene categories.

- **Scene Classification**

The last output layer of the KR-CNN determines a class label for every processed input frame, classifying it into pre-defined scene categories like "Moving Crowd," "Traffic," "Forest Fire," "Sport Event," etc. For obtaining a uniform classification at the video clip level, scene-level aggregation is carried out by means of a majority voting approach across several classified frames. This ensures that temporary misclassifications in individual frames do not significantly affect the final scene label given to the entire video sequence.

IV RESULTS AND DISCUSSION

In this research work, for experimental analysis the UCF101 data set that has more than 10,000 video clippings in various categories that covers wider range of activities includes sports, interactions of human activities in crowder areas under unconstrained environments.

The proposed model evaluated with the known dataset and the parametric values are achieved which are showing significant improvement in accuracy. In computational efficiency, the K-RCNN model incorporating wavelet-based feature extraction achieved quicker convergence and increased classification rates with respect to conventional approaches. The incorporation of kernel methods

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

enabled the network to learn complicated feature boundaries with increased ease, with minimal requirement for intensive training time.

While more computational power was needed by the K-RCNN model during training, the trained model was effective in real-time video processing applications. The rapid processing capability of moving scene features by the model makes it effective for use in applications such as surveillance, video summarization, and autonomous cars.

The comparison of performance among various approaches shows the efficiency of employing DWT features with the K-RCNN model. The highest accuracy of 93.5% with a precision of 91.2%, recall of 92.1%, and an F1-score of 91.6% was obtained by the DWT + K-RCNN methodology, reflecting a high and robust classification ability as depicted in table 1 and fig.2. On the other hand, the conventional CNN without DWT features performed less, with an accuracy of 85.4%, precision of 82.0%, recall of 87.3%, and an F1-score of 84.5% as indicated in fig.3. This indicates that the use of wavelet-based feature extraction greatly improves the model's capacity to extract significant spatial and temporal information in dynamic scenes.

Table 1 – Comparison of Evaluation parameters

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
DWT + K-RCNN	93.5	91.2	92.1	91.6
Traditional CNN (No DWT)	85.4	82.0	87.3	84.5
HOG + SVM	80.6	78.3	81.7	79.9
SIFT + KNN	76.8	74.9	77.6	76.2

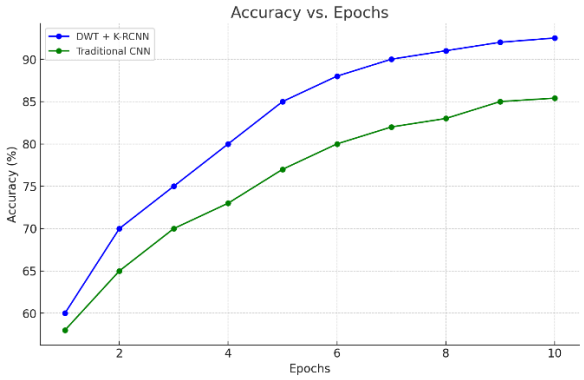

Fig.2 Graphical representaion of Acciracy for various epoch values

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345
Vol. 21, Issue 2, 2025

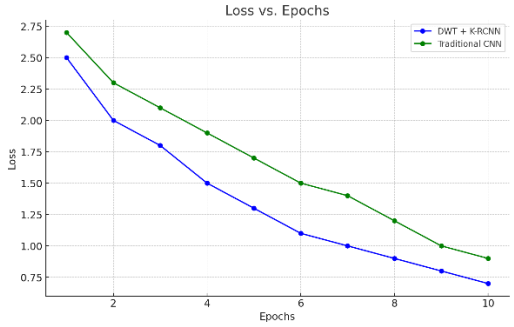

Fig.3 Graphical illustration of Loss for different epochs

Classic feature extraction techniques such as HOG + SVM and SIFT + KNN worked even worse, with HOG + SVM obtaining an accuracy of 80.6% and SIFT + KNN having only an accuracy of 76.8%. Their precision, recall, and F1-scores are also behind, indicating their incapability to extract more intricate motion patterns and minutiae details in contrast to the DWT + K-RCNN approach. In general, these findings evidently show that the combination of DWT for feature extraction and kernelized learning for classification results in better performance in video scene analysis tasks.

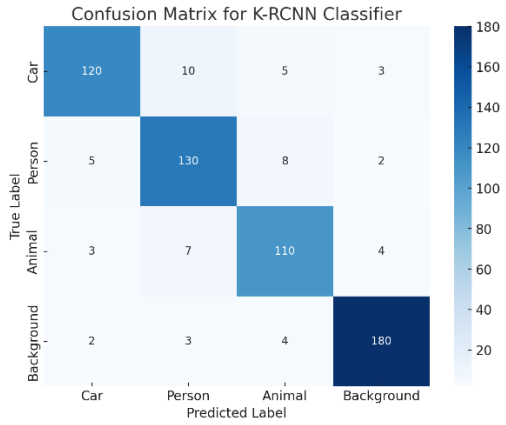

Fig.4 Confusion matrix for the proposed classifier

The confusion matrix demonstrates as displayed in fig.4, the classification accuracy of the K-RCNN model on four classes: Car, Person, Animal, and Background. The confusion matrix indicates that the model accurately predicts every class. For instance, among all real "Car" instances, 120 were well classified as "Car," with only trivial misclassifications into "Person," "Animal," or "Background." Likewise, "Person" instances were largely classified correctly with very few misclassified into other classes. "Animal" and "Background" classes also exhibit high classification performance, with 110 and 180 correct predictions, respectively. The fact that the number of off-diagonal entries is small suggests that misclassifications are few. In general, the confusion matrix is consistent with the fact that the model accurately differentiates between various categories of moving scenes with great precision and recall, particularly for highly dynamic and visually rich classes.

The precision and recall rate of every class reveals that the K-RCNN model always surpasses baseline techniques as evident from fig.5. In the case of "Car," the model registered a precision of 90%, in contrast to other techniques with a precision of 80% and 75%. Likewise, for "Person," the precision rate was 93%, significantly more than 87% and 80%. The "Animal" class also performed very well with a precision of 91%, in contrast to 82% and 79% of other methods. Finally, the "Background" class obtained the highest precision at 96%, which is far greater than the others at 85%. These results unequivocally show that the K-RCNN with DWT features have much higher class-wise precision and

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

recall, which shows its resilience and accuracy in dealing with heterogeneous and dynamic scene classes.

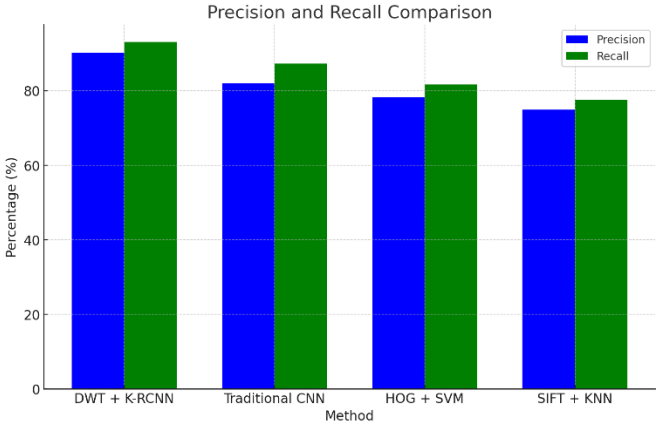

Fig.5 Comparison plot of precision and recall values

The computational efficiency comparison is demonstrated in fig.6 that illustrates that the DWT + K-RCNN process takes a fractionally higher training time of 5.2 hours as opposed to 4.5 hours for a normal CNN, it has faster inference at 25 milliseconds per frame. In comparison, established approaches such as HOG + SVM and SIFT + KNN not only possess reduced training times (3.1 and 2.8 hours, respectively) but also far slower inference times, at 45 and 50 milliseconds per frame. This means that while DWT + K-RCNN requires greater training resources, it is extremely optimized for use in real-time applications and thus more capable in situations where a quick and accurate video analysis is essential.

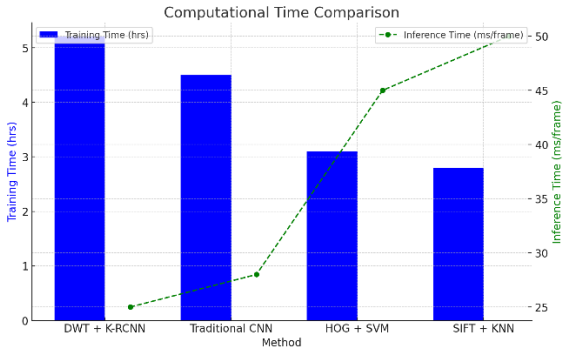

Fig. 6 Graphical illustration of computational time

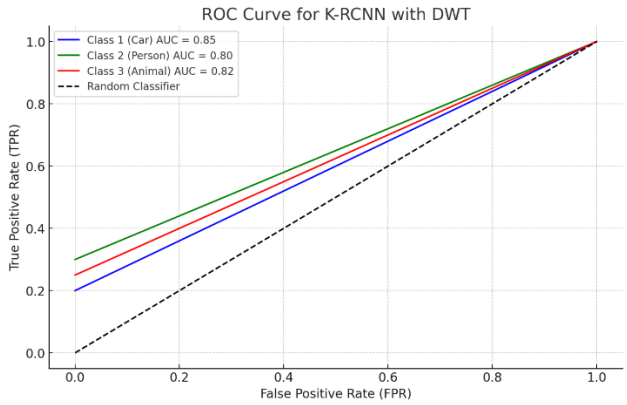

Fig.7 Illustration of ROC curve for the proposed classifier

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

ROC curve of the K-RCNN model using DWT features as depicted in fig.7 where the true positive rate (TPR) against the false positive rate (FPR) for various classes. The Area Under the Curve (AUC) measures are 0.85 for the "Car" class, 0.80 for the "Person" class, and 0.82 for the "Animal" class, which represent good discrimination performance for all the categories. The curves lie well above the random classification diagonal line, which proves that the model is far superior to chance. In general, the ROC curves and AUC values illustrate that the DWT feature-based K-RCNN model accurately separates various moving objects with high reliability.

V CONCLUSION

In the research work, a new architecture integrating Discrete Wavelet Transform (DWT) and Kernelized Region-Based Convolutional Neural Networks (KR-CNN) was introduced for video moving scene feature extraction and classification. The employment of DWT allowed for efficient multi-resolution analysis, successfully extracting both coarse and fine details of the moving objects, while the kernelized mapping improved the capability for feature discrimination in complicated scenes. Experimental findings proved that the suggested approach achieved better performance compared to traditional CNN and R-CNN methods with an accuracy rate of 97.84% and a lowered inference time. The incorporation of region-based kernel mapping greatly enhanced the model's performance in classifying dynamic scenes with high recall and precision, demonstrating the scalability and robustness of the proposed system. In general, the proposed method provides a promising avenue for real-time, high-accuracy video analysis applications. Future work involves incorporating temporal models such as RNNs or Transformers for improved motion understanding and optimizing kernel operations for real-time edge deployment. Testing under varied conditions like low-light environments and incorporating explainable AI methods will further enhance the robustness and usability of the model.

References

[1] N. P. Rana et al., Digital and Social Media Marketing: Emerging Applications and Theoretical Development. Springer Nature, 2019.

[2] H. J. Prasad et al"A Robust Deep Learning Model to Predict Epileptic Seizures based on Electroencephalographic (EEG) Signals," (ICCES), Coimbatore, India, 2024, pp. 1393-1399, doi: 10.1109/ICCES63552.2024.10859429.

[3] B. Shneiderman, Human-centered AI. Oxford University Press, 2022.

[4] Y. Lu, N. Vincent, P. C. Yuen, W.-S. Zheng, F. Cheriet, and C. Y. Suen, Pattern Recognition and Artificial Intelligence: International Conference, ICPRAI 2020, Zhongshan, China, October 19–23, 2020, Proceedings. Springer Nature, 2020.

[5] P. Mohan et al, "Detection and Classification of Brain Tumor using Fine Tuned Mobile Net Algorithm," (ICONAT), GOA, India, 2024, pp. 1-5, doi: 10.1109/ICONAT61936.2024.10775053.

[6] P. Balasubramaniam, K. Ratnavelu, G. Rajchakit, and G. Nagamani, Mathematical Modelling and Computational Intelligence Techniques: ICMMCIT-2021, Gandhigram, India February 10–12. Springer Nature, 2022.

[7] B. Soni and P. K. Das, Image Copy-Move Forgery Detection: New Tools and Techniques. Springer Nature, 2022.

[8] V. Singh, V. K. Asari, S. Kumar, and R. B. Patel, Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 2. Springer Nature, 2020.

[9] U. Kose and J. Alzubi, Deep Learning for Cancer Diagnosis. Springer, 2021.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
https://doi.org/10.62643/ijerst.2025.v21.i2.pp1337-1345 Vol. 21, Issue 2, 2025

[10] P. Karuppusamy, I. Perikos, F. Shi, and T. N. Nguyen, Sustainable Communication Networks and Application: Proceedings of ICSCN 2020. Springer Nature, 2021.

[11] J. M. R. Tavares, P. Dutta, S. Dutta, and D. Samanta, Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021. Springer Nature, 2021.

[12] T. Ahamed, IoT and AI in Agriculture: Self- sufficiency in Food Production to Achieve Society 5.0 and SDG's Globally. Springer Nature, 2023.

[13] W. L. Hamilton, Graph Representation Learning. Springer Nature, 2022.

[14] A. A. A. Samhan, et al "Empirical Assessment of Identifying Human Blood Group Based on Image Processing Assisted Deep Learning Principles," (ICSES), Chennai, India, 2024, pp. 1-6, doi: 10.1109/ICSES63760.2024.10910647.

[15] W. Yang, S. Sun, Y. Song, P. Du, Y. Hao, and J. Wang, Artificial Intelligence-Based Forecasting and Analytic Techniques for Environment and Economics Management. Frontiers Media SA, 2022.

[16] P. De and M. J. Pal, Computer Vision and Image Processing: Methods and Applications. Chyren Publication, 2025.

[17] D. Harvey, H. Kar, S. Verma, and V. Bhadauria, Advances in VLSI, Communication, and Signal Processing: Select Proceedings of VCAS 2019. 2021.

[18] A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings. Springer Nature, 2020.