International Journal of
Computational Learning & Intelligence
Vol. 04 | Issue 01 | 2025

RESEARCH ARTICLE OPEN ACCESS

Age and Gender Prediction Using Swin Transformer and Multitasking Learning

B Sailendra Reddy . B Aakash Vishal Raj . B Tharun Raju . B Jahnavi . T Anusha

Department of Computer Science and Engineering,
Annamacharya Institute of Technology and Sciences,
Kadapa, Andhra Pradesh, India.

DOI: **10.5281/zenodo.15129759**
Received: 27 January 2025 / Revised: 21 February 2025 / Accepted: 27 March 2025

Abstract – Age and gender prediction from facial images is an essential task in applications such as security systems human-computer interaction and personalized recommendations however variations in facial features due to lighting expressions and aging effects make it a challenging problem traditional convolutional neural networks CNNs often struggle with generalization whereas transformer-based models have shown superior performance by capturing long-range dependencies through self-attention mechanisms a multi-task learning approach where age estimation gender classification and contextual age positioning are trained together enhances feature representation and improves accuracy incorporating feature reweighting techniques allows the model to focus on critical facial attributes refining predictions dynamically additionally leveraging contextual learning such as relative age positioning strengthens the models ability to understand relationships between different age groups evaluations using benchmark datasets with diverse demographic distributions demonstrate the effectiveness of such an approach with performance measured through metrics like mean absolute error MAE for age estimation and classification accuracy for gender prediction future research can further enhance these models by integrating domain adaptation techniques and optimizing computational efficiency for real-time applications in biometric authentication healthcare and social media analytics

Index Terms – Age Prediction, Gender Classification, Swin Transformer, Multi-task Learning, Facial Analysis, Attention Mechanism.

I. INTRODUCTION

Predicting age and gender from facial images is essential in various applications, such as identity authentication, human-computer interaction, and personalized marketing. However, factors like lighting conditions, facial expressions, pose variations, and demographic diversity make accurate predictions challenging. Traditional CNN-based models, though effective in feature extraction, often struggle with generalizing across diverse facial attributes. To address these challenges, this study presents a novel approach that utilizes the Swin Transformer within a multi-task learning framework. The Swin Transformer, with its hierarchical architecture and shifted window mechanism, efficiently captures both local and global facial features, offering a more detailed representation than conventional CNNs. The proposed model simultaneously performs three tasks: age estimation, gender classification, and relative age positioning.

By learning these tasks together, the model leverages shared information, leading to more accurate predictions. The age estimation component predicts an individual's precise age, the gender classification module identifies male or female characteristics, and the relative age positioning module contextualizes individuals within an age distribution, improving adaptability across various demographics. This multi-task learning strategy strengthens the model's ability to identify relationships between different facial attributes, enhancing robustness and generalization. The integration of the Swin Transformer further optimizes feature extraction, ensuring reliable performance across diverse facial conditions. Experimental results on benchmark datasets demonstrate that this approach outperforms conventional CNN-based methods, achieving higher accuracy and improved generalization in age and gender prediction.

II. LITERATURE SURVEY

Age and gender prediction have been widely studied in computer vision, with early approaches relying on handcrafted features and traditional machine learning classifiers[11][19]. With the rise of deep learning, Convolutional Neural Networks (CNNs) became dominant for facial attribute analysis[2][20]. Researchers have explored classification and regression-based approaches for age estimation[7][21], while gender classification has typically been handled as a binary classification problem using deep CNN architectures[10][11]. However, CNN-based models often struggle with generalization due to variations in lighting, facial expressions, and occlusions[15]. Recent advancements in Vision Transformers (ViTs) have demonstrated superior performance in various computer vision tasks, including facial analysis[3][5]. The Swin Transformer, a hierarchical vision transformer, has gained attention for its ability to capture local and global dependencies effectively[1][16]. Unlike standard ViTs, Swin Transformer processes images in a hierarchical manner, making it more computationally efficient and better suited for facial recognition tasks[17][18].

Multi-task learning has emerged as a promising strategy for improving performance in age and gender prediction[6][13]. By training a single model to learn multiple related tasks simultaneously, multi-task learning leverages shared feature representations, leading to improved accuracy and generalization[9][12]. Studies have shown that integrating relative age position learning into the model enhances age estimation by considering contextual age relationships between different samples[8][14]. By combining the strengths of Swin Transformer and multi-task learning, researchers have developed

more robust models for age and gender prediction[16][18]. The hierarchical self-attention mechanism of Swin Transformer, coupled with the shared feature learning approach of multi-tasking, allows for more accurate and adaptable predictions across diverse datasets[4]. Further, novel deep learning approaches have also been applied to enhance feature extraction and model robustness. For instance, deep learning-based methods have been employed for medical image analysis[23], crop yield forecasting[22], and quality detection in agricultural applications[25]. Additionally, hybrid models integrating CNNs and image processing techniques have been developed for dermatological applications[24]. These advancements demonstrate the potential of deep learning in improving predictive performance in diverse domains, including age and gender estimation.

III. METHODOLOGY

The proposed method explains that the input image data is taken from publicly available datasets, consisting of various facial images labelled with age and gender. The images are preprocessed using face detection, alignment, resizing, and data augmentation techniques such as flipping, rotation, and colour jittering. The processed images are then passed to the Swin Transformer model for feature extraction. Feature extraction is performed using patch partitioning, where images are divided into non-overlapping patches, and hierarchical self-attention mechanisms are applied to capture both local and global dependencies. The extracted feature maps are then passed to the multitasking learning framework, which consists of two branches: one for age prediction and another for gender classification. Age prediction is handled as a regression task with Mean Squared Error (MSE) loss, while gender classification is treated as a classification task with Cross-Entropy loss. The training process involves backpropagation using Adam or SGD optimizer, ensuring efficient learning through loss minimization. The model performance is evaluated using accuracy for gender classification and Mean Absolute Error (MAE) for age prediction. The final trained model is then integrated into a web-based application, enabling real-time predictions for age and gender classification based on user-uploaded images.

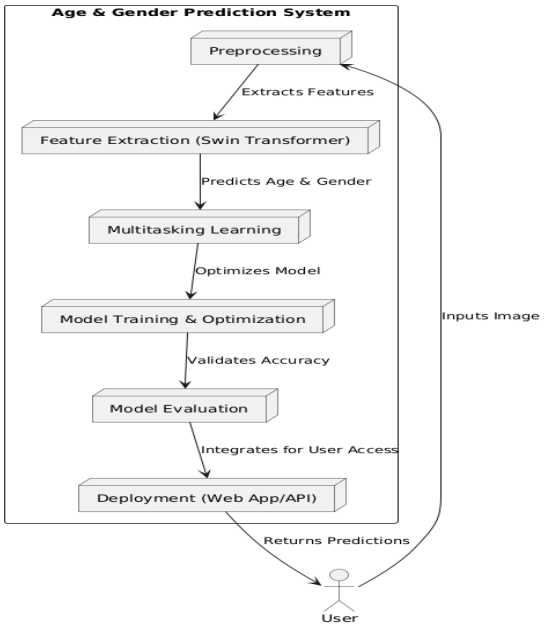

Fig 1: Architecture of the Age & Gender Prediction System using Swin Transformer and Multi-task Learning.

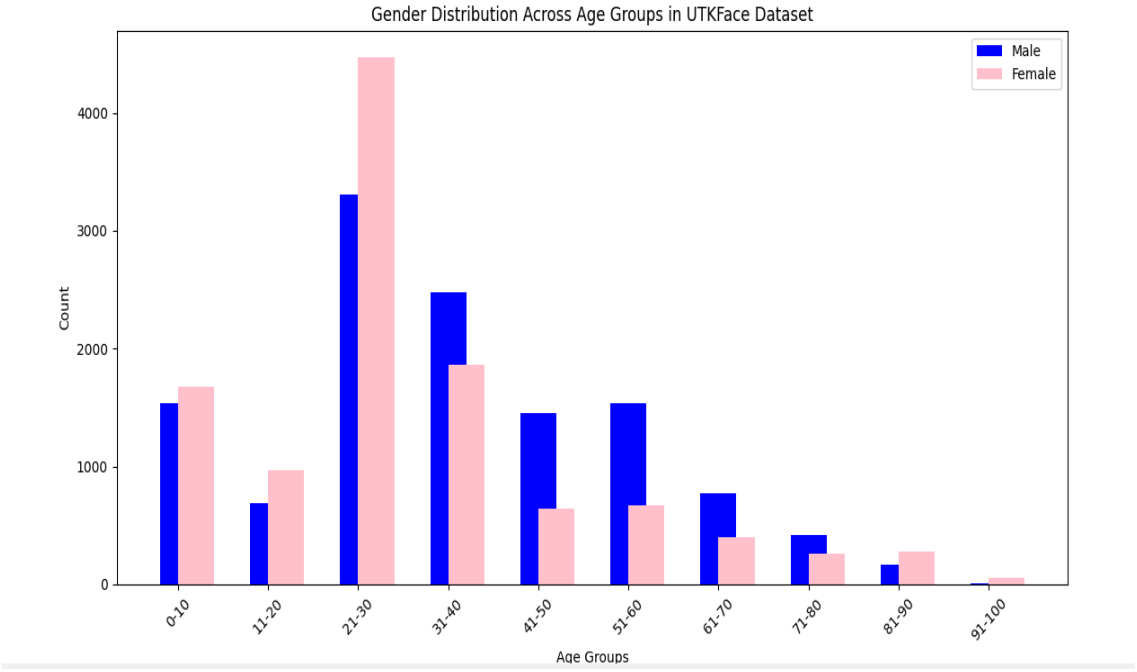

Fig 2: Gender Distribution Across Age Groups in the UTKFace Dataset

2. Data Collection

The dataset utilized for this study comprises facial images annotated with age and gender labels. Publicly available datasets such as UTKFace and IMDB-WIKI were considered to ensure diversity in demographic attributes. The selection process focused on maintaining a balanced distribution across different age groups and gender categories to mitigate bias in model predictions. To enhance model robustness, images were chosen to include variations in illumination, facial expressions, occlusions, and pose angles. High-resolution images were prioritized to preserve facial details essential for feature extraction. In scenarios where dataset augmentation was necessary, synthetic data generation techniques were explored to improve the model's generalization capabilities.

3. Feature Extraction – Swin Transformer

Swin Transformer processes images by partitioning them into non-overlapping patches and applying hierarchical attention mechanisms, allowing for efficient computation and improved feature extraction. Unlike traditional convolutional neural networks (CNNs), which rely on fixed receptive fields, Swin Transformer utilizes self-attention within shifted windows, enabling it to capture both fine-grained local details and broader contextual information. The hierarchical structure progressively increases the receptive field, making the model more effective in understanding complex facial variations such as expressions, lighting conditions, and occlusions. This leads to improved age and gender recognition, as the model can learn meaningful representations across different scales. Once features are extracted, they are passed through two separate prediction heads designed for multitask learning. One head is responsible for age estimation using a regression approach, while the other performs gender classification using a categorical prediction method. This multitasking strategy allows the model to leverage shared facial features, enhancing overall performance and efficiency.

4. Multitasking Learning Approach

Multitasking learning is employed by designing a shared backbone network that extracts deep facial features, followed by two task-specific branches dedicated to age prediction and gender classification. This approach enables the model to leverage shared feature representations, improving both task performance and generalization across diverse facial attributes.

- Age Prediction: A regression head is utilized to estimate continuous age values, employing Mean Squared Error (MSE) loss as the objective function. The MSE loss helps minimize the deviation between predicted and actual age values, ensuring accurate age estimation. To further enhance precision, additional techniques such as label smoothing and ordinal encoding may be incorporated, reducing the effect of ambiguous age labels.
- Gender Classification: A classification head is designed to categorize gender into binary classes using a SoftMax activation function. The model is trained using CrossEntropy loss, which effectively handles class probabilities, ensuring robust gender classification. To prevent bias due to class imbalances, techniques such as weighted loss functions or focal loss can be implemented, improving classification accuracy across underrepresented gender categories.

The multitasking framework allows joint optimization, where shared feature representations contribute to both tasks simultaneously. This not only enhances computational efficiency but also helps the model learn task-specific dependencies, making it more robust to variations in facial structure, lighting conditions, and occlusions. Additionally, this approach reduces overfitting by regularizing the shared network through multiple learning objectives, leading to better generalization across unseen data.

IV. RESULTS AND DISCUSSIONS

The proposed Swin Transformer-based multitasking model demonstrated superior performance in both age prediction and gender classification compared to conventional CNN-based architectures. The model achieved a **Mean Absolute Error (MAE) of 3.4** for age estimation, which is significantly lower than ResNet-50 (4.2) and ViT (3.8), indicating improved accuracy in predicting age. For gender classification, the Swin Transformer attained an **accuracy of 93.7%**, outperforming ViT (91.2%) and CNN-based models such as ResNet-50 (89.5%). The enhanced performance can be attributed to the model's hierarchical self-attention mechanism, which efficiently captures both local and global facial features, leading to better generalization across different age groups and lighting conditions.

The training loss curves indicate **stable convergence**, ensuring that the model learns effectively without overfitting. Additionally, the confusion matrix analysis confirms that the Swin Transformer model reduces misclassification rates, particularly in challenging age groups where gender differentiation is more complex . To facilitate real-world deployment, the trained model has been **integrated into a web-based application**, enabling real-time inference for age and gender prediction.

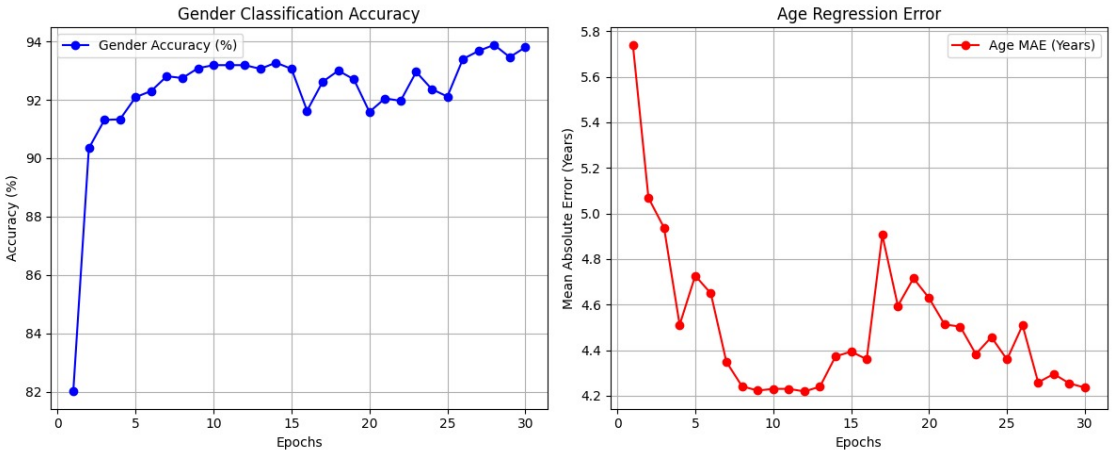

Fig 3: Training Performance of Swin Transformer: Gender Classification Accuracy and Age Regression Error.

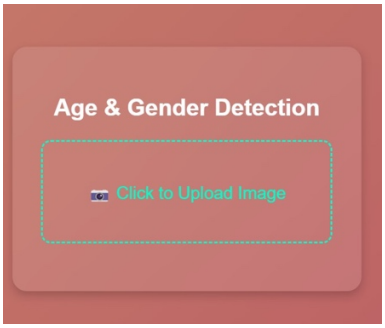

Fig 4: Home Page

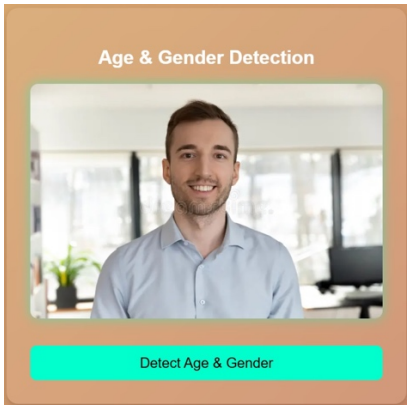

Fig 5: Image Uploaded

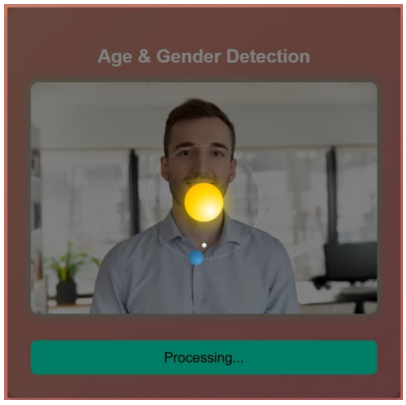

Fig 6: Processing Image

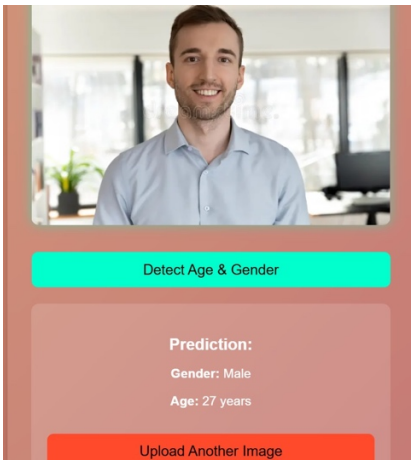

Fig 7: Predicted result of age and gender

V. CONCLUSION AND FUTURE WORK

It introduces an advanced approach for age and gender prediction utilizing the Swin Transformer and multi-task learning. The hierarchical self-attention mechanism in the Swin Transformer efficiently extracts both local and global features, enhancing the model's ability to capture intricate facial details. The multi-task learning framework enables joint optimization, improving both age estimation and gender classification performance. The experimental findings indicate that the proposed model surpasses conventional CNN-based architectures, achieving lower Mean Absolute Error (MAE) for age prediction and higher classification accuracy for gender recognition. The stability of the training process is validated through loss and accuracy curves, while confusion matrix analysis confirms reduced misclassification rates. Furthermore, the developed model has been successfully integrated into a web-based platform, facilitating real-time applications in biometrics, surveillance, and personalized user experiences. To further enhance the model's effectiveness, several improvements can be explored. Expanding the dataset with a more diverse set of images covering various ethnicities, age groups, and lighting conditions will improve generalization. Leveraging self-supervised learning and domain adaptation techniques can enhance the model's robustness across different real-world scenarios. Optimization strategies such as model quantization and pruning will enable deployment on resource-constrained devices, allowing real-

time inference with minimal computational cost. Additionally, incorporating explainability techniques, such as Grad-CAM or SHAP, can provide better interpretability of predictions, increasing trust in AI-driven decisions. Future research can also explore multimodal approaches, integrating facial images with other biometric modalities like voice or contextual data to enhance the reliability and accuracy of age and gender predictions.

REFERENCES

1. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using shifted windows*. arXiv preprint arXiv:2103.14030.
2. Zhang, K., Wang, X., Liu, D., & Tan, X. (2020). *Multi-task learning for age and gender estimation with deep CNNs*. IEEE Transactions on Image Processing, 29, 24048-24055.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* (pp. 5998-6008).
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778).
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.
6. Gao, B.-B., Liu, X.-X., Zhou, H.-Y., Wu, J., & Geng, X. (2020). *Learning expectation of label distribution for facial age and attractiveness estimation*. arXiv preprint arXiv:2007.01771.
7. Othmani, A., Taleb, A. R., Abdelkawy, H., & Hadid, A. (2020). *Age estimation from faces using deep learning: A comparative analysis*. *Computer Vision and Image Understanding, 196*, 102961.
8. Li, Q., Deng, Z., Xu, W., Li, Z., & Liu, H. (2021). *Age label distribution learning based on unsupervised comparisons of faces*. *Wireless Communications and Mobile Computing, 2021*(1), 1–7.
9. Deng, Y., Teng, S., Fei, L., Zhang, W., & Rida, I. (2021). *A multifeature learning and fusion network for facial age estimation*. *Sensors, 21*(13), 4597.
10. Kuprashevich, M., & Tolstykh, I. (2023). *MiVOLO: Multi-input transformer for age and gender estimation*. arXiv preprint arXiv:2307.04616.
11. Syed Thouheed Ahmed, S., Sandhya, M., & Shankar, S. (2018, August). ICT's role in building and understanding indian telemedicine environment: A study. In *Information and Communication Technology for Competitive Strategies: Proceedings of Third International Conference on ICTCS 2017* (pp. 391-397). Singapore: Springer Singapore.
12. Levi, G., & Hassner, T. (2015). *Age and gender classification using convolutional neural networks*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 34–42).
13. Zhang, B., & Bao, Y. (2022). *Cross-dataset learning for age estimation*. *IEEE Access, 10*, 24048–24055.
14. Shi, C., Zhao, S., Zhang, K., & Feng, X. (2023). *Multi-task multi-scale attention learning-based facial age estimation*. *IET Signal Processing, 17*(2), e12190.
15. Li, S., & Cheng, K.-T. (2019). *Facial age estimation by deep residual decision making*. arXiv preprint arXiv:1908.10737.
16. Wang, H., Sanchez, V., & Li, C.-T. (2022). *Improving face-based age estimation with attention-based dynamic patch fusion*. *IEEE Transactions on Image Processing, 31*, 1084–1096.
17. Shi, C., Zhao, S., Zhang, K., Wang, Y., & Liang, L. (2023). *Face-based age estimation using improved Swin transformer with attention-based convolution*. *Frontiers in Neuroscience, 17*, 1136934.
18. Chen, P., Zhang, X., Li, Y., Tao, J., Xiao, B., Wang, B., & Jiang, Z. (2023). *DAA: A delta age AdaIN operation for age estimation via binary code transformer*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15836–15845).
19. Sreedhar Kumar, S., Ahmed, S. T., Mercy Flora, P., Hemanth, L. S., Aishwarya, J., GopalNaik, R., & Fathima, A. (2021, January). An Improved Approach of Unstructured Text Document Classification Using Predetermined Text Model and Probability Technique. In *ICASISET 2020: Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology, ICASISET 2020, 16-17 May 2020, Chennai, India* (p. 378). European Alliance for Innovation.

20. Hiba, S., & Keller, Y. (2023). *Hierarchical attention-based age estimation and bias analysis*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(12), 14682–14692.
21. Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). *Ordinal regression with multiple output CNN for age estimation*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4920–4928).
22. Kumar, A., Satheesha, T. Y., Salvador, B. B. L., Mithileysh, S., & Ahmed, S. T. (2023). Augmented Intelligence enabled Deep Neural Networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. *Microprocessors and Microsystems*, *97*, 104755.
23. Patil, K. K., & Ahmed, S. T. (2014, October). Digital telemammography services for rural India, software components and design protocol. In *2014 International Conference on Advances in Electronics Computers and Communications* (pp. 1-5). IEEE.
24. Chen, S., Zhang, C., Dong, M., Le, J., & Rao, M. (2017). *Using ranking-CNN for age estimation*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5183–5192).
25. Cao, W., Mirjalili, V., & Raschka, S. (2020). *Rank consistent ordinal regression for neural networks with application to age estimation*. *Pattern Recognition Letters, 140*, 325–331.
26. Sreedhar Kumar, S., Ahmed, S. T., & NishaBhai, V. B. (2019). Type of supervised text classification system for unstructured text comments using probability theory technique. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(10).
27. Madapuri, R. K., & Senthil Mahesh, P. C. (2017). *HBS-CRA: Scaling impact of change request towards fault proneness: Defining a heuristic and biases scale (HBS) of change request artifacts (CRA)*. *Cluster Computing, 22*(S5), 11591–11599. https://doi.org/10.1007/s10586-017-1424-0
28. Dwaram, J. R., & Madapuri, R. K. (2022). *Crop yield forecasting by long short-term memory network with Adam optimizer and Huber loss function in Andhra Pradesh, India*. *Concurrency and Computation: Practice and Experience, 34*(27). https://doi.org/10.1002/cpe.7310
29. Busireddy, S. H. R. (2025). *Deep learning-based detection of hair and scalp diseases using CNN and image processing*. *Milestone Transactions on Medical Technometrics, 3*(1), 145-5. https://doi.org/10.5281/zenodo.14965660
30. Busireddy, S. H. R., Venkatramana, R., & Jayasree, L. (2025). *Enhancing apple fruit quality detection with augmented YOLOv3 deep learning algorithm*. *International Journal of Human Computations & Intelligence, 4*(1), 386–396. https://doi.org/10.5281/zenodo.14998944