Notification No: COE/ Ph.D.(Notification)/591/2025

Notification Date: 30-12-2025

Name of Scholar: **Aqib Nazir Mir**

Name of Supervisor: **Dr. Danish Raza Rizvi**

Name of Department: **Computer Engineering**

Topic of Research: **Generalizable and Explainable Deep Learning In Medical Imaging**

FINDINGS

This thesis advances medical image analysis by developing accurate, interpretable, and computationally efficient artificial intelligence methodologies across histopathology, radiology, and neuroimaging. A central finding is that explainability, robustness, and efficiency can be jointly optimized, addressing key barriers to clinical adoption such as model opacity, limited generalization, and resource constraints.

The first study proposed **XViT**, an explainable Vision Transformer for histopathological image analysis that integrates attention-based, gradient-based, and model-agnostic explanation methods. To systematically assess interpretability, quantitative metrics faithfulness, complexity, and sensitivity-were introduced. Experimental evaluation on the **LCS25000 and KBSMC datasets** demonstrated strong classification performance, with **TransLRP** yielding the most faithful and clinically aligned explanations. This confirms that transformer-based models can deliver reliable interpretability when explanation mechanisms are embedded by design.

The second study addressed automated medical report generation through a **self-boosting multimodal framework** that couples report generation with an image–text–label alignment module. This cooperative learning strategy consistently improved report quality on **IU-Xray, MIMIC-CXR**, and related datasets, producing more fluent and clinically coherent reports than baseline methods. Ablation studies verified that the auxiliary alignment task played a critical role, highlighting the importance of multimodal consistency for trustworthy clinical reporting.

The third study focused on deployment efficiency by introducing a **cross-architecture knowledge distillation framework** enabling knowledge transfer between heterogeneous teacher–student models, including CNNs and Transformers. The distilled models maintained diagnostic accuracy while significantly reducing model size and computational cost. Additionally, the student networks exhibited attention maps that were more focused on clinically relevant regions, demonstrating that efficiency gains can be achieved without sacrificing interpretability.

The fourth study developed an **EfficientNet-B0 model augmented with CBAM** for brain tumor MRI classification. The model achieved near state-of-the-art accuracy with low computational overhead. Visual explanations generated using **Grad-CAM and Grad-CAM++** were spatially precise and consistent with expert-defined tumor regions, reinforcing the clinical trustworthiness of the system.

Collectively, these findings establish an **explainability-by-design paradigm**, where interpretability is integrated into model architectures and supported by quantitative evaluation rather than post hoc visualization alone. The proposed methods demonstrate cross-domain and multimodal generalization across histopathology, radiology, ophthalmology, and neuroimaging tasks. Through attention-efficient architectures and knowledge distillation, the work enables resource-efficient deployment while preserving clinical relevance. Overall, the thesis provides a practical foundation for developing trustworthy, interpretable, and deployable AI systems that align closely with clinical reasoning and diagnostic workflows.