IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

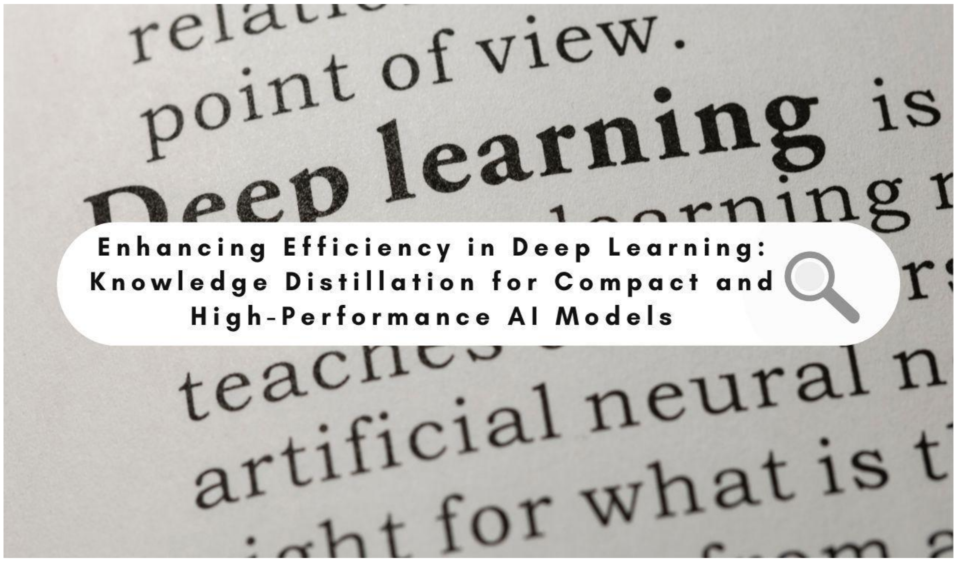
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 7, March 2025

Enhancing Efficiency in Deep Learning: Knowledge Distillation for Compact and High-Performance AI Models

Perumalsamy Ravindran
Anna University, India



Abstract*:* *The exponential growth of transformer-based language models has created significant challenges for their practical deployment, particularly on resource-constrained devices. This article explores knowledge distillation as a solution for creating efficient, compact models while maintaining high performance. We examine various distillation techniques, including logit-based, feature-based, and attention-based approaches, demonstrating their effectiveness in model compression. Through comprehensive case studies of DistilBERT and TinyBERT implementations, we analyze the trade-offs between model size, inference speed, and accuracy. The article also investigates implementation considerations, experimental results, and future directions, including adaptive distillation and reinforcement learning integration. The findings suggest that knowledge distillation offers a promising pathway for democratizing AI by making powerful models accessible across diverse hardware platforms while maintaining acceptable performance levels.*

Keywords: Knowledge Distillation, Model Compression, Transformer Architecture, Edge Computing, Neural Network Optimization