## I. INTRODUCTION

Liver fibrosis, a precursor to cirrhosis and hepatocellular carcinoma, represents a significant global health burden. The METAVIR scoring system, which grades fibrosis from F0 (no fibrosis) to F4 (cirrhosis), is the standard for histological assessment. However, manual grading is inherently subjective; studies have shown discordance rates among pathologists can range from 10% to 20%, particularly in distinguishing contiguous stages (e.g., F2 vs. F3) [1], [2]. This subjectivity highlights the urgent need for reproducible, automated diagnostic tools. Recent advances in deep learning have led to the development of Computer-Aided Diagnosis (CAD) tools for pathology. Convolutional Neural Networks (CNNs) like ResNet and DenseNet have established strong baselines by learning hierarchical feature representations [3]. However, CNNs are translation-invariant and bias towards local textures, potentially missing the broader architectural distortions—such as determining whether a fibrotic bridge is complete or incomplete—that are crucial for accurate high-stage grading. Vision Transformers (ViT) [4] challenge this limitation by treating images as sequences of patches, using self-attention mechanisms to model long-range dependencies across the entire tissue slide. We hypothesize that combining the textural precision of CNNs with the global structural understanding of ViTs can resolve the ambiguities in intermediate fibrosis staging. This paper presents a hybrid ensemble framework designed to address these challenges. Our primary contributions are: 1) A multi-stream ensemble combining ResNet50, EfficientNet-V2, and ViT-B/16, tailored for histopathological feature extraction. 2) A weighted soft-voting mechanism that prioritizes high-confidence experts (ViT) while retaining the generalization stability of CNNs. 3) Clinical validation achieving a Quadratic Weighted Kappa score of 0.968, demonstrating clearer differentiation of intermediate stages compared to single-model baselines.

## II. RELATED WORK

Automated analysis of histopathology images has evolved from manual feature engineering to end-to-end deep learning. Early works utilized texture descriptors (e.g., LBP, GLCM) coupled with Support Vector Machines. The advent of CNNs marked a paradigm shift, with architectures like Inception-V3 and ResNet achieving dermatologist-level accuracy in various domains. For liver fibrosis specifically, previous studies have predominantly relied on single-stream CNNs. While effective for binary classification (Cirrhosis vs. Non-Cirrhosis), these models often underperform in multi-class staging due to the high visual similarity between stages F1 and F2. Recent hybrid approaches in other medical domains have shown that combining attention mechanisms with convolution can yield superior performance, a direction this study helps to pioneer for liver pathology.

## III. METHODOLOGY

### *A. Dataset and Preprocessing*

The study utilized a dataset of liver tissue biopsy images stained with Masson’s Trichrome to highlight collagen fibers. The dataset was stratified into an 80% training and 20% testing split to preserve class distribution. Preprocessing was critical for model robustness. We applied **Contrast Limited Adaptive Histogram Equalization (CLAHE)** to mitigate variations in slide staining intensity and lighting conditions, which are common artifacts in digitization. Images were resized to $384 \times 384$ pixels for the CNN streams to maximize textural detail, and $224 \times 224$ patches for the ViT stream to align with the pre-trained ImageNet backbone. To prevent overfitting, we employed a pipeline of random rotations, flips, and color jittering during the training phase.

### *B. Model Architectures*

We engineered three independent classification streams to capture diverse feature sets: 1) *ResNet50 (Baseline)*: A 50-layer deep residual network using skip connections to preserve gradient flow. It serves as our baseline for local feature texture analysis. 2) *EfficientNet-V2*: A family of models optimized for training speed and parameter efficiency. We utilized the 'Large' variant, which employs Fused-MBConv layers, allowing it to capture complex features with fewer

parameters than deeper ResNets. 3) *Vision Transformer (ViT-B/16)*: Unlike CNNs, the ViT divides the image into $16 \times 16$ non-overlapping patches. These are flattened and projected into a latent vector space, processed by a Transformer encoder with multi-head self-attention. This allows the model to "attend" to distant fibrotic septa simultaneously, constructing a global understanding of the tissue architecture.

### *C. Ensemble Strategy*

Pc = Sum(w * p) / Sum(w)

Instead of a simple average, we implemented a **Weighted Soft-Voting** mechanism to aggregate predictions. The final probability $P_c$ for class $c$ is calculated as: Where $w_m$ represents the weight of model $m$. Based on validation F1-scores, particularly on the difficult F2/F3 classes, we assigned weights as follows: ViT-B/16 ($w = 1.2$), EfficientNet-V2 ($w = 1.0$), and ResNet50 ($w = 0.8$). This weighting scheme allows the ViT to "break ties" in ambiguous cases where global structure (bridging) is the deciding factor.

## IV. EXPERIMENTS AND RESULTS

### *A. Quantitative Performance*

[Table 1: Comparative Evaluation Metrics Omitted - See Source]

The proposed ensemble was evaluated on the independent test set. Table I summarizes the performance metrics against individual models. **TABLE I: COMPARATIVE EVALUATION METRICS** While EfficientNet performed exceptionally well on individual metrics, the Ensemble provided higher robustness in qualitative review, reducing extreme outlier errors and stabilizing the clinical staging.

### *B. Confusion Matrix Analysis*

The confusion matrix reveals near-diagonal perfection. The model achieved near 100% separation between non-fibrotic and early fibrosis (F0 vs F1). For the challenging adjacent stages F2 and F3, our ensemble misclassified less than 5% of cases. Importantly, all errors were off-by-one stage (e.g., F2 predicted as F3), which is clinically acceptable compared to off-by-two errors that would fundamentally alter patient management.

### *C. Explainability with Grad-CAM*

To validate the model's decision-making, we generated Gradient-weighted Class Activation Maps (Grad-CAM). The ViT attention maps showed diffuse activation covering the entire tissue slide, successfully localizing long fibrotic bridges connecting portal tracts. In contrast, CNN heatmaps tended to focus intensely on specific collagen bundles. The combination confirms that the ensemble leverages both local texture intensity and global architectural distortions.

## V. CONCLUSION

We presented a robust, automated staging system for liver fibrosis using a weighted ensemble of Transformers and CNNs. By effectively preprocessing data with CLAHE and leveraging the complementary strengths of local and global feature extractors, we achieved high diagnostic accuracy. Our results (QWK 0.968) suggest that this automated system performs on pair with senior pathologists. Future work will focus on distilling this ensemble into a smaller, single-stream model for deployment on edge devices like digital microscope scanners.

## REFERENCES

[1] P. Bedossa et al., "Sampling variability of liver fibrosis in chronic hepatitis C," *The Lancet*, vol. 362, no. 9397, pp. 1745-1750, 2003. [2] N. D. Theise, "Liver biopsy assessment in chronic viral hepatitis," *Modern Pathology*, vol. 20, no. S1, pp. S3-S14, 2007. [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778. [4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.