



RESEARCH ARTICLE

# HMARNET—A Hierarchical Multi-Attention Residual Network for Gleason scoring of prostate cancer

R. Karthik<sup>1</sup>  | R. Menaka<sup>1</sup>  | M. V. Siddharth<sup>2</sup> | Sameeha Hussain<sup>3</sup> |  
P. Siddharth<sup>3</sup> | Daehan Won<sup>4</sup>

<sup>1</sup>Centre for Cyber Physical Systems,  
Vellore Institute of Technology, Chennai,  
India

<sup>2</sup>School of Mechanical Engineering,  
Vellore Institute of Technology, Chennai,  
India

<sup>3</sup>School of Electronics Engineering,  
Vellore Institute of Technology, Chennai,  
India

<sup>4</sup>System Sciences and Industrial  
Engineering, Binghamton University,  
Binghamton, New York, USA

## Correspondence

R. Karthik, Centre for Cyber Physical  
Systems, Vellore Institute of Technology,  
Chennai, India.  
Email: [r.karthik@vit.ac.in](mailto:r.karthik@vit.ac.in)

## Abstract

Manual delineation of prostate cancer (PCa) from whole slide images (WSIs) demands requires pathologists with adequate domain knowledge. This process is generally strenuous and may be subjected to poor inter-pathologist reproducibility. Accurate Gleason scoring is an important step in the computer-aided diagnosis of PCa. This work proposes a novel lightweight convolutional neural networks (CNN) to extract significant hierarchical features from the histopathology images. It learns meticulous attention-guided feature representations through the convolutional layers for precise scoring of Gleason grades. The Hierarchical Multi-Attention Residual (HMAR) block extracts attention-guided features and fuses the resulting feature maps from multiple levels with a reduced number of trainable parameters. We have also proposed a new lightweight channel-attention module, enhanced channel attention (ECA) to extract inter-channel features from different receptive fields. With the presence of different attention mechanisms, the network learns to focus on more significant features rather than unnecessary ones. Additionally, mixed FocalOHEM loss is proposed to optimize the CNN and efficiently minimize the error. While the focal loss helps to address the class imbalance present in the TCGA-PRAD dataset, the online hard example mining (OHEM) loss focuses on optimizing hard negative samples. It achieved an accuracy of 89.19% with a Kappa score of 86% on the TCGA-PRAD dataset.

## KEYWORDS

attention, deep learning, histopathology, prostate cancer, TCGA-PRAD

## 1 | INTRODUCTION

Prostate cancer (PCa) is the second leading cause of death by cancer in men. Its prevalence has risen significantly in most developed countries in recent years.<sup>1</sup> Like most cancers, it develops as a result of a complex interplay of genetic and epigenetic factors, both of which can be influenced by environmental risk variables.<sup>2</sup> The prostate is responsible for the generation of seminal fluid.<sup>3</sup>

The cancer develops when cells in the prostate gland begin to proliferate uncontrollably.<sup>4</sup> Almost all PCas are adenocarcinomas and therefore a therapeutic intervention is necessary to treat PCa at an earlier stage.

Pathologists employ several screening procedures to identify different types of PCas. They use a digital microscope to analyze whole-slide images (WSIs) of PCa tissue in prior investigations.<sup>5</sup> Numerous challenges are faced in detecting irregularities in biopsy data as manual

processing consumes a significant amount of time. It causes therapy to be delayed and is therefore inefficient. The diagnosis and prognosis of PCa are strongly reliant on early detection and rapid identification of the irregularities. Biomedical imaging is hence critical for the accurate diagnosis and treatment of cancer. Technology is advancing in every industry, including the medical field for automated diagnosis. The usage of computer-aided diagnosis (CAD) has recently risen to assist clinicians in making credible judgments.

The past decade has seen a significant advancement in machine learning (ML) and deep learning (DL) algorithms, which have produced remarkable results in biomedical image processing. This advancement benefit CAD applications. ML researchers have shifted their attention in recent years targeting biological challenges that are difficult to investigate using traditional methods. DL has had a lot of success with large-scale medical imagery like histopathology slides as it makes extraction of high-level information from images possible.<sup>6,7</sup> DL approaches may make it feasible to obtain high detection accuracy and do not require clinical guidance to select hand-crafted features, as features can be extracted during training. With the help of massively parallel computers with GPUs, DL approaches have gained enormous popularity in PCa diagnosis and Gleason grading in recent years.<sup>8,9</sup> The Gleason score (GS) seeks to describe and quantify the regularity of gland patterns.<sup>10,11</sup>

## 2 | RELATED WORKS

Several studies on the detection of PCa have been published in recent years. This study proposes an analysis of various ML and transfer learning-based approaches applied to the Gleason grading of PCa. ML can find hidden patterns and intrinsic structures in input data making it highly efficient in the task proposed to it. They also have the ability to learn without being explicitly programmed to do so. Alkhateeb et al. proposed a ML-based method to classify samples of different GSs using a hierarchical prediction model.<sup>12</sup> DL's capability to automate feature extraction and fine-tune the hyperparameters for the intended result is one of its main benefits over ML. Through their hidden layers, DL networks can progressively learn important features. Additionally, they are adaptable to future problems and therefore have been widely employed in recent works. Sebastian et al. presented a systematic comparison of weakly supervised models which include Shallow CNN, VGG-19, and DenseNet-121 for the classification of the five grades of GS. It was inferred that the DenseNet-121 architecture

offered the best classification results among the others.<sup>13</sup> In Ref. 14, Sebastian et al. presented a semi-weakly supervised model for classifying PCa tissue. The performance of the fully-supervised CNN is significantly improved by the teacher-student approach. This approach was applied at the Gleason pattern level in tissue microarrays (TMAs) as well as at the GS level in WSIs. Among the various CNN models that were tested for the student network, DenseNet-121 displayed the best performance in the validation TMA set. In Ref. 15, Lara et al. proposed an information fusion approach for the automatic classification of prostate histopathology WSIs. The approach employs two versions of weakly-supervised ML models. It could be inferred that the semantic-enhanced M-LSA model outperforms the previous techniques for retrieving images. Wang et al. proposed modern computational analytical methods to identify potential multi-omics biomarkers for the early detection of PCa. Auto-encoders were used with the prediction model for better performance in predicting the prognosis of PCa. Overall, the SVM classifier using the auto-encoder model had the best prediction accuracy and outperformed the ones using multi-omics features.<sup>16</sup> In Ref. 17, Otolara et al. performed Gleason pattern classification based on a pretrained MobileNet architecture with an emphasis on weakly supervised learning. Xu et al. implemented a multi-class support vector machine (SVM) to classify various GS samples and also used VGG16 for feature extraction.<sup>18</sup> Kallen et al. used the OverFeat pre-trained convolutional network for feature extraction. Random forest and SVM were employed to classify benign tumors and tumors with Gleason grades of 3–5.<sup>19</sup>

In Ref. 11, Zhang et al. proposed a novel sampling framework based on spatial and magnification attention for classification tasks. Inception V3 networks were used as the feature extractors. Otolara et al. presented a comparison between two CNN-based regression approaches-DenseNet-BC 121 and ShallowNet, to learn the magnification of histopathology images. For internal evaluation, the best magnification regressor was a linear combination of two DenseNets.<sup>20</sup> Abbasi et al. proposed an efficient DL network based on GoogleNet to detect PCa in Magnetic resonance imaging (MRI).<sup>21</sup> Wildeboer et al. employed a CAD modeling interface to better understand the radiomics of multiparametric PCa imaging.<sup>22</sup> Tolkach et al. developed a DL model based on the NasNetLarge architecture for the detection of PCa.<sup>23</sup> Arvaniti et al. performed Gaussian filtering and Otsu thresholding to better amplify the separated tissue. MobileNet was finally employed to classify four Gleason grades.<sup>24</sup> Using an ensemble neural network, which depended on the inceptionV3 architecture, Strom et al.

accomplished classification between the binary and malignant tumor as well as prediction between the three Gleason grades.<sup>25</sup> Toro et al. did an experimental analysis using various DL models such as LeNet, AlexNet, and GoogLeNet. It could be inferred that GoogLeNet obtained the highest accuracy for the classification of various Gleason grades.<sup>10</sup> Șerbănescu et al. performed Gleason-grade classification for PCa patches using pre-trained networks of AlexNet and GoogleNet.<sup>26</sup> Zhang et al. proposed a high-performance method based on AlexNet to classify multiple sclerosis brain images.<sup>40</sup>

Following the advent of transfer learning for Gleason scoring, many custom CNN models have been implemented to optimize the process. In Ref. 27, Chakraborty et al. employed a custom-built CNN model called Prostate AttentionNet (ProstAttNet) to predict visual attention. Nagpal et al. developed a DL system to classify between three Gleason patterns and non-tumor. The system was based on a custom version of the InceptionV3 architecture and a categorical prediction system based on class selection using the highest calibrated likelihood.<sup>28</sup> Brunese et al. proposed an 8-layer custom CNN based on 71 various radiomic features that aimed to classify the GSs of the affected prostate regions using MRI scans.<sup>29</sup> Shin et al. proposed a custom CNN with a self-attentive normalization (SAN) layer which consists of a center-guided attention module to differentiate between benign tumors and Gleason grades of varying levels.<sup>30</sup> Wang et al. proposed a novel automatic Gleason grading system using local structure model learning and classification. Attributed graphs were used to represent tissue glandular structures. Representative subgraphs were constructed using a learning mechanism based on bag of words features from labeled samples of Gleason grades.<sup>8</sup> Zhang et al. proposed a multi-input deep convolutional neural network, which employed the Convolutional Block Attention Module (CBAM) to provide both spatial and channel attention. The network receives a 3D CT image and a 2D X-ray image as inputs.<sup>39</sup> In Ref. 39, Zhang et al. developed a novel AI-based system for improved detection of ductal carcinoma in thermographs. The CNN utilized exponential linear unit, rank-based weighted pooling, and L-way data augmentation for DCIS detection. Zhang et al. proposed a 13-layer CNN and employed data augmentation techniques for the classification of fruits.<sup>41</sup>

Reda et al. proposed a fusion segmentation model to extract imaging markers and to predict the diagnosis of the input prostate volume through a two-stage classification.<sup>31</sup> Mohsin et al. employed four different CNN architectures which include VGG19, ResNext50, MobileNetV2, and ResNet50 as an encoder based on the UNET model, and achieved optimum results in comparison to other state-of-the-art architectures.<sup>32</sup> In Ref. 33, Bulten et al.

employed a semi-supervised method to label the images and further an extended U-net was used to classify the scores. Li et al. proposed a dual-branch custom architecture consisting of a pretrained DeeplabV3 transfer learning module for image segmentation.<sup>34</sup> Hasan et al. presented an extensive review analysis of the existing methodologies as well as offered unique insights into the detection of PCa using DL.<sup>7</sup>

## 2.1 | Research gaps and motivation

The proposed work addresses the following research gaps in the Gleason-grade scoring of PCa from histopathology images.

1. A large number of channels on different levels contain rich semantic information, which needs to be extracted and refined without utilizing excessive computational power. Most existing deep CNN models overlooked the correlation of hierarchical features in histopathology images, leading to poor feature representation power of the network.
2. While many existing works employ attention modules to improve the performance of the network, using different receptive fields in the attention blocks opens up possibilities to extract salient multi-level features from histopathology images.
3. In the existing works, generic loss functions were utilized for optimizing deep CNN architectures. Due to the presence of class imbalance in the dataset, it is necessary to employ a different paradigm to tackle the issue. The loss function needs to be able to guide the neural network to learn features from hard classes for improved overall performance.

## 2.2 | Research contributions

The following are the main contributions made towards addressing the gaps in Gleason scoring of PCa.

1. The proposed network utilizes residual connections to extract hierarchical features from the input images for precise classification. Furthermore, these feature maps are guided by spatial, channel, and self-attention layers and are fused together for improved feature representation and refinement.
2. The proposed enhanced channel attention (ECA) module applies channel-wise attention to features extracted from different receptive fields. A residual link is also present in the ECA module for improved gradient flow and feature propagation through the HMAR block.

3. We have also proposed a new loss function, ‘Mixed FocalOHEM’ loss for identifying and training on hard negative samples from the dataset. This improves the class-wise results and overall performance of the network.

### 3 | PROPOSED METHODOLOGY

An illustration of the proposed CNN architecture is presented in Figure 1. The network consists of a sequence of multiple HMAR blocks for the effective extraction of multi-level features. Moreover, channel attention, spatial attention, and multi-headed self-attention layers are employed for better feature refinement and network representation. The attention mechanisms aid the network in focusing on salient information from the input histopathology images rather than unnecessary features. We have also proposed a novel channel attention mechanism, ECA block for extraction of channel-wise features from different receptive fields. The residual paths consisting of the self-attention layer in the HMAR block also regulate long-range gradient flow through the network. Convolutions on multiple levels were adopted to (1) facilitate extraction of salient hybrid features enhanced with attention, (2) boost the network representation power

through multi-level feature fusion, and (3) improve gradient and information flow across the network with multiple residual paths. Finally, the network is optimized by a new loss function called ‘Mixed FocalOHEM’ loss and classifies it into one of the GSs.

#### 3.1 | Proposed Hierarchical Multi-Attention Residual Network (HMARNet)

The proposed CNN operates on input images with spatial dimensions of  $128 \times 128$ . The first convolution layer is responsible for extracting low-level features from the images. These features are fed into the HMAR blocks which extract mainstream features progressively. The HMAR blocks facilitate the extraction of attention-guided features from multiple levels. The residual path is responsible for extracting positional information through the self-attention mechanism.<sup>35</sup> The schematic sketch of the proposed network is presented in Figure 2.

Furthermore, channel and spatial attention (CSA) layers are added progressively to the convolution layers in the HMAR block for effective feature extraction. The first feature extraction (FE) block does not contain any attention layers followed by the addition of the ECA module responsible for extracting channel-wise features

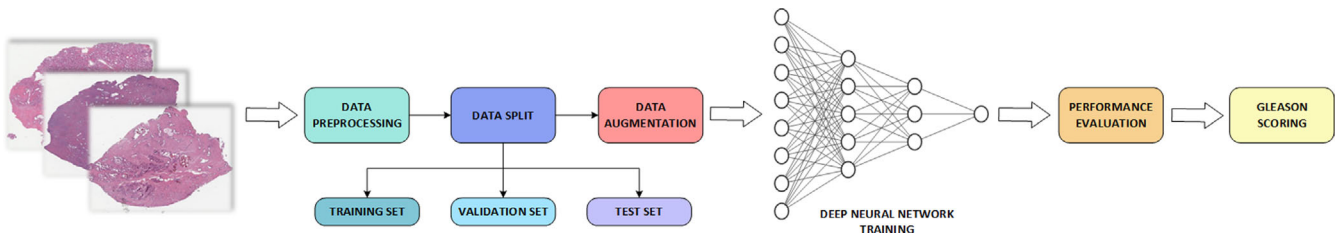


FIGURE 1 The overall workflow of the proposed methodology.

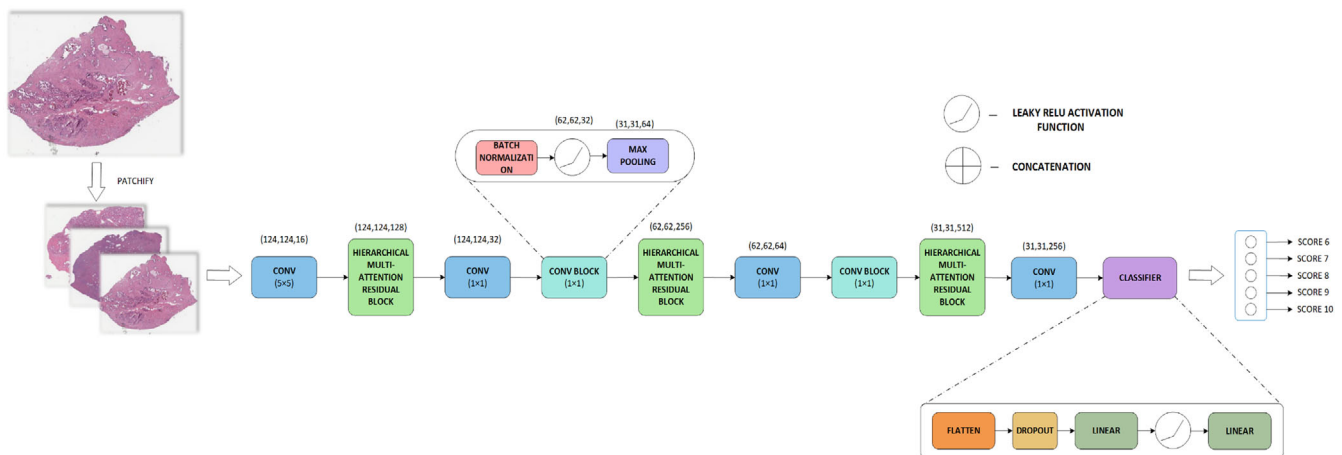


FIGURE 2 Architectural diagram of the proposed Hierarchical Multi-Attention Residual (HMAR) network.

from different receptive fields. Lastly, the third block contains the CSA module consisting of both CSA for the effective extraction of important features from the images.<sup>36</sup> The schematic flow of the different feature extraction blocks is illustrated in Figure 3. Here, the aim is to preserve spatial details from the input feature map while extracting semantic features through multiple convolution layers in the FE blocks. To achieve this, we have employed another path consisting of the C1 and C2 blocks. Specifically, these two blocks reduce the number of filters without sacrificing the spatial and geometric details. This way, our approach mitigates the potential loss of spatial details that may occur when using convolutional layers alone in the FE BLOCK, resulting in a more accurate and detailed output.

### 3.1.1 | Progressive attention-based feature extraction blocks

The HMAR block consists of three feature extraction blocks which extract mainstream features progressively with the addition of attention layers. The first feature extraction block has 2 convolution layers with a  $3 \times 3$  kernel with Leaky Relu activations in between them. The proposed ECA module is added at the end of the next feature extraction block for capturing multi-field channel-wise features. The final feature extraction block consists of the CSA module to efficiently extract inter-channel and inter-spatial features. This improves the overall representation power of the network. The spatial attention module is applied to the channel-refined feature input map calculated by the channel attention module. Important receptive fields of the cells from the histopathology images are selected by the spatial attention module for effective feature extraction. The blocks are illustrated in Figure 4.

### 3.1.2 | Enhanced channel attention (ECA)

The proposed attention module, ECA is responsible for optimizing features from multiple levels as presented in Figure 5. In order to extract important features with a low computation cost, the  $1 \times 3$  and  $3 \times 1$  convolutions are employed to extract features from different receptive fields.

The channel-wise attention is then calculated using the global average pooling layer. Moreover, a residual link is added to improve gradient flow and feature propagation through the HMAR module. Finally, element-wise multiplication is applied to the resulting feature map and the input feature map ( $f_{in}$ ) through the residual link. The above process is represented in Equation (1).

$$f = F_{GAP}(Conv_{1 \times 3}(Conv_{3 \times 1}(f_{in}))) \quad (1)$$

$$f_{out} = f_{in} \odot f \quad (2)$$

where  $f_{in}$  is the input feature map,  $Conv_{3 \times 1}$  and  $Conv_{1 \times 3}$  are the convolution layers with  $3 \times 1$  and  $1 \times 3$  kernels respectively,  $F_{GAP}$  is the global average pooling layer and  $f$  is the feature map generated before element-wise multiplication. The final output feature map of the ECA block is denoted by  $f_{out}$  in Equation (2).

### 3.1.3 | Mixed FocalOHEM Loss

We have proposed a new loss function to optimize and improve the network performance. The Mixed FocalOHEM loss is a combination of focal loss (FL) and online hard example mining (OHEM) loss.<sup>37,38</sup> FL is essentially cross-entropy loss with Alpha ( $\alpha$ ) and Gamma ( $\gamma$ ) parameters to tackle class imbalance in the dataset. The  $\alpha$  parameter in Equation (3) balances the importance of positive and negative samples in the dataset. The

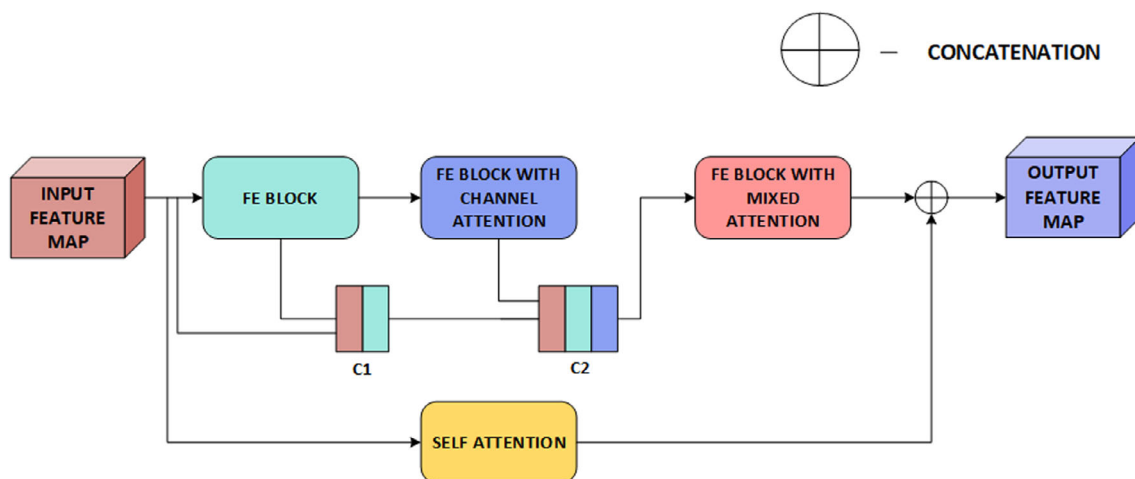


FIGURE 3 Schematic sketch of the Hierarchical Multi-Attention Residual (HMAR) block.



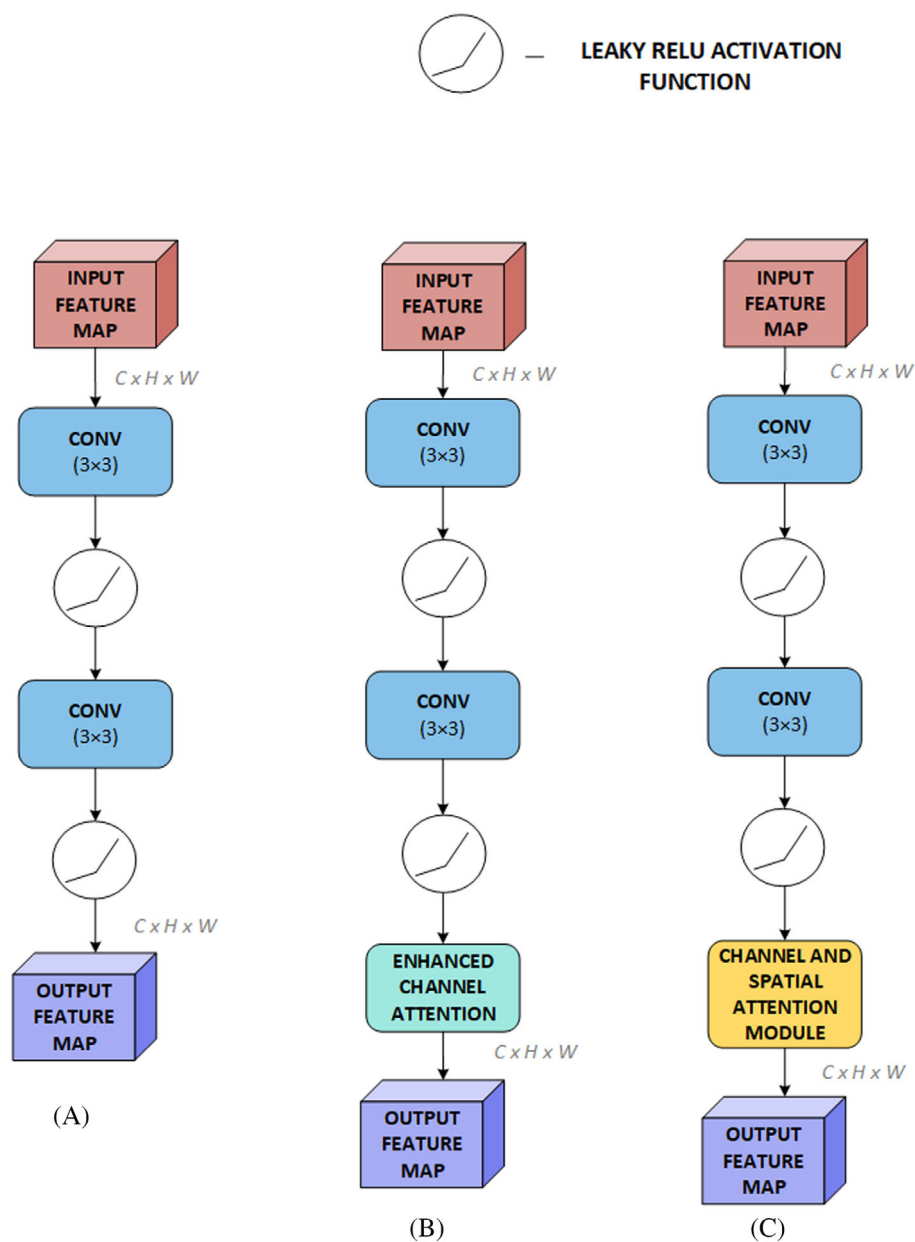


FIGURE 4 Block diagrams of the different feature extraction blocks employed in the HMAR module (A) FE block without attention (B) FE block with enhanced channel attention (ECA) and (C) FE block with channel and spatial attention (CSA).

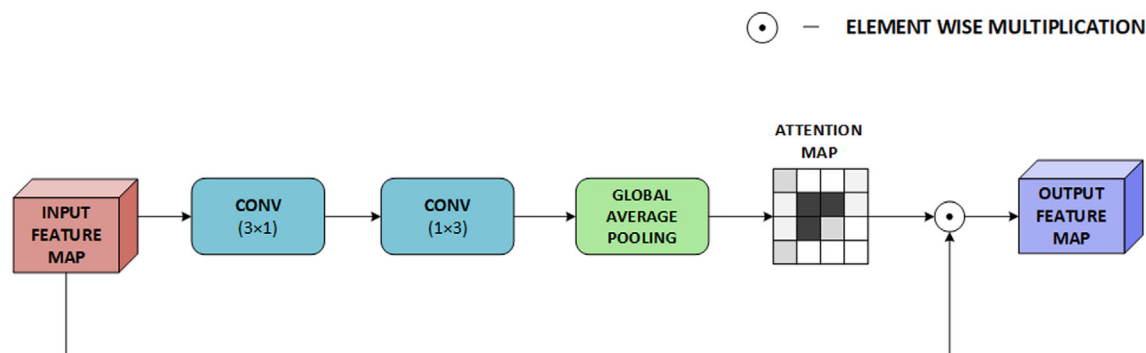


FIGURE 5 Schematic sketch of the enhanced channel attention (ECA) block.

multiplicative factor  $\gamma$  gamma is added to the loss function to reduce the loss of easily-classified examples to prevent these from dominating. The multiplicative factor  $\gamma$

gamma will rescale the modulating factor  $(1 - p_i)^\gamma$  such that the easy examples are down-weighted more than the hard ones.

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log p_t \quad (3)$$

$$p_t = \begin{cases} p & , \text{ if } y=1 \\ 1-p & , \text{ otherwise} \end{cases} \quad (4)$$

where  $(1-p_t)^\gamma$  is the modulating factor,  $\gamma$  is the focusing parameter,  $\alpha_t$  are the class weights and  $p$  is the class probability. Class imbalance was still present even after tuning the alpha and gamma parameters in FL leading to suboptimal performance. To overcome this, we added OHEM loss to improve precision on hard classes. The OHEM loss selects and optimizes negatively classified samples with greater loss values in a batch-wise manner. Given a batch size  $B$ , regular forward propagation is performed and per-instance losses are computed. Then, it finds  $M < B$  hard examples in the batch with high loss values and back-propagates the loss computer over the selected instances. The relation of the proposed loss function is formulated in Equation (5).

$$F_{\text{total}} = FL(p_t) + F_{\text{OHEM}} \quad (5)$$

where  $FL(p_t)$  is focal loss,  $F_{\text{OHEM}}$  is the OHEM loss and  $F_{\text{total}}$  is the total loss to be optimized. Moreover, the cross-entropy loss is applied to the hard negative samples selected in the OHEM loss.

## 4 | EXPERIMENTATION AND RESULTS

An overview of the dataset, experimentation methods and model training are presented. The effectiveness of the proposed network is assessed in this section using ablation experiments.

### 4.1 | Dataset description

The proposed model of this research work was developed using TCGA PRAD dataset sourced from the National

Cancer Institute's GDC data portal. The five main classes include GS 6, GS 7, GS 8, GS 9, GS 10. The distribution of samples in each class of the TCGA-PRAD dataset is presented in Table 1.

A sample is assigned a GS value originally by a pathologist after visual analysis. The process starts by assigning a primary and secondary Gleason grade to the sample after which it is assigned a GS based on the characteristics of the grade. If more than 50% of the sample being observed contains a particular Gleason grade it is assigned as the primary Gleason grade of the sample. When more than 5% and less than 50% of the sample is covered in another pattern of the sample then that is taken to be as the secondary Gleason grade of the sample. Both the primary and secondary values are added up to form the Gleason scoring system.

### 4.2 | Data pre-processing and augmentation

The techniques used for data pre-processing and data augmentation are discussed in this subsection. The dataset that we have taken consists of a total of 286 WSIs each having varying dimensions and split into patches of  $300 \times 300$  pixels accounting for a total of 17 683 patches after processing. Before training, the images are resized to dimensions of  $128 \times 128$  for effective feature extraction. Furthermore, the dataset is split into training, validation, and test sets in the ratio of 80:10:10.

As the dataset contains a majority of the samples belonging to GS 7, class imbalance was observed as reported in Figure 6. This may result in the inferior performance of the proposed network. To overcome class imbalance, it is necessary to perform data augmentation. The following augmentations were carried out using the Torchvision library: (1) Randomly flipping the images horizontally, (2) Randomly rotating the images by  $90^\circ$ , (3) Normalizing using the mean and standard deviation values of the dataset and (4) Resizing the patches into  $128 \times 128$  pixels.

### 4.3 | Patch extraction

Individual directories were assigned for each of the classes (6–10) and the corresponding images were imported into the respective directories. Using the cvtColor function of the cv2 library, the images were converted to the format of BGR2RGB. The images were then initialized as a numpy array and the patchify function was used to split each image into patches of  $300 \times 300$  pixels in the RGB colorspace. The resulting patches were then written into dedicated folders for each class. Once the patch creation

**TABLE 1** Number of samples in each class of the TCGA-PRAD dataset.

Class	Number of WSIs
Gleason score 6	24
Gleason score 7	160
Gleason score 8	38
Gleason score 9	62
Gleason score 10	2

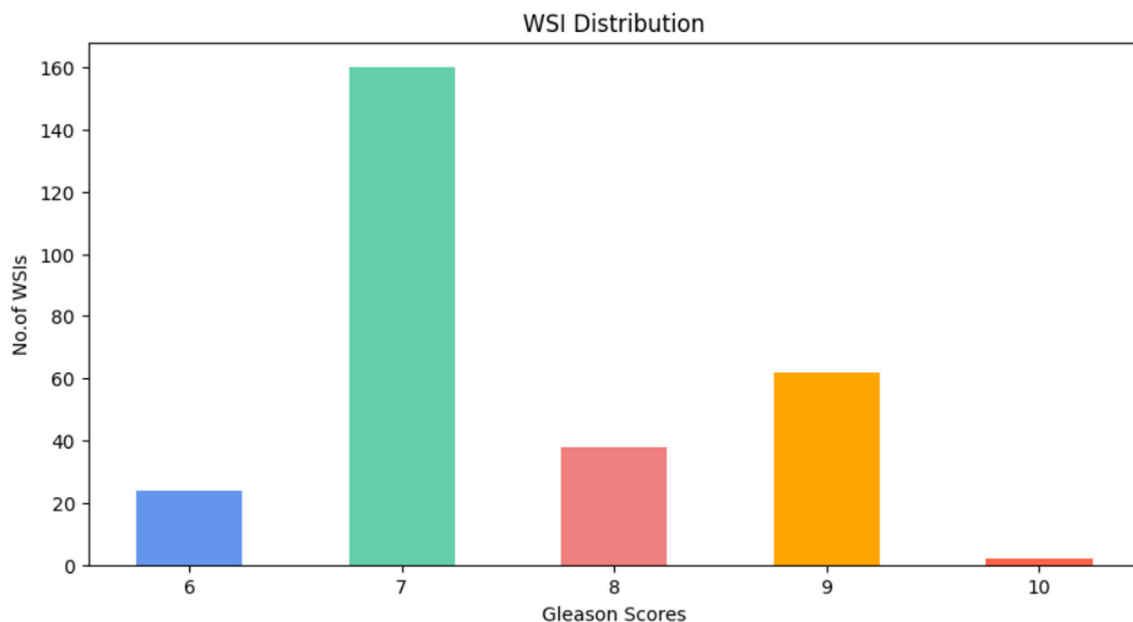


FIGURE 6 Data distribution of the dataset among the five classes.

process was complete, the mean pixel array value of the patches was determined and any images with a value of greater than 240 were eliminated. Finally, during model training, online augmentation was performed where the images were resized to  $128 \times 128$  dimension as it resulted in better performance.

#### 4.4 | Environmental setup

The proposed network was trained on a 24 GB NVIDIA A10G Tensor Core GPU. The DL model was implemented using Pytorch on an AWS EC2 g5.2xlarge instance. The system specifications are Ubuntu 20.0 with 8 AMD vCPUs, and 32 GB RAM. The model was trained with Adam optimizer, with a learning rate of  $1e-5$  and weight decay of  $1e-5$ , chosen after hyperparameter tuning. The training data was sampled in batches of 16 with an imbalanced dataset sampler due to the varied number of samples available in the dataset. Additionally, AMP CUDA library was used for mixed precision training and gradient scaling during backpropagation to prevent the gradients from becoming zero.

#### 4.5 | Hyperparameter tuning

Hyperparameter Tuning was performed using the Ray-Tune framework. The experiment was setup with five hyperparameters in the model (1) alpha (2) gamma (3) weight decay of the optimizer (4) learning rate of the

optimizer (5) batch size. Optimal tuning was attained by iterating through the search space of parameter values in the specified range: alpha value was set between 0 and 1; gamma value was set between 0 and 3; weight decay was one of the following: 0,  $1e-3$ ,  $1e-4$ ,  $1e-5$ ; learning rate of the optimizer was set between  $1e-1$  and  $1e-5$ ; batch size was either 16 or 32. The proposed network resulted in optimal convergence with a learning rate of  $1e-5$ , decay rate of  $1e-5$ , and batch size of 16.

#### 4.6 | Ablation studies

An analysis of the ablation experiments carried out for the proposed architecture is presented in this subsection. The CNN was incrementally trained with various enhancements to validate their effectiveness. The degree of performance improvement is measured by the difference in mean accuracy, precision, F1 score, recall and Kappa score.

##### 4.6.1 | Analysis of a 7-layer CNN

This subsection analyses the performance of a baseline 7-layer convolutional neural network. The model was trained and validated for 50 epochs on the TCGA-PRAD dataset where each epoch took around 40 s to complete. The resultant observations are presented in Figure 7. An accuracy of 84% was obtained on the testing set with the baseline CNN with 33 M trainable parameters. The final



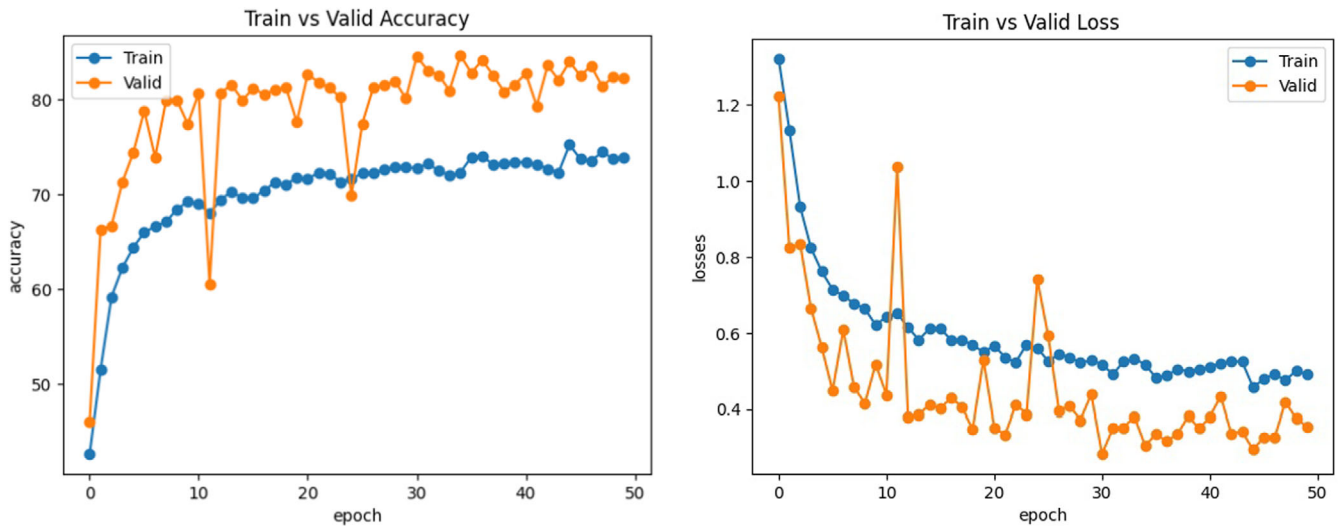


FIGURE 7 Analysis of a baseline 7-layer CNN (A) accuracy (B) loss.

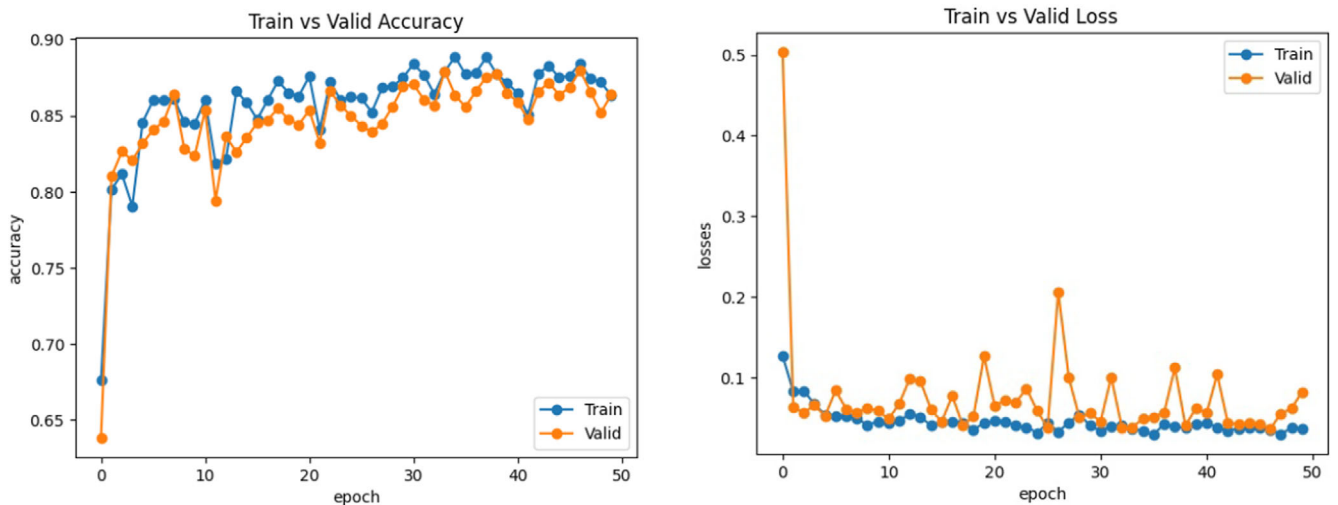


FIGURE 8 Analysis of the proposed network without self-attention layers (A) accuracy (B) loss.

Kappa score obtained by the baseline CNN on the TCGA-PRAD dataset was 84.23%.

#### 4.6.2 | Analysis of the proposed network without self-attention layers

With the inclusion of the self-attention layers to extract important contextual and positional information, the model's performance improves. The proposed network was trained and tested on the TCGA-PRAD dataset without the self-attention layer and the results are illustrated in Figure 8. This addition influences the way attention parameters are tuned. Without a drastic increase in the number of parameters, the attention feature maps are calculated thereby improving the feature representation

power. As a result, the self-attention layer improves the accuracy of the proposed network by 1.01%. More importantly, the precision score has been increased by 1.3% and the F1 score by a factor of 1%. The final accuracy and Kappa score of the proposed network without self-attention are 88.04% and 84.5%.

#### 4.6.3 | Analysis of the proposed network

The proposed network was trained on the TCGA-PRAD dataset for 50 epochs where each epoch took nearly 3 min to complete. The HMAR block is a vital part of the network for extracting salient features from different levels to improve performance. The local residual path with multiheaded self-attention extracts contextual and

positional information. The feature extraction sub-blocks in the HMAR block consist of spatial and channel attention layers to improve the network's feature representation power and overall performance. The observations of the model training are illustrated in Figure 9. The average precision, F1 score, recall, and Cohen's Kappa are 82.4%, 80.4%, 83.6%, and 86%. The total number of trainable parameters in the proposed network is 18 M.

The overall summary of the networks trained and tested on the TCGA-PRAD dataset is tabulated in Table 2. This experimentation is necessary to demonstrate the

performance of the network with and without certain enhancements.

Moreover, the proposed network was trained using cross-entropy loss and FL to validate the efficacy of the proposed loss function. The results are tabulated in Table 3. As it can be inferred from the table, the proposed loss function outperforms the others in all the metrics.

In order to assess the robustness of the proposed network across different samples, five-fold cross-validation was performed and the results of evaluating the proposed network on five-fold cross-validation are presented in Table 4.

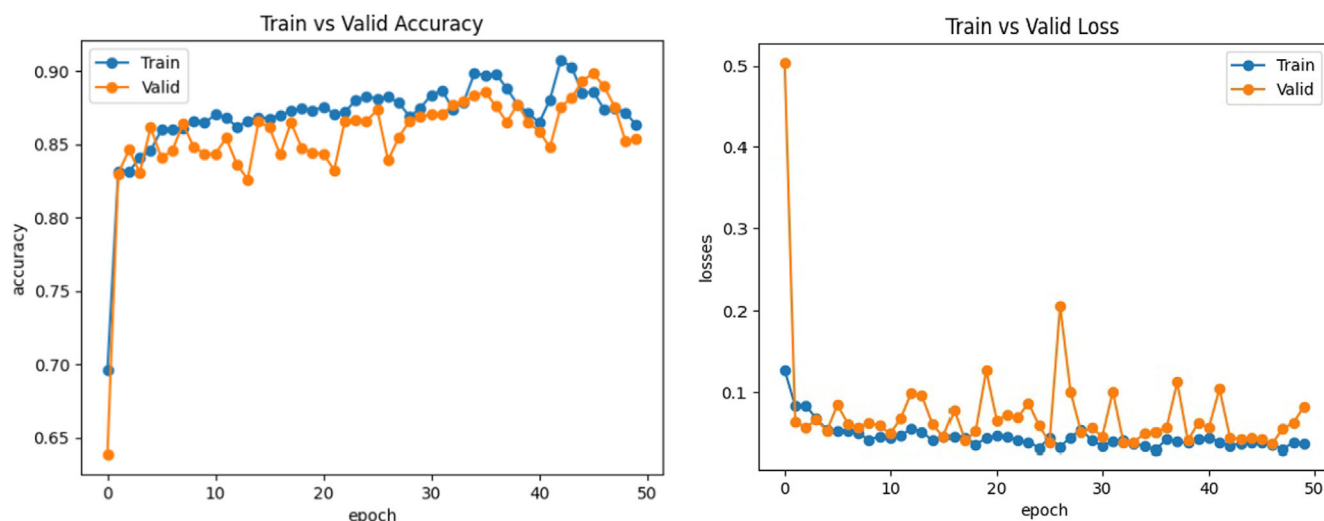


FIGURE 9 Analysis of the proposed network (A) accuracy (B) loss.

TABLE 2 Overall summary of the experiments performed for the ablation studies on the TCGA-PRAD dataset.

Experiments	Balanced accuracy in (%)	Precision in (%)	Recall in (%)	F1-score in (%)
Baseline 7-layer feedforward CNN	84.23	76.8	76.4	75.4
Proposed network without self-attention	88.04	81.0	82.8	79.4
Proposed network with self-attention and w/o Mixed FocalOHEM loss	87.15	79.6	80.1	79.1
Proposed network with self-attention and Mixed FocalOHEM loss	89.19	82.4	83.6	80.4

TABLE 3 Analysis of the proposed network with the Mixed FocalOHEM loss with other loss functions.

Experiments	Balanced accuracy in (%)	Precision in (%)	Recall in (%)	F1-score in (%)
Proposed network with Cross entropy loss	89.10	80.9	83.4	78.22
Proposed network with Focal loss	89.05	81.3	83.4	78.9
Proposed network with Mixed FocalOHEM loss	89.19	82.4	83.6	80.4

**TABLE 4** Fold-wise computation of the proposed model for accuracy, precision, recall, and F1-score on the validation set.

Folds	Balanced accuracy in (%)	Precision in (%)	Recall in (%)	F1-score in (%)
Fold 1	91.40	86.3	88.4	81.7
Fold 2	89.65	82.7	85.2	80.4
Fold 3	90.42	84.1	86.7	81.2
Fold 4	89.26	83.1	84.5	80.5
Fold 5	87.47	81.6	83.1	78.9
Average	89.64	83.5	85.5	80.54
Standard deviation	1.462	1.774	2.041	1.059

**TABLE 5** Quantitative performance comparison of the proposed network with the state-of-the-art architectures for Gleason scoring of prostate cancer.

Methods	Number of parameters	Balanced accuracy in (%)	F1-score in (%)	Precision in (%)	Recall in (%)
AlexNet w/o FocalOHEM loss	61M	85.25	76.86	77.32	77.14
AlexNet with FocalOHEM loss	61M	85.14	76.52	77.21	77.20
ResNet-50 w/o FocalOHEM loss	25M	86.82	74.93	77.98	78.43
ResNet-50 with FocalOHEM loss	25M	86.97	74.97	78.0	78.56
DenseNet-201 w/o FocalOHEM loss	20M	87.44	76.27	78.72	78.93
DenseNet-201 with FocalOHEM loss	20M	87.90	77.74	79.96	79.15
EfficientNet-B0 w/o FocalOHEM loss	5.2M	87.56	72.96	77.78	79.34
EfficientNet-B0 with FocalOHEM loss	5.2M	88.94	73.45	78.21	79.78
VGG-16 w/o FocalOHEM loss	138M	87.89	79.29	79.37	79.47
VGG-16 with FocalOHEM loss	138M	87.7	79.13	79.10	79.40
EfficientNet-B4 w/o FocalOHEM loss	19M	88.12	78.92	79.24	79.22
EfficientNet-B4 with FocalOHEM loss	19M	88.39	79.22	79.51	79.46
Proposed network	18M	89.19	80.4	84.8	79.42

## 5 | DISCUSSION

In this section, a comparison of the proposed work with the existing state-of-the-art architectures and research works is presented. For a fair comparison, all the models were re-implemented and trained on the TCGA-PRAD dataset with the same data distribution. The results were demonstrated on a common test set.

### 5.1 | Comparison with the state-of-the-art networks

Table 5 presents a comparison of the proposed work with the state-of-the-art CNN architectures. The pre-trained

CNN architectures were fine-tuned to adapt to the TCGA-PRAD dataset. The proposed network has shown improvement in performance in all the metrics over the pre-trained architectures. Additionally, the proposed network converges quickly making it an effective method for the scoring of PCa. The networks have also been trained and tested with and without the proposed Mixed FocalOHEM loss to validate its efficacy. Of all the compared architectures, EfficientNet-B4 performed the best followed by VGG-16 in terms of accuracy. However, the VGG network resulted in greater precision, recall, and F1 score than every other state-of-the-art network.

The proposed network outperformed the other networks with a decent margin having comparatively lower training parameters than most other pre-trained architectures.

**TABLE 6** Quantitative performance comparison of the proposed network with the existing works for Gleason scoring of prostate cancer.

Source	Method	Model	Accuracy in (%)
Toro et al. <sup>10</sup>	Transfer learning	GoogLeNet	73.52
Lara et al. <sup>15</sup>	Transfer learning	Multimodal latent semantic alignment, GoogLeNet, AlexNet	77.01
Xu et al. <sup>18</sup>	Custom CNN based on VGG	Custom-VGG	77.12
Kallen et al. <sup>19</sup>	Transfer Learning	Pre-trained CNN, Random Forest, SVM	81.1
Proposed model	Custom CNN with HMAR blocks and ECA module	Hierarchical Multi-Attention Residual Network	89.2

Therefore, a good tradeoff between performance and computational complexity is achieved.

## 5.2 | Performance analysis with the existing works

The performance of the proposed method is compared against the existing works for PCa scoring and the observations are presented in Table 6. We have reported the works that have experimented with the same TCGA-PRAD dataset for a fair comparison. The compared works have applied a variety of ML and CNN techniques for Gleason scoring.

The existing research works have employed both transfer learning and custom CNN for Gleason scoring of PCa histopathology images. Even though transfer learning is a good strategy as it overcomes overfitting and reduces training time, negative transfer may occur. This is due to different initial and target domains. The proposed network comparatively contains less trainable parameters than the majority of the state-of-the-art architectures. Furthermore, the extraction of hybrid features based on context, channel, and positional information is necessary for improved performance and generalization. The classification accuracies obtained by the existing works range from 73.52% to 81.1%. As it can be inferred, the proposed network has shown improved results outperforming the existing state-of-the-art approaches with an overall accuracy of 89.2% and a Kappa score of 86%.

## 6 | CONCLUSION

This research presents a novel hierarchical attention-based residual feature fusion network for Gleason scoring of PCa from histopathology images. Furthermore, a channel attention module and a loss function are also proposed for improved network performance and generalization capability. The existing works have not focused on extracting hierarchical features from multiple

levels for improved precision. Salient attention-guided features are captured with the HMAR blocks present in the network and are finally optimized with the Mixed FocalOHEM loss function. Channel, spatial, and self-attention mechanisms are applied to these hierarchical features gradually for improving the representation power of the network. The residual path in the HMAR block also aids the network by applying self-attention on long-range dependencies captured across the network. The proposed ECA block focuses on extracting channel-wise features from different receptive fields with a low computational cost. Finally, we have proposed Mixed FocalOHEM loss to give importance to optimizing hard negative examples from the dataset. Our proposed method exhibits optimal results in identifying the GS of PCa from histopathology images with an accuracy of 89.19% and a Kappa score of 86%. This framework can be extended for the diagnosis of tumors and other types of cancer. Furthermore, this network can be modified to perform segmentation and detection tasks.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

R. Karthik  <https://orcid.org/0000-0002-5250-4337>

R. Menaka  <https://orcid.org/0000-0002-8652-191X>

## REFERENCES

1. Rawla P. Epidemiology of prostate cancer. *World J Oncol*. 2019; 10(2):63-89. doi:[10.14740/wjon1191](https://doi.org/10.14740/wjon1191)
2. Pienta KJ. Risk factors for prostate cancer. *Ann Int Med*. 1993; 118(10):793. doi:[10.7326/0003-4819-118-10-199305150-00007](https://doi.org/10.7326/0003-4819-118-10-199305150-00007)
3. Britannica T, Editors of Encyclopaedia. *Prostate Gland*. Encyclopaedia Britannica; 2020. <https://www.britannica.com/science/prostate-gland>

4. Key statistics for prostate cancer. <https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html>.
5. Mazhar D. Prostate cancer. *Postgraduate Med J*. 2002;924:590-595. doi:10.1136/pmj.78.924.590
6. Goldenberg S, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol*. 2019;16:391-403. doi:10.1038/s41585-019-0193-3
7. Linkon AHM, Labib MM, Hasan T, Hossain M, Jannat M-E. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: an extensive study. *Inform Med Unlocked*. 2021;24:100582. doi:10.1016/j.imu.2021.100582
8. Wang D, Foran DJ, Ren J, Zhong H, Kim IY, Qi X. Exploring automatic prostate histopathology image Gleason grading via local structure modeling. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2015. doi:10.1109/embc.2015.7318936
9. Zhang Y, Zhang J, Song Y, Shen C, Yang G. Gleason score prediction using deep learning in tissue microarray image. 2020 ArXiv, abs/2005.04886.
10. Jiménez del Toro O, Atzori M, Otálora S, et al. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. In: Gurcan MN, Tomaszewski JE, eds. *SPIE Proceedings*. SPIE; 2017. doi:10.1117/12.2255710
11. Zhang J, Ma K, van Arnam J, Gupta R, Saltz J, Vakalopoulou M, Samaras D. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2021:3776-3784.
12. Alkhateeb A, Atikukke G, Rueda L. Machine learning methods for prostate cancer diagnosis. *J Cancer*. 2020;1(3):70-75.
13. Otálora S, Atzori M, Khan A, Jimenez-del-Toro O, Andrearczyk V, Müller H. A systematic comparison of deep learning strategies for weakly supervised Gleason grading. In: Tomaszewski JE, Ward AD, eds. *Medical Imaging 2020: Digital Pathology*. SPIE; 2020. doi:10.1117/12.2548571
14. Otálora S, Marini N, Müller H, Atzori M. Semi-weakly supervised learning for prostate cancer image classification with teacher-student deep convolutional networks. *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer International Publishing; 2020:193-203. doi:10.1007/978-3-030-61166-8\_21
15. Lara JS, Contreras OVH, Otálora S, Müller H, González FA. Multimodal latent semantic alignment for automated prostate tissue classification and retrieval. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2020*. Springer International Publishing; 2020:572-581. doi:10.1007/978-3-030-59722-1\_55
16. Wang T-H, Lee C-Y, Lee T-Y, Huang H-D, Hsu JB-K, Chang T-H. Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. *Cancers*. 2021;13(11):2528. doi:10.3390/cancers13112528
17. Otálora S, Marini N, Müller H, Atzori M. Combining weakly and strongly supervised learning improves strong supervision in Gleason pattern classification. *BMC Med Imaging*. 2021; 21(1):77. doi:10.1186/s12880-021-00609-0
18. Xu H, Park S, Hwang TH. Computerized classification of prostate cancer Gleason scores from whole slide images. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17(6):1871-1882. doi:10.1109/tcbb.2019.2941195
19. Kallen H, Molin J, Heyden A, Lundstrom C, Astrom K. Towards grading Gleason score using generically trained deep convolutional neural networks. *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2016. doi:10.1109/isbi.2016.7493473
20. Otálora S, Atzori M, Andrearczyk V, Müller H. Image magnification regression using DenseNet for exploiting histopathology open access content. *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer International Publishing; 2018:148-155. doi:10.1007/978-3-030-00949-6\_18
21. Abbasi AA, Hussain L, Awan IA, et al. Detecting prostate cancer using deep learning convolution neural network with the transfer learning approach. *Cogn Neurodyn*. 2020;14:523-533. doi:10.1007/s11571-020-09587-5
22. Wildeboer RR, van Sloun RJG, Wijkstra H, Mischi M. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. *Comput Methods Prog Biomed*. 2020;189:105316. doi:10.1016/j.cmpb.2020.105316
23. Tolkach Y, Dohmgorgen T, Toma M, Kristiansen G. High-accuracy prostate cancer pathology using deep learning. *Nat Mach Intell*. 2020;2(7):411-418. doi:10.1038/s42256-020-0200-7
24. Arvaniti E, Fricker KS, Moret M, et al. Author correction: automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep*. 2021;11(1):23032. doi:10.1038/s41598-021-02195-1
25. Ström P, Kartasalo K, Olsson H, et al. Pathologist-level grading of prostate biopsies with artificial intelligence (version 1). *arXiv*. 2019. doi:10.48550/ARXIV.1907.01368
26. Șerbanescu M-S, Oancea C-N, Streba CT, et al. Agreement of two pre-trained deep-learning neural networks built with transfer learning with six pathologists on 6000 patches of prostate cancer from Gleason2019 challenge. *Rom J Morphol Embryol*. 2020;61(2):513-519. doi:10.47162/rjme.61.2.21
27. Chakraborty S, Ma K, Gupta R, et al. Visual attention analysis of pathologists examining whole slide images of prostate cancer. *ArXiv*. 2022. doi:10.48550/ARXIV.2202.08437
28. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Dig Med*. 2019;2(1):48. doi:10.1038/s41746-019-0112-2
29. Brunese L, Mercaldo F, Reginelli A, Santone A. Radiomics for Gleason score detection through deep learning. *Sensors*. 2020; 20(18):5411. doi:10.3390/s20185411
30. Shin H-K, Hong S-H, Choi Y-J, Shin Y-G, Park S, Ko S-J. Self-attentive normalization for automated Gleason grading system. *2020 IEEE REGION 10 CONFERENCE (TENCON)*. IEEE; 2020. doi:10.1109/tencon50793.2020.9293775
31. Reda I, Khalil A, Elmogy M, et al. Deep learning role in early diagnosis of prostate cancer. *Technol Cancer Res Treat*. 2018;17: 153303461877553. doi:10.1177/1533034618775530
32. Mohsin M, Shaikat A, Akram U, Zarrar MK. Automatic prostate cancer grading using deep architectures. *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*. IEEE; 2021. doi:10.1109/aiccsa53542.2021.9686869
33. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233-241. doi:10.1016/s1470-2045(19)30739-9



34. Li Y, Huang M, Zhang Y, et al. Automated Gleason grading and Gleason pattern Region segmentation based on deep learning for pathological images of prostate cancer. *IEEE Access*. 2020;8:117714-117725. doi:[10.1109/access.2020.3005180](https://doi.org/10.1109/access.2020.3005180)
35. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-alone self-attention in vision models (version 1). *arXiv*. 2019. doi:[10.48550/ARXIV.1906.05909](https://doi.org/10.48550/ARXIV.1906.05909)
36. Karthik R, Menaka R, Siddharth MV. Classification of breast cancer from histopathology images using an ensemble of deep multiscale networks. *Biocybern Biomed Eng*. 2022;42(3):963-976. doi:[10.1016/j.bbe.2022.07.006](https://doi.org/10.1016/j.bbe.2022.07.006)
37. Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(2):318-327. doi:[10.1109/tpami.2018.2858826](https://doi.org/10.1109/tpami.2018.2858826)
38. Shrivastava A, Gupta A, Girshick R. Training Region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. doi:[10.1109/cvpr.2016.89](https://doi.org/10.1109/cvpr.2016.89)
39. Zhang Y-D, Satapathy SC, Wu D, Guttery DS, Górriz JM, Wang S-H. Improving ductal carcinoma in situ classification by convolutional neural network with exponential linear unit and rank-based weighted pooling. *Complex Intell Syst*. 2020;7(3):1295-1310. doi:[10.1007/s40747-020-00218-4](https://doi.org/10.1007/s40747-020-00218-4)
40. Zhang Y-D, Govindaraj VV, Tang C, Zhu W, Sun J. High performance multiple sclerosis classification by data augmentation and AlexNet transfer learning model. *J Med Imaging Health Inform*. 2019;9(9):2012-2021. doi:[10.1166/jmihi.2019.2692](https://doi.org/10.1166/jmihi.2019.2692)
41. Zhang Y-D, Dong Z, Chen X, et al. Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimed Tools Appl*. 2017;78(3):3613-3632. doi:[10.1007/s11042-017-5243-3](https://doi.org/10.1007/s11042-017-5243-3)

**How to cite this article:** Karthik R, Menaka R, Siddharth MV, Hussain S, Siddharth P, Won D. HMARNET—A Hierarchical Multi-Attention Residual Network for Gleason scoring of prostate cancer. *Int J Imaging Syst Technol*. 2024;34(1):e22976. doi:[10.1002/ima.22976](https://doi.org/10.1002/ima.22976)