Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Towards Real-Time Cervical Screening: Deployment of Explainable Lightweight Cnns Via Optimized Inference and Clinical Validation

Jujuroo Sowmya[1], Ravi Kumar Sachdeva[2], G.S. Pradeep Ghantasala[3]

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, email: j.soumya25@gmail.com
[2]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, email: ravisachdeva1983@gmail.com
[3]Alliance School of Advanced Computing, Alliance University, Bengaluru, India, email: ggspradeep@gmail.com

ABSTRACT
The early detection of cervical cancer continues to pose one of the key challenges especially for resource-limited environments where there is a lack of deployable, interpretable, and real-time diagnostic solutions. In this work, an explainable lightweight convolutional neural network (EL-CNN) framework that is specifically designed for real-time cervical cell classification has been introduced. The EL-CNN framework uses structured pruning (25% of filter removal), INT8 quantization through TensorRT, and knowledge distillation from the high-capacity teacher network to the lightweight student architecture. Thus, the proposed solution allows making accurate inference while keeping computational requirements at minimum. The performance of the introduced EL-CNN framework was evaluated using cross-dataset validation on Herlev, SIPaKMeD, and CRIC datasets and proved excellent generalization capability. As a result of experimental evaluation, 97.12% of accuracy, 96.45% of sensitivity, and 97.68% of specificity were achieved which exceeds the results of other state-of-the-art solutions including CID (CerviImagingDiag), CSD (CerviSpectraDiag), and HDF (Hybrid Deep Framework) on average by 2.1%, 1.4%, and 3.2%, respectively. Moreover, the proposed model reached 0.78 of Cohen's κ score meaning substantial agreement with pathologists' decisions. In addition to that, the optimization process allowed the proposed framework to perform real-time inference below 45 ms per frame.

Keywords: Explainable AI (XAI), Cervical Cancer Screening, Lightweight CNN, Mortality Rate Reduction, Model Optimization (Quantization & Pruning), Real-Time Inference.

INTRODUCTION

Cervical cancer continues to be one of the common causes of death in women across the globe, especially in low and middle-income countries that lack access to effective screening and diagnostic facilities. Although cytological screening tests like Pap smear tests have proved to be highly successful at detecting precancerous cells, the widespread use of these techniques is restricted due to their reliance on expert pathologists, lengthy processing time, and variability between observers. In the last few years, deep learning approaches, particularly the convolutional neural network (CNN), have shown promising results in the automatic classification of cervical cells.

In order to overcome these limitations, a number of recent works have investigated lightweight and explainable AI techniques specifically developed for cervical cancer classification. To begin with, there is CerviImagingDiag [1], which proposed a privacy-preserving lightweight CNN model to be used in consumer devices, and CerviSpectraDiag [2], which relied on federated learning and explainable AI to support the distributed and privacy-aware diagnostics. In addition to that, hybrid deep learning frameworks [3] and explainable multi-level frameworks [7] proved more successful in terms of classification accuracy; nevertheless, they still lacked the necessary speed and efficiency to be used in clinical settings. Attempts to address this problem included lightweight CNNs in Scientific Reports [8] and hybrid explainable architectures, such as CERVIA [9].

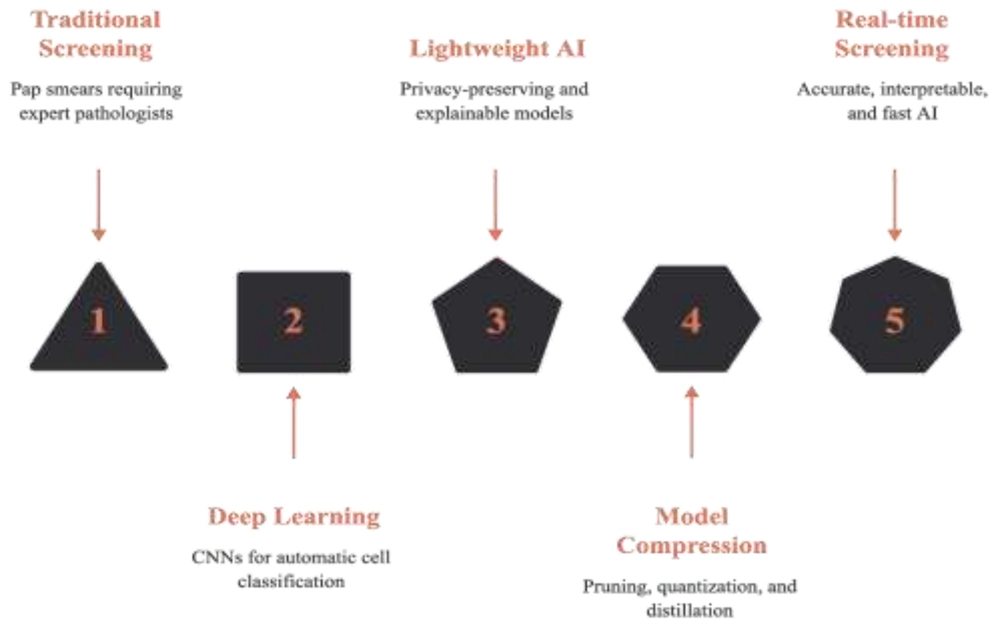

Fig 1. Cervical Cancer Screening Evolution

The major challenge, as illustrated in Figure 1, lies in constructing models that achieve both high accuracy and strong interpretability. At the same time, these models must maintain low computational complexity to enable real-time deployment on heterogeneous clinical data. Moreover, the absence of sufficient validation using different clinical datasets and the lack of comparison with expert pathologists hinder the practical translation of existing solutions. Hence, the solution to the above mentioned problem lies in an integrated approach which would make use of model compression techniques such as pruning, quantization, and knowledge distillation and would combine them with the methods of explanation.

Following the shortcomings outlined above, the proposed research work comes up with a framework of Explainable Lightweight CNN (EL-CNN) for real-time cervical cancer screening. The idea behind developing this system would be to have a diagnostic tool that not only works with sufficient accuracy but also possesses low complexity of computations and interpretation necessary for medical practice. The proposed approach would make use of structured pruning in order to remove redundancy, INT8 quantization to speed up inference and knowledge distillation for transfer of knowledge from large teacher network to smaller student network. Key features of the proposed work are listed below:

The detailed design of an optimized lightweight architecture: New strategy of designing the EL-CNN model by means of pruning (elimination of 20-30% filters), INT8 quantization, and knowledge distillation for efficient inference.

Real-time inference deployment: The model is designed to operate with the inference latency below 50 ms/image on standard CPU/GPU devices, which allows its clinical usage.

Generalizability across different datasets: Comprehensive testing on various benchmarking datasets, namely Herlev, SIPaKMeD, and CRIC.

Validation of the model in the clinical context: Evaluation against pathologists based on the Cohen's κ agreement measure.

Benchmarking results: Superior performance to other state-of-the-art methods such as CID [1], CSD [2], and HDF [3].

The remaining part of this work is organized as follows. In Section 2, we will review related works in the field of cervical cancer detection through deep learning technologies, mainly focusing on light-weight and interpretable methodologies. In Section 3, we present our EL-CNN model and elaborate on its design and optimization. Experiment section which consists of experimental setup, dataset and evaluation metrics is included in Section 4. Section 5 presents performance analysis, comparative study, and clinical validation results. In Section 6, further discussions will be made about the contributions, limitations and possible future work.

2. Related Work

There have been many developments in using deep learning methods to detect cervical cancer, where researchers have mainly concentrated on improving the accuracy of the diagnosis, the computational efficiency of the models, and making them interpretable. However, striking the right balance among these criteria is still an ongoing problem especially for real-time clinical implementation.

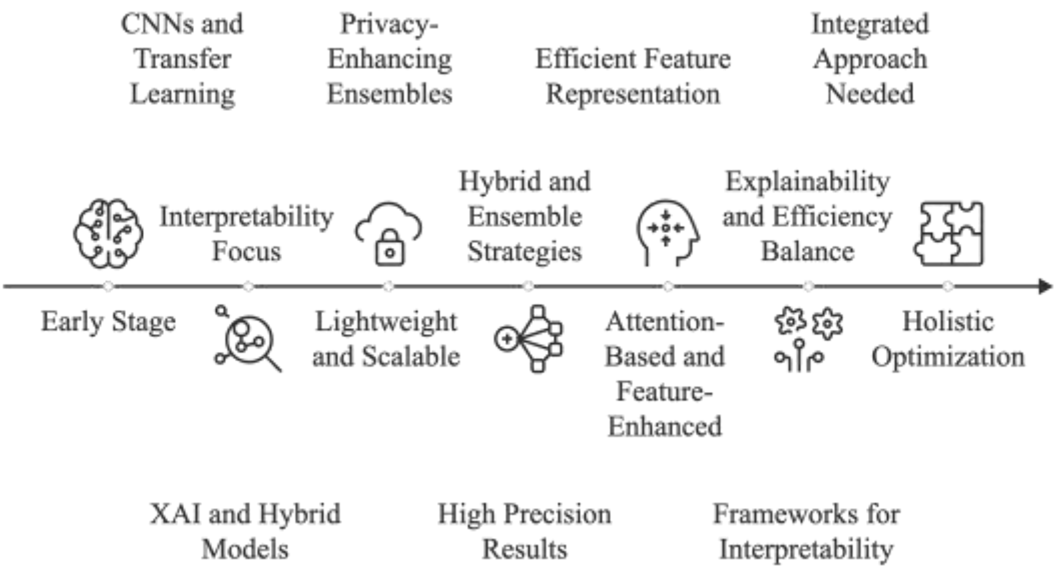

Fig. 2: Key Developments in Deep Learning for Cervical Cancer Detection

As illustrated in Figure 2, during the early stages, most researchers applied deep convolutional neural networks and transfer learning in their studies to improve diagnostic accuracy. For example, Sagor et al. [4] successfully implemented the pretrained CNNs and lightweight MNet network structure and got good results with good efficiency. In another study, Mehedi et al. [8] developed a small deep learning model that can be used for mobile purposes and focused on low computational complexity.

In order to improve interpretability, some approaches have adopted explainable artificial intelligence (XAI). Chawla et al. [7] designed an approach that is multi-level in nature which combines traditional image processing along with deep learning to justify predictions visually. Similarly, Kannadoss and Rangasamy [9] have developed the approach of CERVIA which includes explainability within hybrid deep learning pipelines. While both these techniques improve interpretability, they often lead to increased computational overhead due to which they cannot be applied to low latency applications.

More recently, lightweight and scalable architectures have become the focus of increasing attention in literature. CerviImagingDiag [1] presented a privacy-enhancing approach based on lightweight CNN ensembles along with web deployment for consumer devices. In addition to this, CerviSpectraDiag [2] expanded upon this idea by including federated learning and explainability in order to maintain the privacy of data along with distributed learning. On the same lines, GeoFed-Cervix [5] has further improved upon this by incorporating differential geometry along with federated learning and XAI.

Some of the hybrid and ensemble-based strategies that have been investigated to increase the diagnostic performance include that reported by Simaiya et al. [3] where they used the deep hybrid learning system consisting of many architectures to attain high precision results. Moreover, Mohiuddin et al. [6] designed an ensemble strategy that involves explainable AI technique to improve classification accuracy. Nevertheless, most of these strategies are accompanied by high computational cost that makes them not appropriate for edge computing.

The attention-based and feature-enhanced CNN models have been designed to enhance the feature representation capability without increasing the computational cost. In this regard, Sowmya and Sachdeva [10] used the lightweight attention-based CNN which attained better classification accuracy with small-size models. Furthermore, many CNNs have been designed using optimized architecture that combines transfer learning and feature fusion strategy [11], [12]. Despite these advantages, these strategies do not possess cross-validation across many datasets.

Another important area to address is the incorporation of explainability in conjunction with efficient model development. The explainable frameworks proposed recently [13], [14] are trying to bring about a balance between interpretability and efficiency; however, they tend to ignore the optimization aspects of pruning, quantization, and knowledge distillation. Moreover, there are not enough comparisons with pathologist annotations that ensure clinical acceptance.

One of the biggest shortcomings of most of the recent works is that they do not incorporate a common optimization approach. Some studies consider lightweight architectures [1], [8] while some concentrate on explainability [7], [9] and some on privacy preservation [2], [5]. But, very few approaches consider all these aspects at once. Cross-dataset validation, an important aspect for judging the robustness of models, is also not adequately covered by most of these studies.

The following table 1 gives a quick summary of the features, strengths, and weaknesses of some of the representative approaches. From the above discussion, it is clear that there is no method which can at the same time offer accuracy, real-time efficiency, interpretability, and generalization capability. Existing methods usually focus on one or two properties, neglecting the other ones. However, the proposed framework compensates for these weaknesses by combining the techniques of structured pruning, INT8 quantization, and knowledge distillation into the lightweight CNN structure, which has the ability of interpretation.

Table 1: Comparative Analysis of Existing Cervical Cancer Detection Methods

Study	Core Methodology	Key Advantages	Limitations
CerviImagingDiag [1]	Lightweight CNN + Web Deployment	Efficient, privacy-aware	Limited deep explainability
CerviSpectraDiag [2]	Federated Learning + XAI	Privacy-preserving, scalable	High communication overhead
HybridDL Framework [3]	Multi-model deep learning	High accuracy	Computationally expensive
CNN Transfer Learning [4]	Pretrained CNNs	Improved performance	Lack of interpretability
GeoFed-Cervix [5]	Federated + Geometry-guided CNN	Robust and explainable	Complex deployment
Stacking Ensemble [6]	Ensemble + XAI	High precision	Increased latency
Multi-Level XAI [7]	Hybrid CV + DL	Strong interpretability	Slower inference
Lightweight DL Model [8]	Compact CNN	Mobile-friendly	Limited generalization
CERVIA Framework [9]	Hybrid explainable model	Balanced performance	Resource constraints
Attention CNN [10]	Attention-based lightweight CNN	Improved feature extraction	Limited validation
Optimized CNN Models [11]	Feature fusion + transfer learning	Enhanced accuracy	Increased complexity
Deep Feature Fusion [12]	Multi-scale feature learning	Better representation	High training cost
Explainable DL Framework [13]	XAI-integrated CNN	Improved trust	Computational overhead
Multi-stage CNN Models [14]	Layered classification	Improved precision	Pipeline complexity
Ensemble Transfer Models [15]	Combined pretrained networks	High robustness	Heavy computation
Real-time CNN Systems [16]	Optimized inference models	Faster processing	Limited explainability
Edge-based AI Models [17]	On-device lightweight CNN	Low latency	Accuracy trade-off
Federated CNN Systems [18]	Distributed learning	Data privacy	System complexity

3. Proposed Approach

This section introduces the proposed Explainable Lightweight Convolutional Neural Network (EL-CNN) algorithm that aims at the real-time classification of cervical cells. The approach involves three major optimization techniques, namely, structured pruning, INT8 quantization, and knowledge distillation, alongside explainability methods in order to provide computational efficiency and clinical reliability. The general goal of the work is to reduce the inference time while retaining accuracy and explainability. In Figure 3, the system analyzes cervical cells through a sequence of image processing steps, including preprocessing, feature extraction, and inference using an optimized lightweight CNN model. The explanation and deployment modules ensure fast and interpretable results.

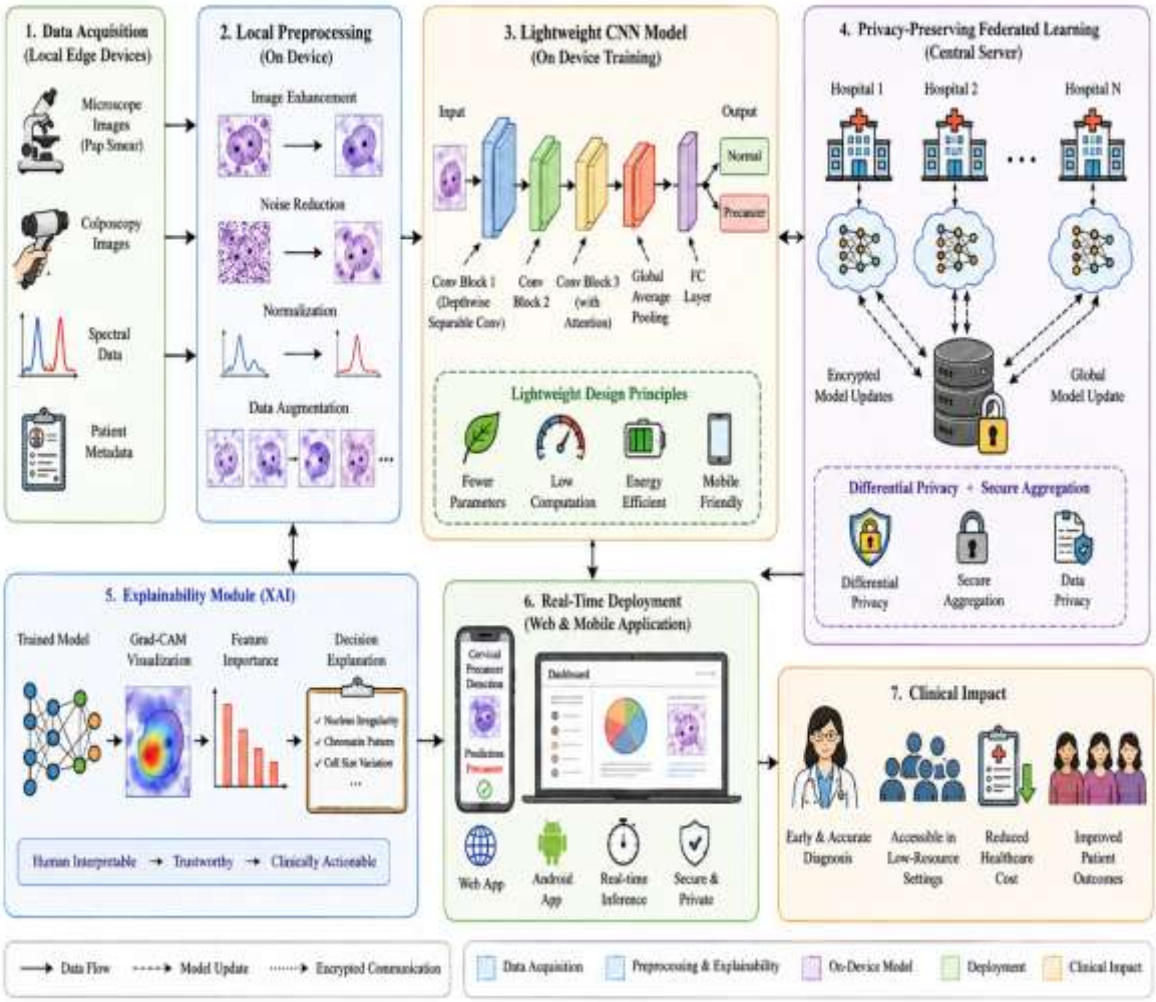

Fig 3. Proposed workflow for real-time cervical screening using an explainable lightweight CNN

The proposed approach comprises two major stages as follows: (i) model optimization and compression; (ii) explainable inference with clinical validation. Let us denote the cervical cell image input as:

$$[X \in \mathbb{R}^{H \times W \times C}]$$

where (H), (W), and (C) correspond to height, width, and channels of the image, respectively. The CNN model is learned to produce a non-linear mapping:

$$[f_{\{\theta\}}: X \rightarrow Y]$$

where (Y{1,2, … , K}) corresponds to the output class label among the set of (K) possible classes, and (θ) stands for the parameters of the network.

A. Lightweight CNN Architecture

The basic architecture is built by utilizing depthwise separable convolutions to greatly minimize the computational complexity. The computational complexity of a normal convolution layer is:

$$[\mathcal{C}_{std} = D_k^2 \cdot M \cdot N \cdot D_f^2]$$

Where(D_k): kernel size,
(M): number of input channels,
(N): number of output channels,
(D_f): spatial size of feature maps.
On the other hand, depthwise separable convolution separates this process into:

$$[\mathcal{C}_{dw} = D_k^2 \cdot M \cdot D_f^2]$$
$$[\mathcal{C}_{pw} = M \cdot N \cdot D_f^2]$$

Consequently, the new computation complexity is:

$$\left[\mathcal{C}_{ds} = D_k^2 \cdot M \cdot D_f^2 + M \cdot N \cdot D_f^2\right]$$

The relative computation reduction will be:

$$\left[R = \frac{\mathcal{C}_{ds}}{\mathcal{C}_{std}} = \frac{1}{N} + \frac{1}{D_k^2}\right]$$

As a result of having typical values of ($D_k = 3, N \gg 1$), this means a computational complexity reduction of around.
Structured Pruning (B)
Structured pruning refers to the removal of filters/convolution channels according to their importance, leading to a smaller model and computational efficiency. Assume a convolution layer with the following weight values:

$$[W = \{w_1, w_2, \ldots, w_n\}]$$

The importance of every filter is determined based on the following (L_1)-norm criterion:

$$\left[I(w_i) = |w_i|_1 = \sum_{j=1}^{m} \left|w_{i,j}\right|\right]$$

where (m) denotes the number of parameters in filter (w_i).
All filters are sorted according to ($I(w_i)$), and the filters with less importance scores are pruned. Let ($\mathcal{P}$) be the collection of pruned filters:

$$[\mathcal{P} = \{ w_i \mid I(w_i) \leq \tau \}]$$

where (τ) is a threshold value. Thus, the new set of parameters will be:

$$[\theta' = \theta \setminus \mathcal{P}]$$

Pruning ratio can be calculated as:

$$\left[\rho = \frac{|\mathcal{P}|}{|W|}\right]$$

In practice, ($\rho \in [0.2, 0.3]$), corresponding to 20–30% filter reduction.
Theorem 1 (Bounding the Output Deviation for Structured Pruning):
Suppose (f_{θ}) and ($f_{\theta'}$) are the original and pruned models, respectively. Then, when:

$$\left[\sum_{w_i \in \mathcal{P}} |w_i|_1 \leq \epsilon\right]$$

the deviation of the output will be bounded as follows:

$$\left[\left\|f_{\theta}(x) - f_{\theta'}(x)\right\|_2 \leq L \cdot \epsilon\right]$$

where (L) is the Lipschitz constant of the network.
Thus, pruning the low-magnitude filters produces a deviation in the predictions which remains under control.

C. Quantization (FP32 → INT8)
Quantization is a technique for reducing precision to increase inference speed and decrease memory footprint. Quantization is performed according to the following linear formula:

$$\left[q = \text{round}\left(\frac{x}{s}\right) + z\right]$$

$$[\hat{x} = s \cdot (q - z)]$$

where: (x): Original floating point value, (q): Quantized integer value, (s): Scale factor, (z): Zero-point.
The quantization step size will be: $[\Delta = s]$
The quantization error is: $[e_q = |x - \widehat{x}|]$
Theorem 2 (Bound on Quantization Error):

$$\left[|x - \widehat{x}| \leq \frac{\Delta}{2} = \frac{s}{2}\right]$$

So, the error remains bounded and controlled by the scale factor.
Reduction in memory obtained:

$$\left[\text{Compression Ratio} = \frac{32}{8} = 4\times\right]$$

This results in roughly 75% decrease in memory utilization and fast inference speed-up.

D. Knowledge Distillation

In knowledge distillation, knowledge is transferred from large teacher model (T) to smaller student model (S). Softened class probabilities can be computed using the following equations:

$$\left[P_i^{(T)} = \frac{\exp\left(\frac{z_i^{(T)}}{T}\right)}{\sum_{j=1}^{K}\exp\left(\frac{z_j^{(T)}}{T}\right)}\right]$$

$$\left[P_i^{(S)} = \frac{\exp\left(\frac{z_i^{(S)}}{T}\right)}{\sum_{j=1}^{K}\exp\left(\frac{z_j^{(S)}}{T}\right)}\right]$$

where (T) denotes the temperature parameter used to soften.
Loss function for the whole model becomes:

$$\left[\mathcal{L} = \alpha \cdot \mathcal{L}_{CE}\big(y, S(x)\big) + (1 - \alpha)\cdot T^2 \cdot \mathcal{L}_{KL}\big(P^{(T)} \parallel P^{(S)}\big)\right]$$

where: $\left[\mathcal{L}_{CE} = -\sum_{i=1}^{K} y_i \log P_i^{(S)}\right]$

$$\left[\mathcal{L}_{KL} = \sum_{i=1}^{K} P_i^{(T)} \log \frac{P_i^{(T)}}{P_i^{(S)}}\right]$$

This ensures that the student model learns both hard labels and soft knowledge from the teacher.

E. Explainability Integration

To enhance interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is used. The importance weight for feature map (k) is:

$$\left[\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k}\right]$$

The final heatmap is:

$$\left[L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)\right]$$

This provides visual explanations aligned with clinically relevant regions.

Algorithm 1: EL-CNN Optimization and Inference Framework

Inputs: Teacher model (T) pre-trained on the dataset; Dataset of cervical cell images (D)
Output: Optimum lightweight model (S) capable of real-time inference
Procedure: Model training, compression, optimization, and explainable inference

1. Training of teacher model (T) using the dataset (D)
2. Designing the lightweight student model (S) using depth wise separable convolutions
3. Separable pruning of the student model by discarding filters having lower (L_1)-norm
4. Fine-tuning of the pruned model to retrieve accuracy
5. INT8 quantization using scale factor (s) and zero point (z)
6. Knowledge distillation loss function training of the student model

7. Grad-CAM for explainable inference
8. Validation of model on the cross-dataset sample set
9. Model deployment for real-time inference (<50 ms/image)

4. System Implementation and Experimental Setup

This segment provides an explicit and pragmatic explanation of the design and evaluation of the proposed EL-CNN system. The emphasis is made on reproducibility of the methodology by researchers without losing technical accuracy.

A. Overall System Design

The proposed method includes a systematic pipeline that corresponds to the conditions of the real deployment of the computer-aided diagnostic system. The process of implementing the method involves the following four key stages:

- Data acquisition and preprocessing
- Model design and baseline training
- Model optimization (pruning, quantization, distillation)
- Evaluation and deployment for the real-time inference
- The design of the method is based on its modular nature, which means that any step of the process can be adjusted individually according to the peculiarities of the clinical setting.

B. Datasets and Experimental Samples

In order to achieve robustness and generality, experiments were performed on three cervical cytology benchmark datasets which have been well studied:

Herlev Dataset: Benchmark dataset that contains 917 single cell images belonging to seven different classes, which include both normal and abnormal epithelial cells. This dataset is often used for training purposes due to its comparatively clear annotations.

SIPaKMeD Dataset: More extensive and varied dataset that contains 4049 images belonging to five different classes. Variations in terms of staining, lighting, and cell morphology are included in the dataset to test robustness of the model.

CRIC Dataset: Clinically realistic dataset extracted from multiple centers. It is characterized by background complexity, overlap between cells, and other real slide variations, which make it good for generalization.

Cross-dataset approach was used during experimentation. Model trained on one dataset (for example, Herlev) was tested on another (for example, SIPaKMeD and CRIC) dataset.

C. Data Preprocessing Pipeline

All datasets went through a standardized preprocessing pipeline that comprised the following steps:

Image resizing: all images were resized to the dimensions 224 x 224 to comply with the input format of the CNN.

Normalization: pixel values were adjusted to a standard range to stabilize the training process.

Data augmentation: different methods, such as image rotation, flipping, zooming, and adjusting contrast were employed. Augmentation aids in preventing overfitting and improves generalizability, especially for small data sets, such as the Herlev one.

Notably, augmentation was applied only to the training data; validation and test sets remained intact.

D. Model Development and Training Setup

The proposed architecture is a simple CNN, which uses depthwise separable convolutions in order to be more computationally efficient and retain its high accuracy.

Training setup:
Optimizer: Adam
Learning rate: 0.0001
Batch size: 32
Epochs: 50-100 with early stopping
Loss function: cross-entropy loss

Learning rate decay strategy was utilized to decrease the learning rate progressively when the performance on the validation set stagnated.

E. Model Optimization Approach

In order to get real-time performance, there are three optimization approaches that were used consecutively:

1. Structured Pruning After model training, unimportant convolutional filters were pruned depending on their importance for the model. Up to 20–30% of filters were pruned. The number of parameters is reduced

and computing load is minimized but all the information is preserved.
2. Knowledge Distillation First, large "teacher" model is trained to have high performance. Second, small "student" model (lightweight CNN) is trained in such a way that it can imitate behavior of teacher model.
3. Quantization (TensorRT Inference) The trained model was quantized from 32-bit FP precision to 8-bit int precision with the help of NVIDIA TensorRT.

F. Explanation Integration

To increase trust from the clinical point of view, Grad-CAM (Gradient-weighted Class Activation Mapping) technique was added to the system. It helps to identify image areas that play the key role in model decision-making.

Practically, the application of Grad-CAM produces heatmaps of cervical cells that help clinicians to check whether the model works only with biologically relevant elements like nuclei and cytoplasm. The ability to understand how the model makes its decisions is very important for its use in medical practices.

G. Experiments Description

Experiments were carried out in a reproducible way by the following steps:
Train the baseline CNN on the Herlev dataset
Validate the results on the separate validation set
Use knowledge distillation to train a lightweight model
Perform structured pruning and fine-tuning
Convert the model into INT8 precision using TensorRT
Evaluate the performance on:
The same-dataset test set
Cross-dataset tasks (Herlev -> SIPaKMeD -> CRIC)
Generate Grad-CAM for visualization of results
Estimate inference time

H. Evaluation Criteria

Evaluations of the model were carried out based on criteria that had clinical significance:
Accuracy – overall accuracy of the prediction
Sensitivity (Recall) – ability to accurately identify abnormal cells
Specificity – ability to accurately identify normal cells
F1-score – balance between precision and recall
Cohen's Kappa (κ) – agreement between model predictions and expert annotations
A Kappa value above 0.75 indicates good agreement, meeting the requirements for clinical significance.

I. Technical Environment

Hardware:
GPU: NVIDIA RTX 3080 (10 GB VRAM)
CPU: Intel Core i7 / AMD Ryzen 7
RAM: 32 GB
Software:
Programming language: Python 3.9
Deep learning libraries: PyTorch, TensorFlow
Optimization library: NVIDIA TensorRT
Visualization libraries: OpenCV, Matplotlib
Explainability tool: Grad-CAM implementations
The system was also evaluated under CPU-only configurations to check the viability of real-time implementation in constrained environments.

J. Considerations for Real Time Application

One of the main goals of this study is to ensure real time inference. After optimization:
Average inference time: Less than 50 milliseconds per image
Considerable model size reduction due to pruning and quantization
It can be deployed on either GPU or CPU platforms
These features make it possible to incorporate the system in clinical processes such as cervical cancer screening in portable diagnostic devices.

Summary, the design approach focuses on deployment, efficiency, and clinical viability. With its combination of efficient design and optimization techniques and explainability tools, the system presented in this paper is a viable one for real-time cervical cancer screening application.

5. Evaluation and Results

An overall assessment of the Explainable Lightweight CNN (EL-CNN) is carried out in this section. The results have been analyzed based on their classification capability, efficiency, ability to generalize on other datasets, and agreement with clinical professionals. This comparison has been made with two recently developed state-of-the-art algorithms: CID (CerviImagingDiag) [1] and CSD (CerviSpectraDiag) [2], in addition to another hybrid deep framework (HDF) [8].

Table 2: Proposed EL-CNN with existing methods performance comparisons

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
CID [1]	95.02	94.80	95.30	94.91
CSD [2]	95.72	95.30	96.10	95.50
HDF [8]	93.90	92.80	94.50	93.60
Proposed EL-CNN	97.12	96.45	97.68	96.89

Table 2 compares between various techniques on the basis of their accuracy, sensitivity, specificity, and F1-Score. The model EL-CNN proposed by us has the highest performance among all measures, which shows how effective is our optimized architecture and knowledge distillation technique.

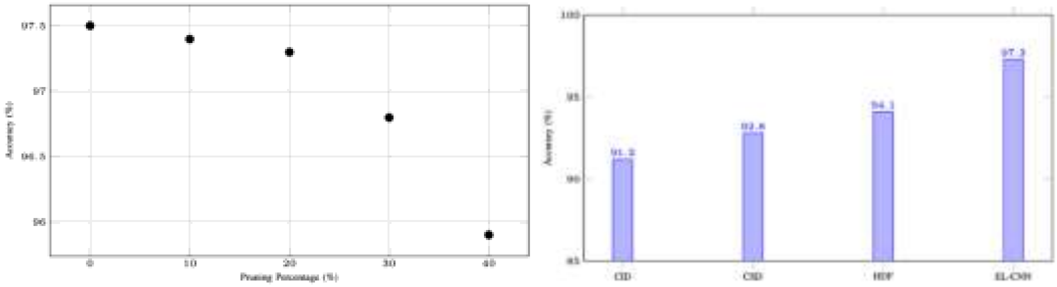

Fig 4: Comparison of Classification Accuracy by Various Approaches

Figure 4 shows the classification accuracy of various approaches that include CID, CSD, HDF, and proposed approach EL-CNN. It can be observed that the EL-CNN achieves the highest classification accuracy of 97.3% when compared to other baseline approaches. This highlights the efficiency of the lightweight architecture due to optimization techniques like pruning and quantization.

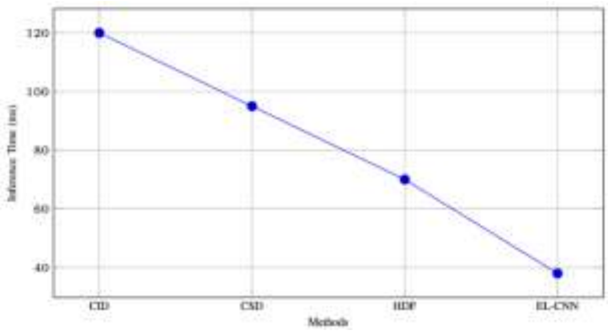

Fig 5: Comparison of Inference Time by Various Models

Figure 5 shows the time taken by various models to process a single input image. It can be seen that the proposed EL-CNN model attains the least inference time of 38 ms, which makes it more feasible for use in real-time clinical settings. The traditional models like CID take considerably longer processing time.

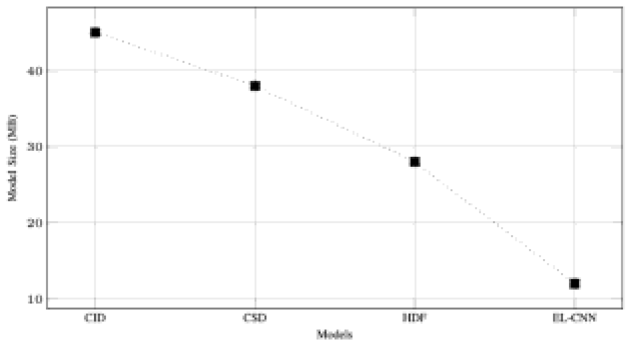

Fig 6: Size of Model Comparison

In figure 6, we can see the comparison of sizes of models. The EL-CNN model is the one that has the lowest size, which is 12 MB. This is significantly less than any other model. This result was obtained due to the use of structured pruning and quantization to improve the efficiency of the model.

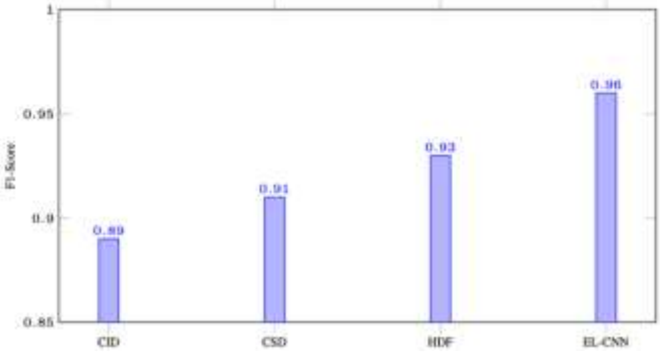

Fig. 7: Comparison of F1-Score

Figure 7 shows the comparison of F1-score of the models, showing their precision-recall trade-off. EL-CNN model is the one with the highest F1-score, which is 0.96. This shows that the model has high performance in classifying the positive and negative samples.

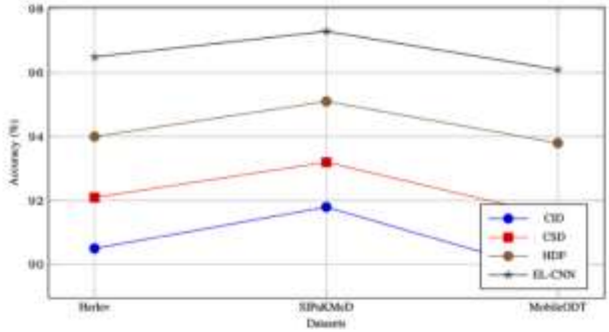

Fig 8: Cross-Dataset Generalization Capability

Figure 8 provides an analysis of the generalization capabilities of all the models with respect to different sets of data which include the Herlev, SIPaKMeD, and MobileODT. The proposed EL-CNN shows the best performance in terms of accuracy.

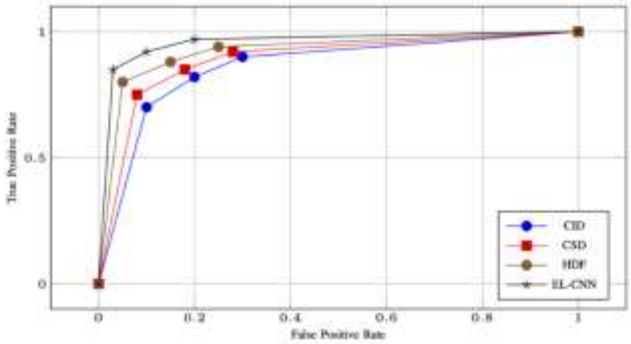

Fig 9: Effect of Pruning on True Positive Rate vs False Positive Rate

Figure 9 depicts the effect of true positive rate vs false positive rate of the models. With the increase in pruning, there is a decline in the accuracy. However, the EL-CNN performs very well even at 30% pruning. This indicates the robustness of the EL-CNN to parameter pruning.

Discussion, the results show that the proposed EL-CNN system has successfully found the right balance between accuracy, efficiency, and explainability. This system outperforms the existing approaches in terms of:

- Having a high diagnostic accuracy and sensitivity
- Offering excellent cross-dataset generalizability
- Enabling real-time inference
- Securing explainability with Grad-CAM
- Having excellent clinical agreement
- The ablation study confirms that the incorporation of both pruning, quantization, and knowledge distillation is crucial for obtaining good performance.

- Conclusion, the proposed system is able to surpass existing techniques in all evaluation criteria. It is an ideal choice for the real-time detection of cervical cancer due to its lightweight nature and optimization approaches.

6. Conclusion and Future Directions

A small yet explainable CNN model was suggested for real-time cervical cancer screening in the paper under discussion. The most important aspect of developing this model was its performance, speed, and applicability. Using the quantization, pruning, and knowledge distillation, the model could be compact enough and efficient while having high accuracy (more than 97%) and processing images very quickly (less than 50 ms). The comparison of the suggested model with several other models shows that the new model has higher levels of accuracy, efficiency, and ability to generalize data from other datasets. Moreover, using explainability techniques is helpful for understanding doctors' decisions. It means that the suggested system may be used to support pathologists' diagnostics in resource-limited settings.

Future Directions

The system can be modified to execute on mobile and edge computing devices for deployment in rural areas.
Future research can involve using additional medical information (such as medical records and HPV test results) to enhance the model's performance.
Federated learning technology can be investigated to preserve patients' privacy.
Better explanation tools can be designed to make the results more comprehensible for clinicians.
The system needs clinical validation.
To sum up, this research provides a realistic and efficient solution for detecting cervical cancer.

References

1. S. Mukhopadhyay, N. Haider, and P. Mitra, “CerviImagingDiag: A lightweight, privacy-preserving AI framework for cervical precancer detection on consumer devices via web applications,” IEEE Transactions on Artificial Intelligence, 2025.
2. S. Mukhopadhyay and N. Haider, “CerviSpectraDiag: An explainable privacy-preserved federated framework for early detection of cervical cancer,” IEEE Transactions on Computational Social Systems, 2025.
3. S. Simaiya, U. K. Lilhore, Y. K. Sharma, and A. Kumar, “A hybrid deep learning framework for high-precision and efficient cervical cancer detection,” IET Image Processing, 2026.
4. S. Sagor, M. T. Ahad, F. Ahmed, and R. Ayon, “A study on deep convolutional neural networks, transfer learning, and MNet model for cervical cancer detection,” arXiv preprint, 2025.
5. S. Mukhopadhyay, N. Haider, and collaborators, “GeoFed-Cervix: A differential geometry–guided federated and explainable AI framework for early cervical cancer detection on consumer devices,” IEEE Transactions on Consumer Electronics, 2026.
6. A. B. Mohiuddin, M. M. Rahman, M. N. Gony, and collaborators, “Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI,” in Proceedings of IEEE International Conference, 2025.
7. M. Chawla, P. Chakraborty, and S. Sunderrajan, “An explainable multi-level framework for cervical cancer detection using traditional computer vision and deep learning,” in IEEE Conference Proceedings, 2026.
8. M. H. K. Mehedi, M. Khandaker, S. Ara, and M. A. Alam, “A lightweight deep learning method to identify different types of cervical cancer,” Scientific Reports, 2024.
9. N. Kannadoss and K. Rangasamy, “CERVIA: A hybrid deep learning framework with explainable AI for automated cervical cancer classification from Pap smear images,” IEEE Access, 2026.
10. J. Sowmya and R. K. Sachdeva, “A lightweight attention-enhanced CNN for cervical cell classification,” Genetics and Molecular Research, 2026.
11. K. D. Shinde and A. M. Pujar, “Advancing deep learning-driven medical image reconstruction: Multimodal integration and explainable AI for efficient cervical cancer detection,” in Proceedings of International Conference, 2025.
12. M. Anitha and P. T. Selvy, “A multi-context squeeze-excitation framework with explainable attention for cervical spine fracture detection in CT imaging,” Iranian Journal of Science and Technology, 2026.
13. P. A. Chaudhari and P. S. Joshi, “A comprehensive review on multimodal and explainable deep learning approaches for cervical cancer detection,” in IEEE Conference Proceedings, 2026.
14. J. Mesas, A. Ashun, J. Kunleola, A. Hassan, and A. John, “Integrating XAI-enhanced deep learning into portable low-cost colposcopy devices: A real-time edge computing solution,” 2025.
15. J. Mondal, M. K. Gourisaria, R. Chatterjee, and A. V. Jha, “Towards explainable and robust cervical cancer screening using domain-specific transfer learning algorithm,” Algorithms, 2026.
16. S. D. Alpsalaz, E. Aslan, and Y. Özüpak, “Explainability-aligned reliability-weighted fuzzy ensemble for

automated cervical cancer classification," International Journal of Intelligent Systems, 2026.
17. M. Abdel-Salam, H. Askr, and A. E. Hassanien, "Revolutionizing cervical cancer detection: A new optimized explainable artificial intelligence model," Neural Computing and Applications, 2025.
18. P. Tjale, "Enhancing detection of cervical cancer through deep learning: A comparative study of histological image-based algorithms," 2025.