Array 27 (2025) 100468

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

NEDL-GCP: A nested ensemble deep learning model for Gynecological cancer risk prediction

Kamal Berahmand [a], Xujuan Zhou [b,*], Yuefeng Li [a], Raj Gururajan [b], Prabal Datta Barua [b], U Rajendra Acharya [c], Srinivas Kondalsamy Chennakesavan [d]

[a] Department of Science and Engineering, Queensland University of Technology, Brisbane, Australia
[b] School of Business, University of Southern Queensland, Ipswich, Australia
[c] School of Mathematics, Physics and Computing, University of Southern Queensland, Ipswich, Australia
[d] Rural Clinical School, University of Queensland, Toowoomba, Australia

ARTICLE INFO

Keywords:
Cancer diagnostics
Gynecological cancer
Deep learning
Neural networks
Ensemble learning

ABSTRACT

Gynecological cancer remains a critical global health concern, where early detection significantly improves patient outcomes. Despite advances in deep learning for medical diagnostics, existing models often struggle with feature redundancy, lack of generalizability, and suboptimal integration of diverse feature representations, limiting their effectiveness in clinical applications. In this study, we present NEDL-GCP, a Nested Ensemble Deep Learning model for Gynecological Cancer Risk Prediction, which uses a hierarchical ensemble framework to improve the accuracy of the classification. NEDL-GCP integrates CNNs, RNNs, and SVMs as base learners, extracting diverse feature representations, while a meta-classifier combining J48 and Stochastic Gradient Descent (SGD) refines predictions. Evaluated on the Herlev and SIPaKMeD Pap Smear datasets, NEDL-GCP achieved state-of-the-art accuracy scores of 99.1% and 98.5%, outperforming existing methods. These results demonstrate the robustness and reliability of the model, making it a valuable tool for the early detection of cervical cancer. By enhancing diagnostic accuracy and optimizing clinical workflows, NEDL-GCP supports timely decision-making, ultimately improving patient care.

1. Introduction

Cancer remains a leading cause of death worldwide, with 9.6 million deaths recorded in 2018, accounting for one in six global deaths [1]. Its incidence is projected to increase by 70% over the next two decades, posing a significant public health challenge [2]. In Australia, a person is diagnosed with cancer every four minutes, and in 2014, cancer-related deaths accounted for nearly 30% of all deaths [3,4]. Cancer imposes a substantial social burden and remains the leading cause of disease-related mortality in the nation [5]. Gynecological cancers, including ovarian, uterine, cervical, vaginal, and vulval cancers, contribute significantly to cancer-related morbidity and mortality among women [3,6]. More than half of these cancers are classified as rare, making clinical trials and treatment standardization challenging [5]. In addition, societal stigma often prevents women from discussing gynecological health concerns, leading to delays in screening and late-stage diagnoses, thus increasing the risk of preventable deaths.

Cervical cancer, a prevalent gynecological malignancy, originates in the cervix, the lower part of the uterus that connects to the vagina [7,8]. It is mainly caused by persistent infection with high-risk human papillomavirus (HPV), a common sexually transmitted pathogen. The disease progresses gradually, beginning with precancerous cellular changes that, if not diagnosed or treated, can develop into invasive cancer and spread to other tissues [9]. However, cervical cancer is highly treatable when detected early. Effective screening programs, including HPV vaccination and routine screenings, enable timely interventions and early-stage treatment. By identifying high-risk individuals, healthcare providers can implement personalized screening strategies, improve patient outcomes, and reduce mortality rates.

Fig. 1 illustrates a motivating example comparing traditional and machine learning-based diagnostic pathways for cervical cancer, highlighting the limitations of manual interpretation and the potential for automation in risk prediction. Accurate models of cervical cancer risk prediction are therefore essential to enable personalized medicine,

* Corresponding author.
E-mail addresses: kamal.berahmand@hdr.qut.edu.au (K. Berahmand), xujuan.zhou@unisq.edu.au (X. Zhou), y2.li@qut.edu.au (Y. Li), raj.gururajan@unisq.edu.au (R. Gururajan), prabal.barua@unisq.edu.au (P.D. Barua), rajendra.acharya@unisq.edu.au (U.R. Acharya), s.kondalsamychennakes@uq.edu.au (S.K. Chennakesavan).

https://doi.org/10.1016/j.array.2025.100468
Received 15 May 2024; Received in revised form 12 July 2025; Accepted 14 July 2025
Available online 23 July 2025
2590-0056/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

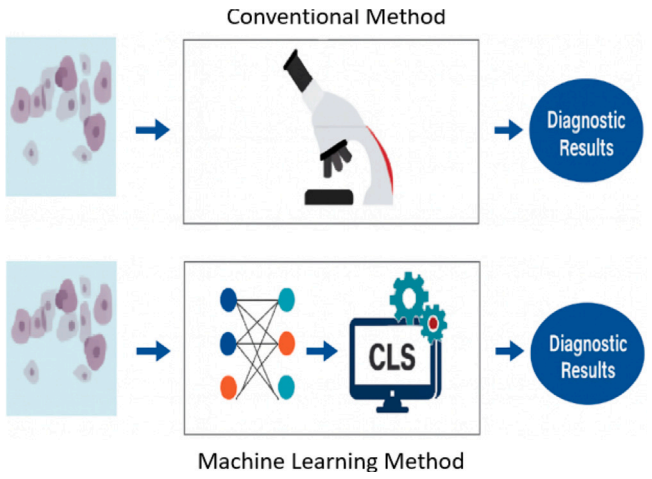


Fig. 1. Comparison between conventional (microscope-based) diagnostics and machine learning–driven cervical cancer risk prediction.

allowing interventions to be tailored to individual risk profiles [10]. Such models enhance diagnostic efficiency, reduce unnecessary testing for low-risk individuals, and prioritize high-risk cases, ultimately improving clinical outcomes while lowering overall healthcare costs [11]. Despite advances in predictive modeling within oncology, the application of machine learning and deep learning techniques to develop and validate prognostic tools remains limited [12–14]. To address this gap, this study introduces a novel machine learning–based approach for cervical cancer risk prediction, contributing to the advancement of gynecological oncology.

Machine learning (ML), a core area within artificial intelligence, has emerged as a powerful tool for uncovering complex patterns in medical data, driving major advances in cancer prediction and recurrence assessment [15–17]. Both traditional non-deep learning algorithms (e.g., with a single hidden layer) and deep learning architectures (with multiple layers) have demonstrated effectiveness in cancer diagnostics across various modalities [18–21]. While Support Vector Machines (SVMs) continue to be widely adopted due to their robustness in high-dimensional spaces [22,23], they often demand labor-intensive feature engineering steps [24]. Deep learning addresses this limitation by automatically extracting relevant features through hierarchical representations, enabling efficient end-to-end training pipelines with minimal manual intervention [25,26]. Architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are especially well-suited for modeling complex, nonlinear relationships in biomedical data. To further enhance predictive performance and generalizability, ensemble learning approaches—such as stacking, boosting, and majority voting—are increasingly used to integrate multiple base learners [27–29], improving overall reliability in clinical decision support.

Nested ensembling is an advanced extension of traditional ensemble methods. Although standard ensembles typically aggregate predictions from individual models, nested assembly introduces an additional layer of complexity [30,31]. The first layer of this approach consists of diverse base models, similar to conventional ensembles. However, instead of directly combining their predictions, the outputs from these base models serve as inputs for a second layer of models. This second layer processes and refines the aggregated outputs, applying additional learning mechanisms to enhance predictive accuracy. The strength of nested assembly lies in its ability to capture intricate relationships between base models, often leading to superior performance compared to traditional assembly techniques. Building on its success in image classification, this study uses nested ensembles for the prediction of gynecological cancer risk, with the objective of improving the diagnostic accuracy and robustness of the model.

The study presents a robust nested ensemble learning framework designed to improve cervical cancer risk prediction through deep learning. Unlike conventional ensemble methods that simply aggregate predictions from base models, our approach employs a hierarchical stacking mechanism to refine feature extraction and enhance classification performance. In the first phase, we integrate multiple deep learning architectures, including CNNs, RNNs, and SVMs, each capturing distinct patterns within medical imaging data. The second phase utilizes a MetaClassifier with a voting mechanism that strategically combines J48 and SGD to optimize final predictions. This structured two-layer learning paradigm enhances the robustness of the model, mitigates overfitting, and improves generalizability between datasets.

The key contributions of this study are as follows.

- We propose NEDL-GCP, a nested ensemble deep learning framework that integrates CNNs, RNNs, and SVMs as base learners. This approach effectively combines spatial, sequential, and statistical patterns in medical images, improving feature representation and classification accuracy.
- We introduce a two-tier MetaClassifier that utilizes J48 and SGD for adaptive decision fusion. By dynamically weighting predictions from multiple base models, this hierarchical structure enhances classification performance, improves model robustness, and reduces overfitting.
- We validate NEDL-GCP on the Herlev and SIPaKMeD Pap Smear datasets, achieving accuracy rates of 99.1% and 98.5%, respectively. These results demonstrate the reliability and potential of the model for advancing automated cervical cancer diagnostics.

The structure of the paper is organized as follows: Section 2 offers a comprehensive review of the related literature in cancer prediction models. Section 3 outlines our model development methodology. The experimental results are presented in Section 4. Section 5 discusses the implications, providing a comparative analysis with prior studies. Finally, Section 6 summarizes the main contributions of our study and outlines future research directions.

2. Related work

Gynecological cancers, including ovarian, uterine, cervical, vaginal, and vulvar cancers, represent a significant global health concern. Among them, cervical cancer remains one of the most preventable malignancies, primarily caused by persistent infection with high-risk human papillomavirus (HPV). Early detection and accurate risk prediction are essential to reduce mortality rates, where machine learning and deep learning techniques have increasingly played a vital role. This section reviews existing methods for the prediction of cervical cancer risk, categorizing them into single-model approaches and ensemble-based learning strategies.

2.1. Single methods for gynecological cancer risk prediction

Traditional approaches for cervical cancer detection, such as Papanicolaou (Pap) smear test and HPV DNA screening, remain the gold standard for early diagnosis. However, these methods are limited by interobserver variability, high false negative rates, and dependency on expert interpretation, prompting the need for automated and more reliable risk prediction techniques [2]. Machine learning has emerged as a promising alternative that offers data-driven models capable of improving diagnostic accuracy and minimizing human bias. Several studies have explored single-model machine learning approaches to improve cervical cancer detection. Kaushik et al. (2021) [32] developed a predictive model utilizing logistic regression, ridge classifiers and Gaussian Naive Bayes classifiers to analyze cytokine gene variants and sociodemographic risk factors. Their study highlighted the importance of genetic predisposition and environmental factors in assessing the risk

of cervical cancer, demonstrating the potential of statistical machine learning models in clinical decision making.

Neural network-based cytology image analysis has also shown promise in automating cervical cancer screening. Singh et al. (2015) [33] introduced a neural network-driven classification system that employed image processing techniques to extract morphological features such as the nucleus-to-cytoplasm ratio, color intensity, and shape irregularities. The system, trained using the Backpropagation algorithm, effectively differentiated between non-cancerous, low-grade, and high-grade cancerous cells, offering an efficient alternative to manual cytology assessment. Deep learning techniques, particularly convolutional neural networks (CNNs), have further improved classification performance. Mingshi et al. (2019) [34] designed a CNN-based model for the classification of cervical exfoliated cells, leveraging hierarchical feature extraction to distinguish between normal and malignant structures. Similarly, Li et al. (2019) [35] employed a CNN model based on transfer learning, pre-trained in large-scale medical datasets, to improve the representation of features and achieve higher accuracy in the classification of cervical cancer images. These studies underscore the effectiveness of deep learning in automating cervical cancer diagnostics.

Despite their advancements, single-model approaches are often susceptible to overfitting, data variability, and limited generalizability between diverse populations. These challenges have driven the adoption of ensemble learning strategies, combining multiple models to improve robustness, enhance classification accuracy, and ensure reliable real-world clinical applications.

2.2. Ensemble methods for gynecological cancer risk prediction

Ensemble learning techniques have emerged as a powerful approach to improve cervical cancer prediction by combining multiple models to improve classification accuracy, robustness, and generalizability. Unlike single-model classifiers, ensemble approaches leverage diverse learning architectures, allowing for more comprehensive feature extraction and decision-making. Ensemble methods have demonstrated superior performance in automated cervical cancer detection by integrating various machine learning and deep learning models.

An ensemble-based cervical cancer prediction model was introduced by Lu et al. (2020) [36], utilizing a voting mechanism among five classifiers, including logistic regression, decision trees, support vector machines (SVMs), multilayer perceptrons, and k-nearest neighbors. To enhance classification robustness, the model incorporated a gene-assistance module that integrates genetic biomarkers into the prediction process. Curia et al. (2021) [37] developed an explainable ensemble framework that combined machine learning classifiers with interpretability techniques such as LIME and Shapley values. The integration of explainable AI significantly improved the transparency of cervical cancer risk prediction, making the model more suitable for clinical decision-making.

For cervical cancer diagnosis using colposcopy images, Chandran et al. (2021) [38] proposed an ensemble deep learning architecture named CYENET. The model utilized VGG19 for transfer learning and incorporated a novel classification fusion approach to enhance feature extraction and improve diagnostic accuracy. Ali et al. (2024) [39] presented a machine learning ensemble classifier integrating Random Forest, SVM, Gaussian Naïve Bayes, and Decision Tree models for cervical cancer prediction. The study emphasized model interpretability by incorporating SHapley Additive exPlanations (SHAP), ensuring greater transparency in clinical applications.

A stacked ensemble learning framework was proposed by Aljrees et al. (2024) [40], combining Random Forest, SVM, and XGBoost. The approach also employed KNN imputation to handle missing data, improving classification robustness for real-world cervical cancer diagnosis. Uddin et al. (2024) [41] designed an ensemble machine learning framework using hybrid feature selection techniques. By integrating Principal Component Analysis (PCA) and XGBoost for optimal feature selection, the model leveraged a voting-based ensemble that combined Random Forest and Multilayer Perceptron classifiers. Additionally, Random Oversampling was applied to mitigate class imbalance, further enhancing predictive performance. Kwatra et al. (2025) [42] introduced an ensemble deep learning architecture integrating ResNet50 and Inception V3. The model exploited the complementary strengths of both architectures to improve feature extraction and classification accuracy in gynecological cancer detection.

Despite the advancements of ensemble-based methods in cervical cancer prediction, challenges such as computational complexity, the need for large-scale annotated datasets, and standardization of clinical evaluation remain significant. Future research should focus on optimizing ensemble architectures, incorporating multi-modal data sources, and improving interpretability to enhance real-world clinical applicability.

3. NEDL-GCP method

Ensemble learning enhances classification by integrating multiple models to improve generalization and mitigate overfitting. Traditional methods, such as bagging and boosting, aggregate independent classifiers but lack direct interaction between models, limiting their ability to exploit diverse feature representations. Nested ensemble learning addresses this limitation by introducing multiple refinement layers, where base classifiers generate predictions further optimized by a meta-classifier, enabling deeper feature extraction and adaptive learning.

To improve cervical cancer risk prediction, we propose NEDL-GCP (Nested Ensemble Deep Learning for Gynecological Cancer Prediction), a two-layer ensemble framework that integrates deep learning and traditional classifiers. The first layer extracts spatial, sequential, and statistical features using CNNs, RNNs, and SVMs. In contrast, the second layer refines predictions through a meta-classifier combining J48 decision trees and stochastic gradient descent (SGD). This structured approach improves classification accuracy and robustness. The following subsections detail the architecture, training strategy, and optimization.

3.1. Architectural design of NEDL-GCP

The NEDL-GCP framework consists of a two-layer ensemble structure designed to improve the accuracy of the classification. The first layer, the base classification layer, extracts diverse feature representations, while the second layer, the meta-classification layer, refines predictions for improved decision-making. Fig. 2 illustrates the overall framework.

(1) Base Classification Layer: This layer consists of independent classifiers that extract distinct feature representations: CNNs: Capture spatial patterns, texture, and morphological structures from cervical smear images. RNNs: Model sequential dependencies and structured imaging patterns. SVMs: Provide robust decision boundaries for linear and nonlinear classification.

(2) Meta-Classification Layer: Predictions from the base classifiers are aggregated and refined by the meta-classifier for final decision-making: J48 Decision Tree: Captures hierarchical relationships between predicted class probabilities. SGD: Iteratively updates model weights to enhance classification performance.

3.2. Training and hyperparameter optimization

The hyperparameters of the NEDL-GCP framework are optimized using grid search and cross-validation to enhance classification performance. CNNs employ ReLU activation with the Adam optimizer, while RNNs utilize tanh activation and RMSprop. The meta-classifier integrates J48 decision trees and stochastic gradient descent (SGD) for iterative learning. The key hyperparameters for all models are summarized in Table 1.

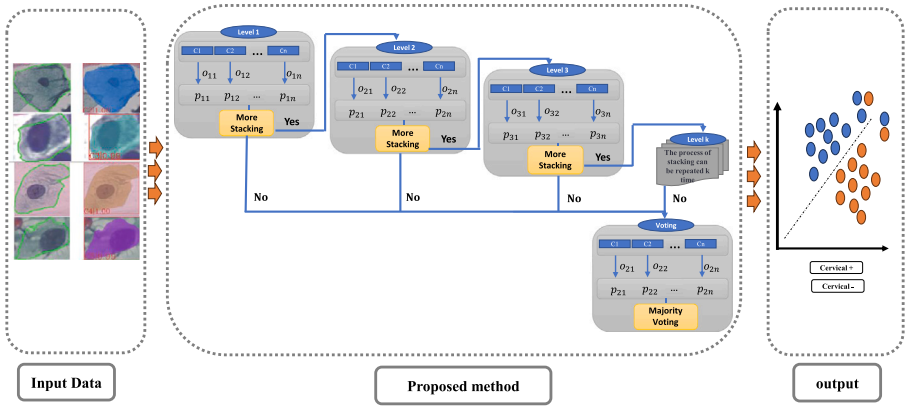


Fig. 2. Nested ensemble framework for cervical cell image classification.

Table 1
Key training parameters for base models and meta-classifier.

Model	Layers	Activation function	Optimizer	Learning rate
CNNs	5 Conv, 2 FC	ReLU	Adam	0.001
RNNs	3 LSTM, 1 FC	Tanh	RMSprop	0.0005
SVMs	–	RBF Kernel	–	–
J48	–	Decision Tree Splitting	–	–
SGD	–	Softmax	Stochastic Gradient Descent	0.01

3.3. Training algorithm for NEDL-GCP

The training process follows a structured two-stage approach. First, base classifiers (CNNs, RNNs, and SVMs) are trained independently to extract distinct feature representations. Their predictions are then refined by the meta-classifier (J48, SGD) to optimize classification accuracy. Cross-validation ensures robustness and prevents overfitting. The detailed training algorithm is outlined below.

Algorithm 1: Training Algorithm for NEDL-GCP

Input: Labeled dataset D, Cross-validation folds (K), Base models (M), Meta-classifiers (P), Feature set $\mathcal{F}$, Target classes (C)
Output: Optimized nested ensemble model $\mathcal{E}^*$

Split D into K folds for cross-validation
for each fold $i = 1$ to K **do**
  Partition D into training set D_{train} and validation set D_{val}
  for each base model $m = 1$ to M **do**
    Train CNN, RNN, or SVM on D_{train}
    Save trained model $\mathcal{M}_m$
  end for
  for each meta-classifier $p = 1$ to P **do**
    Generate validation predictions $\mathcal{P}_m$ from each $\mathcal{M}_m$
    Aggregate $\mathcal{P}_m$ using majority voting
    Train meta-classifier $\mathcal{M}_p$ (J48, SGD) on aggregated predictions
    Evaluate $\mathcal{M}_p$ using accuracy, precision, recall, and F1-score
  end for
end for
Train final nested ensemble model $\mathcal{E}^*$ using optimized parameters
Evaluate $\mathcal{E}^*$ on an independent test set and report performance metrics

The structured learning approach in NEDL-GCP enhances classification performance by integrating multiple classifiers, ensuring robust and interpretable predictions. Using various learning paradigms, this model sets a new benchmark in automated cervical cancer detection, optimizing both sensitivity and specificity.

4. Experiments

In this section, we present the experimental findings of our proposed method. We conducted a series of experiments to thoroughly evaluate the performance of our approach. The comparison methods were compiled in the Python programming language and implemented on a supercomputer with high-performance computing capabilities.

4.1. Datasets description

We used two publicly available datasets to develop and evaluate our cervical cancer prediction model: the Herlev dataset and SIPaKMeD Pap-Smear dataset. We carried out experiments on each dataset separately.

4.1.1. Herlev dataset

The Herlev dataset comprises 917 cervical smear images collected from the Department of Pathology at Herlev University Hospital in Denmark [43]. These images are categorized into seven distinct classes based on cervical cell morphology: superficial squamous epithelia (A, 70 images), intermediate squamous epithelia (B, 98 images), columnar epithelial cells (C, 74 images), mild squamous non-keratinizing dysplasia (D, 182 images), moderate squamous non-keratinizing dysplasia (E, 150 images), severe squamous non-keratinizing dysplasia (F, 146 images), and squamous cell carcinoma in situ (G, 197 images). To ensure consistency in image analysis, all images were pre-processed by normalizing their size and converting them to grayscale. This standardization helps improve feature extraction and improves the robustness of deep learning models. Fig. 3 illustrates the distribution of images in the seven categories of the Herlev dataset.

4.1.2. SIPaKMeD Pap-Smear dataset

The SIPaKMeD Pap Smear dataset contains 4049 isolated cell images extracted from 966 whole slide images [44]. These images are categorized into five distinct classes according to cytomorphological characteristics: Normal Superficial-Intermediate (831 images), Normal

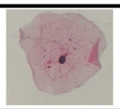
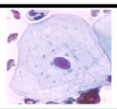
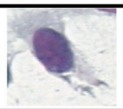
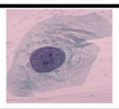
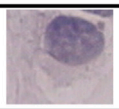
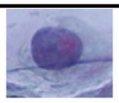
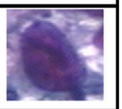
Class	1	2	3	4	5	6	7
Class type	A	B	C	D	E	F	G
Cell image							
Category	Normal	Normal	Normal	Abnormal	Abnormal	Abnormal	Abnormal
Number of Images	74	70	98	182	150	146	197

Fig. 3. Distribution of images in the Herlev dataset.

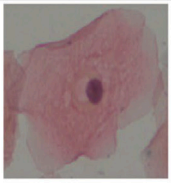
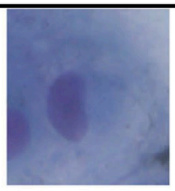
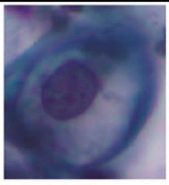
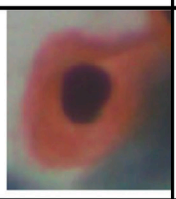
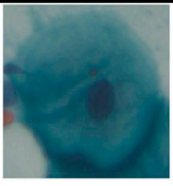
Class	1	2	3	4	5
Class type	Superficial-intermediate	Parabasal	Koilocytotic	Dyskeratotic	Metaplastic
Cell image					
Category	Normal	Normal	Abnormal	Abnormal	Benign
Number of Images	831	787	825	813	793

Fig. 4. Distribution of images in the SIPaKMeD Pap Smear dataset.

Parabasal (787 images), Abnormal Koilocytotic (825 images), Abnormal Dyskeratotic (813 images), and Benign Metaplastic (793 images). This dataset provides a diverse representation of cervical cytology, capturing both normal and abnormal cell types. This variability improves the robustness of deep learning models trained for automated cervical cancer detection. Fig. 4 illustrates the distribution of images in the five classes of the SIPaKMeD Pap smear dataset.

4.2. Data pre-processing

Data pre-processing is a crucial step in preparing datasets for machine learning models. In this study, data augmentation techniques were applied to enhance the training dataset by generating transformed versions of the original cell images. These transformations improve model generalization and mitigate overfitting. The augmentation process included horizontal flipping, rotations, and random scaling. Specifically, the images were rotated within a range of $\theta = -60$ to 60 degrees, scaled by a factor α ranging from 1.0 to 1.1, with a probability of $P_A = 0.75$ for scaling and $P_B = 0.5$ for horizontal flipping. These transformations introduced diversity into the training dataset, ensuring that the model learns invariant features and improves its performance on unseen data.

4.3. Baselines

In this subsection, we compare the proposed method with widely used deep learning models—VGG-16, VGG-19, ResNet-50, and XceptionNet—along with ensemble-based classifiers, including Late Fusion (LF), ML-EnsCC, and BRFEC. The evaluation is conducted on two benchmark datasets: Herlev and SIPaKMeD Pap Smear. To ensure fair comparison, all deep learning models were fine-tuned using the Stochastic Gradient Descent (SGD) optimizer and Rectified Linear Unit (ReLU) activation function.

- **VGG-16** [45]: A deep convolutional neural network (CNN) with 13 convolutional layers and 3 fully connected layers. It uses 3×3 filters and has been widely applied for image classification due to its strong feature extraction capabilities.
- **VGG-19** [45]: An extension of VGG-16 with 16 convolutional layers, maintaining the same filter sizes and improving feature representation.
- **ResNet-50** [46]: A residual deep network with 50 layers, incorporating skip connections to ease training and mitigate gradient vanishing issues. ResNet architectures have achieved state-of-the-art performance in image recognition tasks.
- **XceptionNet** [47]: A CNN that utilizes depthwise separable convolution layers for efficient parameter utilization while maintaining high classification accuracy. XceptionNet has demonstrated robust performance across multiple image processing tasks.
- **Late Fusion (LF)** [48]: An ensemble technique that aggregates predictions from multiple classifiers using majority voting. The final classification is determined by the highest vote count among the models, mathematically represented as:

$$\sum X(m,n) = \max Y_n \sum X(m,n)$$

where $X(m,n)$ represents the number of classifiers, Y_n denotes the number of classes, and $E(m,n) \in (0,1)$ indicates the decision of the ith classifier.
- **ML-EnsCC** [36]: An ensemble method that combines deep learning models (CNNs, RNNs) with traditional classifiers (SVMs) using stacking or majority voting to enhance cervical cancer detection accuracy.
- **BRFEC** [39]: An ensemble method for cervical cancer prediction using behavioral risk factors. It integrates Random Forest, SVM, Naïve Bayes, and Decision Tree in a stacking framework, employing feature selection, SMOTE, and SHAP for improved interpretability.

4.4. Evaluation metrics

The performance of NEDL-GCP is evaluated using four key metrics: accuracy, precision, recall, and F1-score [49,50]. These metrics provide a comprehensive assessment of classification effectiveness, balancing correct identification and misclassification rates.

Accuracy: Measures the proportion of correctly classified instances among all samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Table 2
Performance comparison of two layers in Herlev dataset.

Model	Precision	Recall	F1-Score	Accuracy
First Layer	0.939	0.941	0.931	0.952
Nested Ensemble	0.989	0.982	0.986	0.985

Table 3
Performance comparison of two layers on SIPaKMeD Pap Smear dataset.

Model	Precision	Recall	F1-Score	Accuracy
First Layer	0.933	0.942	0.930	0.930
Nested Ensemble	0.994	0.991	0.992	0.991

Precision: Indicates how many predicted positive cases are actually correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: Represents the model's ability to identify actual positive cases:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score: A harmonic mean of precision and recall, balancing both measures:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics collectively evaluate model reliability, ensuring a balance between sensitivity and specificity.

4.5. Results

This section presents the experimental findings of our proposed model, evaluating its performance across multiple datasets. We provide a comparative analysis with existing methods, assess statistical significance, and discuss key observations from the results.

4.5.1. Quantitative results on Herlev dataset

The effectiveness of the proposed method for detecting cervical cancer in the Herlev Pap smear dataset was evaluated through a series of experiments. A two-layer nested ensemble approach was utilized, achieving an F1 score of 0.986 and an accuracy of 0.985. These results highlight the robustness of the proposed method and its potential for improving cervical cancer detection.

Nested ensemble with two layers on Herlev dataset. The proposed nested ensemble framework consists of two layers. The first layer includes three base classifiers: CNN, RNN, and SVM, which extract spatial, sequential, and statistical features from cervical smear images. Their outputs are combined using a majority voting strategy to generate preliminary predictions. The second layer refines these predictions using an ensemble classifier composed of J48 decision trees and SGD. This hierarchical structure improves decision making by leveraging multiple learning mechanisms.

Table 2 presents the classification performance of each layer in terms of precision, recall, F1-score, and accuracy. The first layer achieved an F1-score of 0.931 and an accuracy of 0.952. The nested ensemble, which integrates the output from the first layer, demonstrated superior performance with an F1 score of 0.986 and an accuracy of 0.985.

The results confirm the effectiveness of the nested ensemble approach, demonstrating improved classification performance compared to individual base classifiers.

5-Fold cross-validation and hold-out 80:20 on Herlev dataset. To further validate the robustness of the proposed method, two evaluation techniques were employed: 5-fold cross-validation and hold-out 80:20. In 5-fold cross-validation, the dataset was divided into five equal subsets, where four were used for training and one for testing. This process was repeated five times to ensure stability in the results. The hold-out 80:20 method randomly split the dataset into 80% training and 20% testing.

The performance metrics for both techniques are depicted in Fig. 5. The 5-fold cross-validation experiment yielded an average accuracy of 0.985, precision of 0.985, recall of 0.986, and F1-score of 0.985. Similarly, the hold-out 80:20 experiment achieved an accuracy of 0.978, precision of 0.975, recall of 0.982, and F1-score of 0.981. These results indicate the consistency and reliability of the proposed method.

Comparison with state-of-the-art methods on Herlev dataset. In this experiment, we evaluated the performance of our proposed method on the Herlev Pap Smear dataset and compared it with seven state-of-the-art methods: Late Fusion (LF), VGG16, VGG19, ResNet50, XceptionNet, ML-EnsCC, and BRFEC. We used the same dataset division as in Experiment 2, where we randomly split the dataset into a training set (80% of the data) and a testing set (20% of the data). We trained our proposed method on the training set and evaluated its performance on the testing set. To ensure a fair comparison, we compared our results with the results reported in the literature for the seven state-of-the-art methods using the same evaluation metrics.

The results of the experiment are shown in Fig. 6. Our proposed method achieved a precision of 0.989, a recall of 0.985, an F1-score of 0.986, and an accuracy of 0.985. In particular, our method outperformed all the state-of-the-art methods in terms of precision, recall, F1 score, and accuracy.

4.5.2. Quantitative results on SIPaKMeD Pap Smear dataset

To further assess the generalizability of the proposed method, we evaluated its performance on the SIPaKMeD Pap Smear dataset. This publicly available benchmark dataset contains 4049 cervical cell images categorized into five diagnostic classes. Using the same experimental setup as in the Herlev dataset evaluation, we aimed to validate the robustness and adaptability of our approach in cervical cancer classification.

Nested ensemble with two layers on SIPaKMeD Pap Smear dataset. The nested ensemble approach was implemented with two layers to classify cervical cells within the SIPaKMeD dataset. The first layer consisted of three base classifiers: CNN, RNN, and SVM. Each classifier was trained independently on the dataset and their outputs were aggregated using a majority voting strategy to form initial predictions.

In the second layer, predictions from the base classifiers served as input features for another ensemble classifier, utilizing J48 and SGD algorithms to refine the final decision-making process. Majority voting was applied at this stage to further enhance classification accuracy.

Table 3 presents the performance of both classification layers. The first layer achieved an accuracy of 0.930 and an F1-score of 0.930, demonstrating strong baseline performance. The nested ensemble approach further improved classification results, achieving an accuracy of 0.991 and an F1-score of 0.992, confirming its effectiveness in cervical cell classification.

5-Fold cross-validation and hold-out 80:20 on SIPaKMeD Pap Smear dataset. To ensure a thorough evaluation, we tested the proposed method using two different validation strategies: 5-fold cross-validation and an 80:20 hold-out method. In 5-fold cross-validation, the dataset was partitioned into five subsets, four of which were used for training and one for testing in each iteration. This process was repeated five times to obtain an average performance measure. In the hold-out approach, the dataset was randomly split into 80% training and 20% testing.

The results are summarized in Fig. 7. The 5-fold cross-validation experiment yielded an average accuracy of 0.9918, precision of 0.9919, recall of 0.9919, and F1-score of 0.9918. Similarly, the hold-out 80:20 experiment achieved an accuracy of 0.9901, precision of 0.9899, recall of 0.9909, and an F1-score of 0.9903. These results confirm the consistency and robustness of the proposed method in cervical cancer classification.

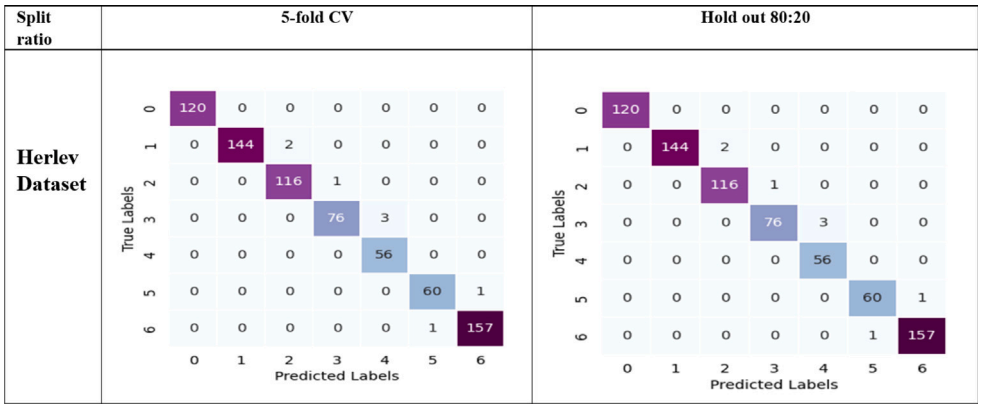


Fig. 5. Performance comparison of evaluation techniques on the Herlev dataset: 5-fold cross-validation and hold-out 80:20.

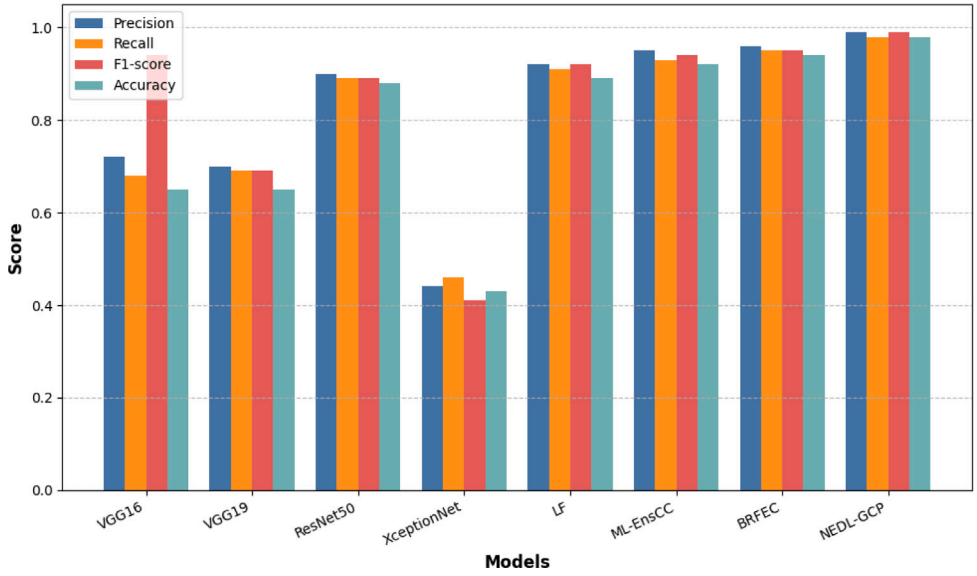


Fig. 6. Performance comparison of our proposed method with baseline models on the Herlev dataset in terms of precision, recall, F1-score, and accuracy.

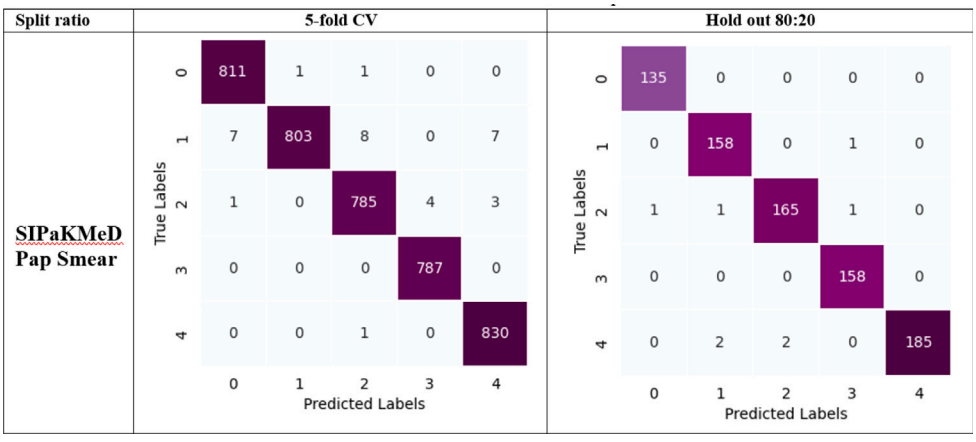


Fig. 7. Performance comparison of evaluation techniques on SIPaKMeD Pap Smear dataset: 5-fold cross-validation vs. hold-out 80:20 split.

Comparison with state-of-the-art methods on SIPaKMeD Pap Smear dataset. In this experiment, we evaluated the performance of our proposed method on the SIPaKMeD Pap Smear dataset and compared it with seven state-of-the-art methods: VGG16, VGG19, ResNet50, XceptionNet, Late Fusion (LF), ML-EnsCC, and BRFEC. We used the same dataset division as in Experiment 2, where we randomly divided the dataset into a training set (80% of the data) and a testing set (20% of the data). We trained our proposed method on the training set and evaluated its performance on the testing set. We then compared our results with the results reported in the literature for the seven state-of-the-art methods.

Fig. 8 shows the performance of each method in terms of precision, recall, F1-score, and accuracy. Our proposed method achieved

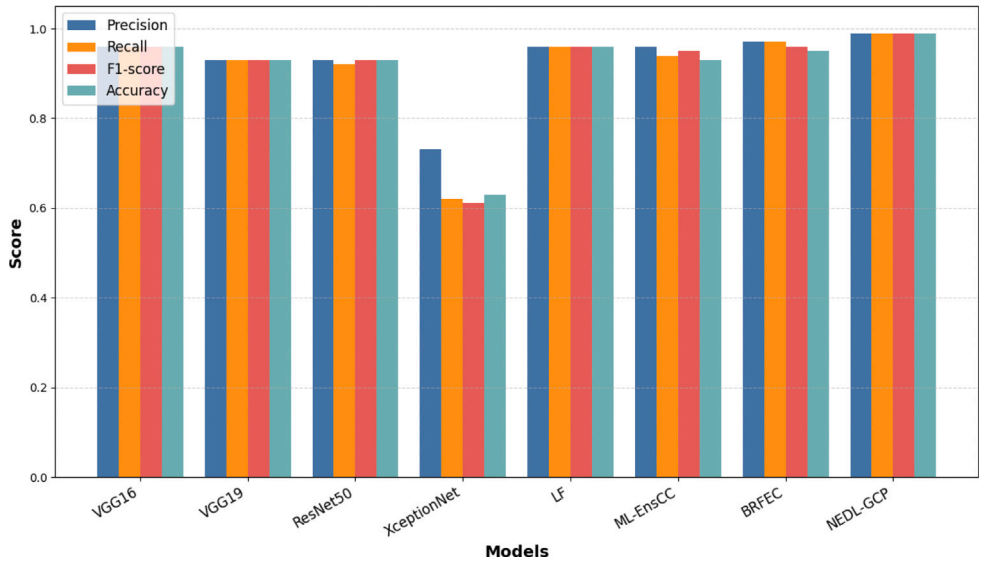


Fig. 8. Performance comparison of our proposed method with baseline models on the SIPaKMeD Pap Smear dataset in terms of precision, recall, F1-score, and accuracy.

a precision of 0.992, a recall of 0.990, an F1-score of 0.990, and an accuracy of 0.991, outperforming all of the state-of-the-art methods in all evaluation metrics.

VGG16 achieved a precision of 0.983, a recall of 0.981, an F1-score of 0.980, and an accuracy of 0.982. VGG19 achieved a precision of 0.966, a recall of 0.962, an F1-score of 0.964, and an accuracy of 0.964. ResNet50 achieved a precision of 0.964, a recall of 0.958, an F1-score of 0.960, and an accuracy of 0.960. XceptionNet achieved a precision of 0.751, a recall of 0.650, an F1-score of 0.639, and an accuracy of 0.657. LF achieved a precision of 0.986, a recall of 0.986, an F1-score of 0.986, and an accuracy of 0.986. ML-EnsCC achieved a precision of 0.96, a recall of 0.94, an F1-score of 0.95, and an accuracy of 0.93. BRFEC achieved a precision of 0.97, a recall of 0.97, an F1-score of 0.96, and an accuracy of 0.95.

Our proposed method achieved high precision, recall, and F1-score, demonstrating its effectiveness in accurately predicting gynecological cancer from Pap smear images. The achieved accuracy of 0.991 indicates the potential of our proposed method in clinical settings for screening and early detection of cervical cancer.

4.6. Statistical significance analysis

To validate the performance improvements of the proposed nested ensemble model, we conducted statistical significance testing using independent two-sample t-tests and computed 95% confidence intervals for accuracy scores. The tests compare NEDL-GCP against seven widely used models, including deep learning architectures (VGG-16, VGG-19, ResNet-50, XceptionNet), Late Fusion (LF), and ensemble-based methods ML-EnsCC and BRFEC, across the Herlev and SIPaKMeD Pap Smear datasets.

Table 4 presents the statistical significance results, including mean accuracy, confidence intervals, and p-values. The confidence intervals indicate the range within which the true accuracy values likely fall, while the p-values measure the statistical significance of the differences between NEDL-GCP and the baseline models.

The results indicate that NEDL-GCP consistently outperforms the baseline models on both datasets. The p-values for most comparisons are below 0.05, suggesting statistically significant improvements, particularly over VGG-16, VGG-19, ResNet-50, and XceptionNet. The Late Fusion classifier exhibits competitive performance in SIPaKMeD, but NEDL-GCP still achieves superior results. The inclusion of ML-EnsCC and BRFEC provides a fairer comparison with other ensemble-based methods. As shown in Table 4, while both models perform well, NEDL-GCP still achieves the highest accuracy with statistically significant improvements.

The confidence intervals further confirm the robustness of NEDL-GCP, as it consistently maintains a high accuracy range compared to the baseline models. These findings provide strong statistical evidence supporting the efficacy of the proposed nested ensemble approach for the prediction of cervical cancer risk prediction.

5. Discussion

Deep learning models have shown significant potential in medical applications, particularly for the prediction of cervical cancer risk. In this study, we introduce NEDL-GCP, a nested ensemble deep learning approach that integrates deep learning architectures with ensemble techniques to enhance prediction accuracy. Our model surpasses several state-of-the-art classifiers, achieving remarkable performance on the Herlev and SIPaKMeD Pap Smear datasets. Specifically, NEDL-GCP achieved F1 scores of 0.986 and 0.992, and accuracies of 0.991 and 0.985, respectively. To ensure a comprehensive evaluation, Table 5 presents a comparative analysis of NEDL-GCP against existing classification models, emphasizing its superiority in precision, accuracy, F1-score, and recall.

Our study introduces an advanced nested ensemble deep learning framework (NEDL-GCP) that significantly improves classification performance compared to traditional ensemble classifiers. Unlike conventional ensemble methods that only aggregate predictions from base models, our approach refines output through multiple learning layers, enhancing feature representation and robustness. Using deep learning architectures alongside decision-based classifiers, NEDL-GCP achieves superior predictive performance. The results in Table 5 demonstrate that our model outperforms both traditional ensemble techniques and individual deep learning classifiers for cervical cancer detection.

Despite these promising results, some limitations should be acknowledged. First, dataset bias remains a concern. Although the Herlev and SIPaKMeD Pap Smear datasets are well-structured benchmarks, they may not fully reflect real-world clinical variability between different populations, imaging conditions, and medical institutions. Future research should incorporate more diverse datasets to improve model generalizability.

Second, model overfitting is a potential issue due to the complexity of the nested ensemble architecture. Although cross-validation and regularization methods were applied, additional techniques such as

Table 4
Statistical significance analysis of model performance.

Model	Herlev dataset			SIPaKMeD dataset		
	Accuracy	CI (95%)	*p*-value	Accuracy	CI (95%)	*p*-value
NEDL-GCP	0.991	(0.981, 1.001)	–	0.985	(0.975, 0.995)	–
VGG-16	0.887	(0.877, 0.897)	0.0001	0.986	(0.976, 0.996)	0.2103
VGG-19	0.860	(0.850, 0.870)	0.00001	0.986	(0.976, 0.996)	0.1845
ResNet-50	0.838	(0.828, 0.848)	0.000001	0.960	(0.950, 0.970)	0.0034
XceptionNet	0.397	(0.387, 0.407)	0.000001	0.657	(0.647, 0.667)	0.000001
Late Fusion (LF)	0.860	(0.850, 0.870)	0.00001	0.960	(0.950, 0.970)	0.0029
ML-EnsCC	0.930	(0.920, 0.940)	0.0042	0.948	(0.938, 0.958)	0.0095
BRFEC	0.950	(0.940, 0.960)	0.0028	0.970	(0.960, 0.980)	0.0053

Table 5
Performance comparison of cervical cancer prediction models.

Dataset	Model	Precision	Accuracy	F1-Score	Recall
Herlev Dataset	**NEDL-GCP**	**0.989**	**0.985**	**0.986**	**0.982**
	LF	0.887	0.860	0.877	0.872
	XceptionNet	0.412	0.397	0.380	0.425
	ResNet-50	0.860	0.838	0.853	0.850
	ML-EnsCC	0.920	0.930	0.925	0.918
	BRFEC	0.945	0.950	0.948	0.943
SIPaKMeD Pap Smear	**NEDL-GCP**	**0.994**	**0.991**	**0.992**	**0.991**
	LF	0.986	0.986	0.986	0.986
	XceptionNet	0.751	0.657	0.639	0.650
	ResNet-50	0.964	0.960	0.960	0.958
	ML-EnsCC	0.950	0.948	0.949	0.947
	BRFEC	0.970	0.965	0.968	0.966

dropout layers, data augmentation, and semi-supervised learning could further enhance robustness in future studies.

Third, this study primarily focuses on cervical cancer prediction. The applicability of NEDL-GCP to other general or gynecological cancers remains unexplored. Future research should investigate whether this framework can be extended to broader oncological applications by validating its effectiveness in various cancer datasets.

In conclusion, NEDL-GCP demonstrates state-of-the-art performance in cervical cancer risk prediction, outperforming baseline classifiers in precision, accuracy, recall, and F1-score. Addressing the aforementioned limitations will further strengthen its applicability and reliability. Future work should focus on incorporating additional datasets, refining regularization techniques, and expanding the application of the model to other types of cancer, maximizing its clinical impact.

6. Conclusion

This study introduced a new ensemble deep learning framework for predicting cervical cancer risk, using a stacking-based model that integrates multiple neural network architectures. By effectively combining complementary features from different classifiers, our approach improves predictive accuracy and robustness. Our model outperformed conventional machine learning approaches, achieving an F1-score of 0.986 and an accuracy rate of 0.985, demonstrating state-of-the-art performance. Furthermore, the proposed method exhibited strong generalizability, maintaining high predictive accuracy when tested on an independent dataset. This robustness highlights its potential for real-world clinical applications, offering physicians a reliable tool for personalized risk assessment and treatment planning. By improving early detection, our approach can contribute to more precise interventions, ultimately enhancing patient outcomes. Future research should focus on extending this framework to other types of cancer to assess its broader applicability. Furthermore, integrating multimodal data sources, such as genomic, histopathological and clinical records, could further refine the accuracy of cancer prediction and improve personalized healthcare strategies.

CRediT authorship contribution statement

Kamal Berahmand: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Xujuan Zhou:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Yuefeng Li:** Writing – review & editing, Supervision, Software, Methodology, Investigation, Conceptualization. **Raj Gururajan:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Prabal Datta Barua:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **U Rajendra Acharya:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Srinivas Kondalsamy Chennakesavan:** Writing – review & editing, Validation, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We used two publicly available datasets to develop and evaluate our cervical cancer prediction model: the Herlev dataset and SIPaKMeD Pap-Smear dataset. any one can use it.

References

[1] Sharma R, Sharma S, Dhiman A, Singh SP, Thakur S, Garg VK, et al. Global epidemiological status of breast cancer burden: Potential reduction strategies and preventive measures. In: Cancer of the breast. Elsevier; 2025, p. 1–10.
[2] Gültekin M, Ramirez P, Broutet N, Hutubessy R. World health organization call for action to eliminate cervical cancer globally. Int J Gynecol Cancer 2020;30(4):426–7.
[3] Wilson LF, Antonsson A, Green AC, Jordan SJ, Kendall BJ, Nagle CM, et al. How many cancer cases and deaths are potentially preventable? Estimates for Australia in 2013. Int J Cancer 2018;142(4):691–701.
[4] Australian Institute of Health and Welfare. Cancer in Australia: Actual incidence data from 1982 to 2013 and mortality data from 1982 to 2014 with projections to 2017. Asia-Pac J Clin Oncol 2018;14(1):5–15.
[5] Harrison LA, Baum SH, Bhatara AK, Freedman EG, Lerner MD. Over-responsivity in autism spectrum disorder: A review of physiological reactivity domains and measures. In: International organization for autism research annual meeting. 2015.
[6] Australian Institute of Health and Cancer Australia, et al. Gynaecological cancers in Australia: An overview. 2012.
[7] Sankaranarayanan R, Ferlay J. Worldwide burden of gynaecological cancer: The size of the problem. Best Pr Res Clin Obs Gynaecol 2006;20(2):207–25.
[8] Hanzala A, Akter T, Rahman MS. A hybrid approach for cervical cancer detection: Combining D-CNN, transfer learning, and ensemble models. Array 2025;25:100434.
[9] Stumbar SE, Stevens M, Feld Z. Cervical cancer and its precursors: A preventative approach to screening, diagnosis, and management. Prim Care: Clin Off Pr 2019;46(1):117–34.
[10] Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: Epidemiology, risk factors and screening. Chin J Cancer Res 2020;32(6):720.

[11] Abotchie PN, Shokar NK. Cervical cancer screening among college students in Ghana: Knowledge and health beliefs. Int J Gynecol Cancer 2009;19(3):412–6.
[12] Craddock M, Crockett C, Mcwilliam A, Price G, Sperrin M, van der Veer S, et al. Evaluation of prognostic and predictive models in the oncology clinic. Clin Oncol 2022;34(2):102–13.
[13] Nejadshamsi S, Bentahar J, Eicker U, Wang C, Jamshidi F. A geographic-semantic context-aware urban commuting flow prediction model using graph neural network. Expert Syst Appl 2025;261:125534.
[14] Sharifi S. Stratified approaches for sustainable decision-making under uncertainty. Appl Soft Comput 2025;113239.
[15] Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. Brain Inform 2023;10(1):10.
[16] Vasheghani S, Sharifi S. Dynamic ensemble learning for robust image classification: A model-specific selection strategy. 2025, Available At SSRN 5215134.
[17] Abdollahi S, Deldari A, Asadi H, Montazerolghaem A, Mazinani SM. Flow-aware forwarding in SDN datacenters using a knapsack-PSO-based solution. IEEE Trans Netw Serv Manag 2021;18(3):2902–14.
[18] Zhou T, Ruan S, Canu S. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 2019;3:100004.
[19] Afroj M, Mondal MRH, Hassan MR, Akter S. MobDenseNet: A hybrid deep learning model for brain tumor classification using MRI. Array 2025;100413.
[20] Gopalakrishnan A, Gururajan R, Zhou X, Venkataraman R, Chan KC, Higgins N. A survey of autonomous monitoring systems in mental health. Wiley Interdiscip Rev: Data Min Knowl Discov 2024;14(3):e1527.
[21] Nejadshamsi S, Karami V, Ghourchian N, Armanfard N, Bergman H, Grad R, et al. Development and feasibility study of HOPE model for prediction of depression among older adults using Wi-Fi-based motion sensor data: Machine learning study. JMIR Aging 2025;8:e67715.
[22] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genom Proteom 2018;15(1):41–51.
[23] Sharifi S. Enhancing kidney transplantation through multi-agent kidney exchange programs: A comprehensive review and optimization models. 2025, arXiv preprint arXiv:2502.07819.
[24] Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustainability 2021;13(3):1224.
[25] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[26] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.
[27] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput System Sci 1997;55(1):119–39.
[28] Rabby SF, Arafat MA, Hasan T. BT-Net: An end-to-end multi-task architecture for brain tumor classification, segmentation, and localization from MRI images. Array 2024;22:100346.
[29] Rojas MG, Olivera AC, Vidal PJ. Optimising multilayer perceptron weights and biases through a cellular genetic algorithm for medical data classification. Array 2022;14:100173.
[30] Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, et al. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit Lett 2020;132:123–31.
[31] Kamalov F, Sulieman H, Moussa S, Reyes JA, Safaraliev M. Nested ensemble selection: An effective hybrid feature selection method. Heliyon 2023;9(9).
[32] Kaushik M, Joshi RC, Kushwah AS, Gupta MK, Banerjee M, Burget R, et al. Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: A machine learning approach. Comput Biol Med 2021;134:104559.
[33] Singh S, Tejaswini V, Murthy RP, Mutgi A. Neural network based automated system for diagnosis of cervical cancer. In: Deep learning and neural networks: Concepts, methodologies, tools, and applications. IGI Global; 2020, p. 1422–36.
[34] Li M, Feng A, Yan Y, You S, Li C. Deep convolutional neural network based cervical cancer exfoliated cell detection. In: International conference on image, vision and intelligent systems. Springer; 2022, p. 589–98.
[35] Li C, Xue D, Zhou X, Zhang J, Zhang H, Yao Y, et al. Transfer learning based classification of cervical cancer immunohistochemistry images. In: Proceedings of the third international symposium on image computing and digital medicine. 2019, p. 102–6.
[36] Lu J, Song E, Ghoneim A, Alrashoud M. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Gener Comput Syst 2020;106:199–205.
[37] Curia F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. Heal Technol 2021;11(4):875–85.
[38] Chandran V, Sumithra M, Karthick A, George T, Deivakani M, Elakkiya B, et al. Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. BioMed Res Int 2021;2021(1):5584004.
[39] Ali MS, Hossain MM, Kona MA, Nowrin KR, Islam MK. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. Heal Anal 2024;5:100324.
[40] Aljrees T. Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. PLoS One 2024;19(1):e0295632.
[41] Uddin KMM, Al Mamun A, Chakrabarti A, Mostafiz R, Dey SK. An ensemble machine learning-based approach to predict cervical cancer using hybrid feature selection. Neurosci Inform 2024;4(3):100169.
[42] Kwatra CV, Kaur H, Potharaju S, Tambe SN, Jadhav DB, Tambe SB. Harnessing ensemble deep learning models for precise detection of gynaecological cancers. Clin Epidemiol Glob Heal 2025;101956.
[43] Jantzen J, Norup J, Dounias G, Bjerregaard B. Pap-smear benchmark data for pattern classification. In: Nature inspired smart information systems. 2005, p. 1–9.
[44] Plissiti ME, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A. SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: 2018 25th IEEE international conference on image processing. ICIP, IEEE; 2018, p. 3144–8.
[45] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.
[46] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.
[47] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1251–8.
[48] Benzebouchi NE, Azizi N, Ashour AS, Dey N, Sherratt RS. Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis. J Exp Theor Artif Intell 2019;31(6):841–74.
[49] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manage 2009;45(4):427–37.
[50] Powers DM. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020, arXiv preprint arXiv:2010.16061.