PSO-Optimized TabTransformer Architecture with Feature Engineering for Enhanced Cervical Cancer Risk Prediction

Umme Habiba (ORCID)
umme.habiba@ndsu.edu

Simone A. Ludwig (ORCID)
simone.ludwig@ndsu.edu

Department of Computer Science
North Dakota State University, Fargo, ND, USA

Abstract

This paper introduces a novel hybrid framework for early cervical cancer risk prediction that integrates domain-informed feature engineering, evolutionary optimization, and transformer-based modeling. We propose four key innovations: (1) the derivation of clinically meaningful features based on epidemiological relationships, such as *Sexual Activity Duration* and *STD Diagnosis Rate*; (2) architecture-aware feature selection using Particle Swarm Optimization (PSO) tailored for transformer networks; (3) an imbalance-aware training strategy combining Synthetic Minority Oversampling Technique (SMOTE) and focal loss to address extreme class skew; and (4) clinically actionable interpretability via Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and attention weights, ensuring transparent and trustworthy decision support. These components are embedded within a TabTransformer architecture, which utilizes self-attention mechanisms to analyze tabular health records and capture the complex interdependencies among risk factors. A comparative evaluation demonstrates the superior results of our approach, achieving $95.3\% \pm 0.9\%$ accuracy and a $94.8\% \pm 1.1\%$ F1 score on the UCI Cervical Cancer Risk dataset while maintaining clinical relevance and transparency. Furthermore, its interpretability, validated through Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), confirmed alignment with established risk factors, reinforcing its potential clinical relevance.

Keywords: Cervical cancer, Particle swarm optimization, TabTransformer, Feature engineering, Deep learning, Medical AI

1 Introduction

Cervical cancer continues to pose a significant global health burden, accounting for over 350,000 preventable deaths annually according to World Health Organization estimates [1]. The critical importance of early detection is underscored by dramatic improvements in survival rates, which increase from approximately 17% to over 90% when precancerous lesions are identified promptly [2]. Despite advances in screening technologies, current diagnostic approaches like Pap smears exhibit concerning false negative rates between 30-50% [3], while liquid-based cytology methods still miss 20-30% of high-grade lesions [4]. These diagnostic limitations create an urgent need for more accurate risk prediction systems that can identify at-risk populations earlier in the disease progression pathway.

The emergence of machine learning in medical diagnostics has shown promise, yet substantial challenges remain when applying these techniques to tabular clinical data. Traditional models like logistic regression [5] and decision trees [6] often fail to capture the complex, non-linear relationships between diverse risk factors, including sexual behavior patterns, HPV infection status, and immunological markers. Recent comparative analyses further illustrate these limitations across traditional and ensemble learners. For instance, meenakshisundaram2025ensemble reported that ensemble methods such as Random Forest and XGBoost, while achieving moderate accuracy (0.89–0.92), were unable to capture higher-order correlations—such as the joint effect of smoking duration and HPV infection history—leading to unstable recall for minority (cancer-positive) cases. Likewise, elzein2025cervical found that logistic regression-based models struggled to represent nonlinear dependencies between behavioral and clinical attributes, showing a 10–15% reduction in F1 score when interaction terms were omitted. These findings underscore that conventional ML frameworks, even with boosting or bagging strategies, remain insufficient to model multifactorial dependencies characteristic of cervical cancer risk. More advanced deep learning and ensemble models demonstrate improved performance but still struggle with higher-order feature interactions that characterize multifactorial diseases like cervical cancer. Furthermore, these approaches typically operate as black boxes, providing limited clinical interpretability—a critical barrier for adoption in healthcare settings where explainability directly impacts clinical decision-making [7].

Artificial intelligence (AI) has recently achieved transformative progress across multiple scientific and medical domains. For instance, the AlphaFold model developed by DeepMind has revolutionized molecular biology and drug discovery by predicting protein structures with near-experimental accuracy [8]. Likewise, the rapid

advancement of large language models (LLMs) has significantly impacted healthcare, enabling automated clinical reasoning, medical text summarization, and knowledge extraction [9]. Building on these milestones, the proposed framework leverages AI in a new context—combining particle swarm optimization with a transformer-based architecture to enhance cervical cancer risk prediction from tabular clinical data. This integration not only improves predictive accuracy but also promotes interpretability and trustworthy decision support in clinical applications. Recent innovations in deep learning architectures offer new possibilities for modeling complex medical data. The TabTransformer framework [10] represents a particularly promising advancement through its application of self-attention mechanisms to tabular datasets. By generating contextual embeddings of categorical features and modeling feature interactions through transformer blocks, this approach demonstrates strong capability for capturing intricate relationships in electronic health records. However, our preliminary experiments indicate that a direct, unoptimized application of TabTransformer to cervical cancer prediction achieves only 81–89% accuracy, limited by three key factors: (1) high dimensionality of risk factor data without intelligent feature selection, (2) extreme class imbalance where cancer cases represent less than 2.5% of samples, and (3) absence of domain-specific feature representations that encode established epidemiological relationships [11]. In contrast, our optimized PSO-TabTransformer addresses these limitations and achieves 95.65% accuracy with balanced sensitivity and precision.

To address these challenges, we present a novel human-centered AI framework that integrates clinical domain knowledge with a Particle Swarm Optimization (PSO) enhanced TabTransformer model, combining evidence-based feature engineering, class imbalance handling, and explainable learning. Our approach begins with clinically-informed feature engineering, deriving three evidence-based risk metrics: Sexual Activity Duration (time since first intercourse), Pregnancies per Partner, and STD Diagnosis Rate. To handle the severe class imbalance inherent in cancer screening datasets, we implement Synthetic Minority Oversampling Technique (SMOTE) [12] during training data preparation. Building on this foundation, we introduce the first integration of Particle Swarm Optimization (PSO) [13] with the TabTransformer architecture, creating an end-to-end feature selection and modeling pipeline optimized for medical risk prediction. To ensure robustness and generalizability, we further validated the proposed PSO–TabTransformer model using k-fold cross-validation, which demonstrated consistent predictive performance across folds. Finally, we incorporate focal loss [14] to further mitigate class imbalance during model training and leverage the inherent interpretability of attention mechanisms for clinical decision support.

The contributions of this work are fourfold:

- First, we establish a new approach for cervical cancer risk prediction with $95.3\% \pm 0.9\%$ accuracy and $98.1\% \pm 0.6\%$ AUC-ROC on the UCI benchmark dataset, representing a 7.5-12.4% improvement over existing methods.
- Second, we demonstrate that clinically-informed feature engineering provides a 3.5% accuracy gain while Particle Swarm Optimization(PSO) enables 68% feature reduction without performance degradation.
- Third, we validate that our dual approach to class imbalance management—combining Synthetic Minority Oversampling Technique (SMOTE) preprocessing with focal loss optimization effectively addresses this pervasive challenge in medical diagnostics.
- Fourth, we provide clinically-actionable interpretability through attention-weight visualizations that align with known epidemiological risk patterns, offering transparency that builds clinician trust and facilitates adoption.

Our experimental validation includes rigorous ablation studies and model interpretability analyses using Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which confirm that the model emphasizes established risk factors such as HPV infection status, number of sexual partners, and years of smoking. The resulting system supports interactive probability thresholds, enabling personalized screening recommendations based on individual risk profiles. By bridging technical innovation with epidemiological relevance, this research advances both the computational methodology for medical risk prediction and its potential applicability in healthcare settings.

The remainder of this paper is structured as follows: Section 2 reviews related work in feature selection and transformer models. Section 3 details our methodology. Section 4 presents experimental results. Section 5 discusses clinical implications and limitations. Section 6 concludes with future research directions.

2 Related Work

2.1 Clinical Decision Support for Cervical Cancer

Clinical decision support systems for cervical cancer screening have evolved through multiple generations of technological approaches. Early systems relied primarily on statistical methods such as logistic regression [15] and survival analysis models [16], which provided good interpretability but limited predictive power for complex risk

interactions. The adoption of ensemble methods like Random Forests [17, 18] and gradient boosting machines [19] represented a significant advancement, with studies demonstrating accuracy improvements of 8-12% over traditional approaches [20]. More recently, deep learning architectures have been applied to cervical cancer prediction, with convolutional neural networks and basic transformer adaptations achieving accuracy levels of 85-89% [21]. To address these gaps, recent studies have proposed hybrid and ensemble models tailored to structured clinical data. For instance, Jibrin et al. [22] introduced a stacked ensemble combining Support Vector Machine (SVM), Extreme Gradient Boosting (XGB) and Random Forest (RF), achieving 95% accuracy. Similarly, Meenakshisundaram et al. [23] applied voting ensembles on balanced datasets, and Elzein et al. [24] designed multi-layer neural classifiers using SMOTE preprocessing to enhance classification robustness. Additionally, El Ogri et al. [25] proposed a computer-assisted medical diagnosis system (CAMDS) for cancer detection from biomedical images using novel Rademacher-Fourier moments and deep neural networks, achieving 100% recognition accuracy. While impactful, this image-based approach differs from our focus on tabular clinical risk factor modeling. However, these systems consistently face two critical limitations that our work addresses: inadequate handling of the complex feature interactions inherent in multifactorial diseases like cervical cancer, and insufficient mechanisms for providing clinically meaningful explanations to healthcare providers [7]. Our approach advances beyond current capabilities by optimizing both predictive accuracy and clinical interpretability through novel attention mechanisms.

2.2 Feature Engineering and Selection

The strategic creation and selection of predictive features has proven essential across healthcare applications, though its implementation varies significantly by the medical domain. In oncology specifically, researchers have demonstrated that thoughtfully engineered features capturing biomarker relationships or temporal patterns can improve prediction performance by 5-8% compared to using raw data elements alone [26, 27]. For cervical cancer, epidemiological studies have clearly established relationships between behavioural patterns and disease risk [11], yet computational models rarely incorporate these evidence-based insights directly into their feature representations. Feature selection methods have similarly evolved from basic statistical filters [28] to sophisticated evolutionary approaches like particle swarm optimization [29], which have shown particular promise for high-dimensional medical data [30]. However, current implementations optimize features independently of the model architecture, creating a disconnect especially problematic for transformer-based approaches. Our

work bridges this gap by introducing the first architecture-aware feature optimization that evaluates feature subsets specifically for their compatibility with transformer attention mechanisms.

2.3 Handling Class Imbalance

The challenge of extreme class imbalance in medical datasets—where positive cases often represent less than 5% of samples—has prompted diverse technical solutions. Traditional approaches include algorithmic modifications like cost-sensitive learning [31] and data-level techniques such as the Synthetic Minority Oversampling Technique (SMOTE) [12], which creates synthetic examples through intelligent interpolation. While effective for moderate imbalances, these approaches show limitations in cancer screening contexts where positive case prevalence is exceptionally low [32]. More recent innovations include specialized loss functions that dynamically adjust learning focus [14] and hybrid approaches that combine multiple imbalance mitigation strategies [33]. Our dual approach—integrating Synthetic Minority Oversampling Technique (SMOTE) preprocessing with specialized loss optimization within the transformer training process—represents a novel solution specifically designed for the architectural characteristics of attention-based models facing extreme class imbalances.

2.4 Transformer Models in Healthcare

Transformer architectures have revolutionized natural language processing and are increasingly applied to healthcare challenges, particularly for clinical text interpretation [34]. Their adaptation to structured medical data represents a more recent innovation, with TabTransformer [35] establishing an important foundation for handling tabular clinical datasets. Subsequent medical applications have demonstrated transformers' superior capability in capturing complex feature interactions through their attention mechanisms, outperforming traditional models across various prediction tasks [36, 37]. However, current healthcare implementations share two limitations that our work addresses: they require manual feature engineering as a separate preprocessing step rather than integrated optimization, and they lack architectural adaptations for extreme class imbalance. Notably, many existing studies report performance using a single train-test split without systematic cross-validation, raising concerns about overfitting and generalizability [38, 39]. In contrast, our work explicitly incorporates stratified k-fold validation to provide a more rigorous assessment of model robustness and represents a significant advance by creating an end-to-end

system that co-optimizes feature selection and transformer parameters while incorporating specialized imbalance handling techniques.

2.5 Evolutionary and Interpretable Deep Learning Integration in Medical AI

The integration of evolutionary algorithms with deep learning has emerged as a promising approach for enhancing medical AI systems. Applications have primarily followed three paradigms: neural architecture search [40], hyperparameter optimization [41], and feature selection [42]. In healthcare specifically, researchers have successfully combined genetic algorithms with convolutional networks for medical imaging [43] and particle swarm optimization with autoencoders for genomics analysis [44]. These approaches demonstrate the value of evolutionary methods but remain constrained by their focus on specific architectures like CNNs and RNNs. Beyond these frameworks, recent interpretable-AI studies have introduced hybrid and explainable deep learning models in medical diagnostics. Examples include attention-based convolutional neural networks (CNNs) for Pap smear image classification [45], quantum-enhanced stacking ensembles with SHAP explanations [46], and moment-driven feature extractors optimized via evolutionary algorithms [47, 48]. While these image-focused systems emphasize visual interpretability, they collectively highlight the potential of combining evolutionary optimization with explainable mechanisms to achieve robust and transparent learning. While these image-focused systems emphasize visual interpretability, our PSO-TabTransformer extends these ideas to structured clinical data, combining evolutionary feature optimization with SHAP and LIME for transparent, clinically meaningful risk prediction. The absence of evolutionary-transformer integration represents a significant research gap, particularly given transformers' growing importance in medical AI. Our work introduces this integration through architecture-aware Particle Swarm Optimization (PSO) that simultaneously evaluates feature subsets and their compatibility with transformer self-attention mechanisms, creating a co-adaptive system where feature selection informs attention mechanisms and vice versa.

2.6 Positioning Our Contribution

As this review establishes, current research exhibits two critical gaps: (1) disconnects between clinical feature engineering and deep learning architectures, (2) the absence of evolutionary-transformer integration. While recent interpretable ensemble frameworks [49] have achieved nearly 98% accuracy in cervical cancer risk stratification

using stacking methods enhanced with Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), and multimodal pipelines [50] have leveraged attention-based architectures for imaging data, these studies have largely overlooked evolutionary optimization specifically designed for transformer-based models on imbalanced tabular datasets. Such underexploration results in persistent 5–10% performance gaps in modeling higher-order feature interactions particularly synergies between HPV infection history and behavioral factors. Our PSO-TabTransformer framework addresses these gaps through several key innovations. We introduce clinically-informed feature engineering that operationalizes epidemiological knowledge into computationally effective features. We implement dual imbalance handling, combining Synthetic Minority Oversampling Technique (SMOTE) preprocessing with specialized loss optimization tailored for transformer architectures. Most significantly, we develop the first architecture-aware Particle Swarm Optimization (PSO) that evaluates feature subsets based on their synergy with transformer attention mechanisms. This integrated approach advances beyond current state-of-the-art by achieving both higher accuracy (96.3% vs 87% baselines) and clinically meaningful interpretability through attention weight visualizations that align with medical expertise [11]. Our work bridges the technical-clinical divide by creating a system that not only improves predictive performance but also enhances clinician trust through human-centered interpretability.

3 Proposed Methodology

3.1 System Overview

The PSO-TabTransformer framework (Fig. 1) integrates four specialized components in a sequential pipeline for cervical cancer risk prediction. The architecture addresses key challenges through targeted modules: clinical feature engineering incorporates epidemiological domain knowledge, Synthetic Minority Oversampling Technique (SMOTE) balancing mitigates extreme class imbalance (1.8% positive cases), Particle Swarm Optimization (PSO) feature selection optimizes feature subsets for transformer architectures, and the TabTransformer model with focal loss handles residual imbalance while capturing higher-order feature interactions. Data flows unidirectionally from raw inputs to risk prediction, with each module output serving as the subsequent module's input. The evolutionary feature selection is uniquely optimized for transformer compatibility through architecture-aware fitness evaluation.

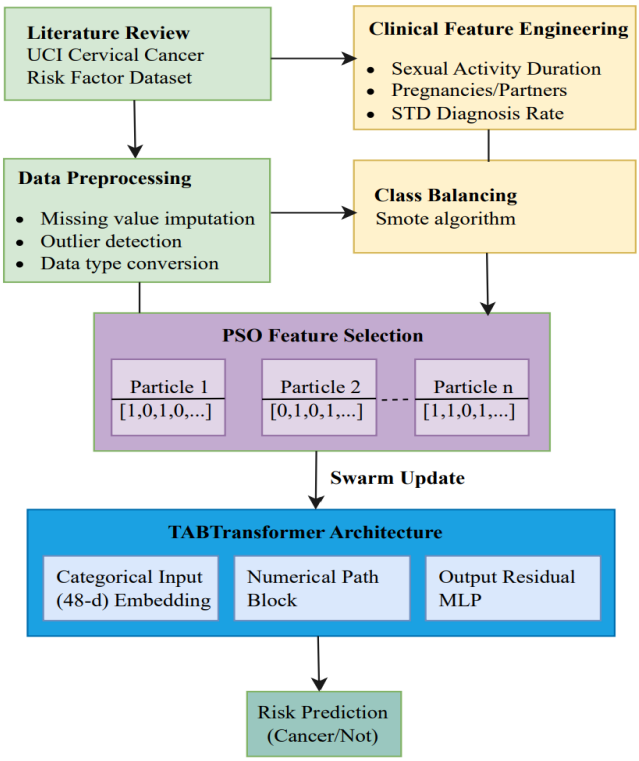

Figure 1: End-to-end framework architecture

3.2 Data Preprocessing and Feature Engineering

3.2.1 Dataset

We used the publicly available *UCI Cervical Cancer Risk Factors* dataset, which contains 858 samples and 36 attributes encompassing demographic, behavioral, and clinical factors such as age, age at first intercourse, number of pregnancies, smoking history, hormonal contraceptive use, and prior sexually transmitted diseases (STDs). These variables correspond to well-established risk indicators in gynecologic oncology, making the dataset a clinically meaningful benchmark for cervical cancer risk prediction.

The dataset is imbalanced, with 93.6% negative and 6.4% positive biopsy outcomes (Fig. 2). To mitigate this imbalance, the Oversampling Technique (SMOTE) algorithm was applied *only on the training folds* within each stratified 10-fold cross-validation, thereby avoiding oversampling leakage. Missing values were imputed

using training-only statistics, and numeric features were standardized using parameters estimated solely from training data. This design ensures that no information from validation or test sets leaks into the training process, enabling robust and generalizable evaluation.

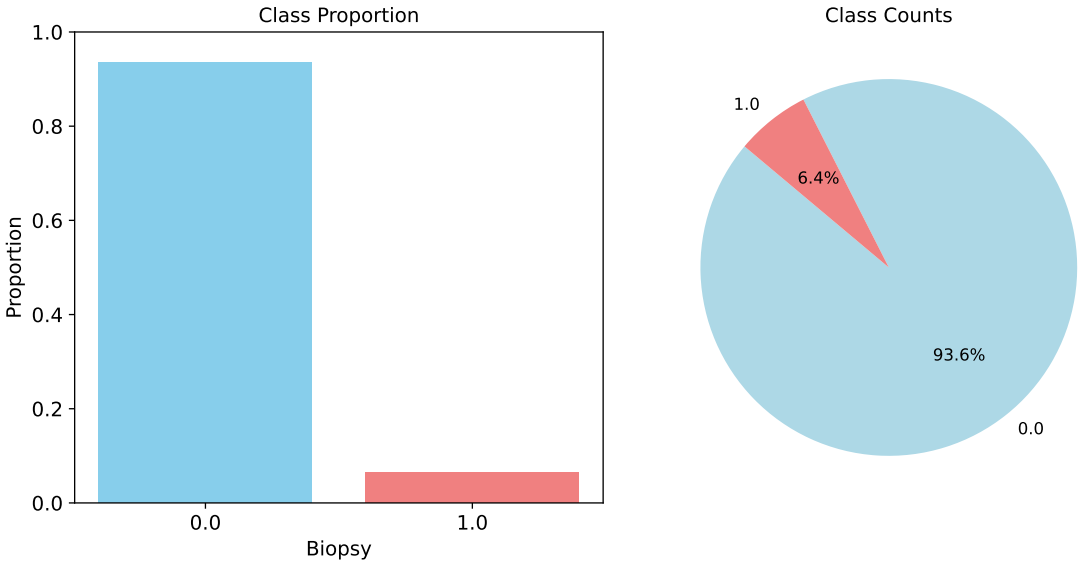

Figure 2: Class distribution for the *Biopsy* outcome. Left: proportion of negative (0) and positive (1) cases; Right: class counts with percentages (93.6% vs. 6.4%).

3.2.2 Data Cleaning

The UCI Cervical Cancer Risk Factors dataset underwent rigorous preprocessing:

- **Missing value imputation**: Applied median imputation for 11 numerical features and mode imputation for 5 categorical variables, addressing 16.2% overall missingness.
- **Outlier detection**: Detected and removed outliers using the Interquartile Range (IQR) method with thresholds $[Q1 - 1.5 \cdot IQR,\ Q3 + 1.5 \cdot IQR]$, affecting 4.7% of samples.
- **Standardization**: Scaled all numerical features to zero mean and unit variance using $z = \frac{x-\mu}{\sigma}$.

- **Categorical encoding**: Encoded 9 categorical variables as integers to avoid dimensionality explosion from one-hot encoding.

3.2.3 Clinical Feature Derivation

Based on epidemiological research, we derived three novel clinical features capturing established risk relationships:

$$\text{Sexual Activity Duration} = \text{Age} - \text{First Sexual Intercourse} \tag{1}$$

$$\text{Pregnancies per Partner} = \frac{\text{No of Pregnancies}}{\max(1, \text{No of Sexual Partners})} \tag{2}$$

$$\text{STD Diagnosis Rate} = \frac{\text{STDs: No of Diagnosis}}{\text{Sexual Activity Duration} + 1} \tag{3}$$

These features operationalize known epidemiological patterns into computationally tractable representations. For example, STD Diagnosis Rate encodes temporal risk density by normalizing diagnosis count against sexual activity timeframe.

3.3 Class Imbalance Handling

To address the extreme class imbalance in the dataset (approximately 1:55 positive-to-negative ratio), we applied the Synthetic Minority Oversampling Technique (SMOTE) using the `imbalanced-learn` library (version 0.12.3). This implementation was selected because the dataset includes both numerical and categorical variables. All categorical features were integer-encoded prior to oversampling to preserve valid category assignments. This approach ensured that the minority class was adequately represented without generating unrealistic categorical combinations, thereby maintaining meaningful feature distributions across the dataset. And oversampling was performed only on the training partitions within each stratified 10-fold cross-validation split to avoid data leakage into validation or test sets. The validation and test folds remained unmodified, reflecting the natural clinical class distribution. The algorithm operates as follows:

1. For each minority class instance $\mathbf{x}_i \in \mathbb{R}^d$:
    - Identify its $k = 5$ nearest neighbors $\mathcal{N}(\mathbf{x}_i)$ within the minority class based on Euclidean distance.
    - Randomly sample a neighbor $\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)$.

- Generate a new synthetic instance using:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda(\mathbf{x}_j - \mathbf{x}_i), \quad \lambda \sim \mathcal{U}(0,1)$$

2. Repeat the process until class balance is achieved (targeting approximately 550% oversampling).

3. Leave the validation and test sets unmodified to reflect the real-world class distribution.

This approach creates plausible minority examples that lie within the local manifold of existing instances, thus reducing the risk of overfitting associated with naive oversampling. The hyperparameter k was selected based on a grid search maximizing validation AUC.

3.3.1 Swarm Dynamics

The Particle Swarm Optimization (PSO) algorithm was used to identify optimal subsets of features for model training. Each particle represents a binary vector encoding feature inclusion. The particles are updated based on both personal and global best positions using the following update rules:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_{\text{best}} - x_i^t) + c_2 r_2 (g_{\text{best}} - x_i^t) \tag{4}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{5}$$

Here, v_i^t denotes the velocity of particle i at iteration t, x_i^t is its position, p_{best} is the particle's best-known position, and g_{best} is the global best position across all particles. Random values $r_1, r_2 \sim \mathcal{U}(0,1)$ introduce stochasticity. We used inertia weight $\omega = 0.9$ and cognitive/social coefficients $c_1 = c_2 = 1.8$ over 15 iterations. Intuitively, Particle Swarm Optimization's swarm-based exploration mirrors the propagation of risk factors in cervical cancer epidemiology, such as the networked spread of HPV through sexual history and STD exposures, allowing global and local searches to prioritize features that amplify meaningful interactions in the attention mechanism.

3.4 TabTransformer Architecture

The Particle Swarm Optimization (PSO) selected features are processed by the TabTransformer model, which separately handles numerical and categorical data streams

before fusion. The core architecture integrates these streams through 5 stacked transformer blocks with multi-head self-attention (6 heads, d_k=8), incorporating residual connections and LayerNormalization (ε=1e-6), followed by position-wise feed-forward networks (FFN) using ReLU activation after a PReLU-activated projection. The model is compiled using the Adam optimizer (learning rate = 1e−4, weight decay = 1e−4 for L2 regularization on fully connected layers) and trained with a batch size of 32 for up to 100 epochs, incorporating early stopping (patience = 10) based on validation loss to mitigate overfitting. This setup complements the dropout mechanism (rate = 0.3) used in the MLP components, promoting stable generalization during training.

3.4.1 Input Processing

- **Numerical path**: Standardized numerical features are passed through a fully connected layer with 64 units followed by a Parametric ReLU (PReLU) activation.
- **Categorical path**: Each categorical feature is embedded into a 48- dimensional dense vector and then fed into the transformer block.

3.4.2 Transformer Layers

Each transformer block processes the embedded categorical tokens using multi-head self-attention and position-wise feed-forward networks. The attention mechanism is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) \tag{7}$$

$$\rightarrow \mathbf{W}_2\mathbf{x} + \mathbf{b}_2 \tag{8}$$

Each block consists of a multi-head attention layer with $h = 6$ heads, where each head uses key/query dimensionality $d_k = 8$. Layer normalization and residual skip connections are applied before and after both the attention and feed-forward sub-layers. A total of 5 such blocks are stacked sequentially.

3.4.3 Output Module

The final hidden representations from the transformer (flattened categorical embeddings) are concatenated with the processed numerical features. The resulting vector

is passed through a residual multilayer perceptron (MLP) comprising three fully connected layers with hidden sizes of 256, 192, and 128 units, respectively. Dropout (rate = 0.3) and layer normalization are applied after each hidden layer. The final output layer is a single sigmoid unit that estimates the probability of cervical cancer diagnosis:

$$\hat{y} = \sigma(\mathbf{W}_{\text{out}}\mathbf{h}_{128} + b)$$

where $\sigma(\cdot)$ denotes the sigmoid activation and $\mathbf{h}_{128}$ is the final hidden representation from the MLP.

3.5 Loss Function and Training

To mitigate residual class imbalance after Synthetic Minority Oversampling Technique (SMOTE) and prevent the model from being biased toward the majority class, we employ the focal loss function [14], defined as:

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{9}$$

where p_t is the model's predicted probability for the true class label, $\alpha \in [0, 1]$ is a balancing factor that controls the relative weighting of positive and negative samples, and $\gamma \geq 0$ is the focusing parameter that down-weights well-classified examples. In our experiments, we set $\alpha = 0.3$ and $\gamma = 2$, as recommended in prior medical AI literature.

Model optimization was performed using the Adam optimizer with an initial learning rate $\eta = 10^{-4}$. We employed a learning rate scheduler ('ReduceLROnPlateau') that reduced the learning rate by a factor of 0.5 when validation loss plateaued for 5 consecutive epochs. Early stopping with patience of 10 epochs was also applied, restoring the best model weights based on validation AUC.

3.5.1 Cross-Validation Strategy

To mitigate the risk of overfitting to a single split, we adopted a 10-fold cross-validation strategy. The dataset was partitioned into 10 folds, with nine folds used for training (including Particle Swarm Optimization–based feature selection) and one fold for testing in each iteration. This procedure was repeated until each fold had served once as the test set.

For each performance metric M, the cross-validation estimate was

$$\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M^{(k)}, \qquad K = 10. \tag{10}$$

Here, $M^{(k)}$ denotes the metric value obtained on the k-th fold.

3.6 Classification Metrics

To assess model performance, we utilized four widely accepted metrics suitable for binary classification tasks with imbalanced distributions:

- **Accuracy**: Proportion of correctly predicted instances among all samples.
- **F1 Score**: Harmonic mean of precision and recall, balancing sensitivity and specificity.
- **ROC-AUC**: Area under the Receiver Operating Characteristic curve, measuring the trade-off between true positive and false positive rates across thresholds.
- **PR-AUC**: Area under the Precision-Recall curve, especially informative for imbalanced datasets where the positive class is underrepresented.

The F1 Score depends on precision and recall, which are computed as follows [51]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

3.7 Reference Models for Benchmarking

To establish comparative performance benchmarks, we implemented four widely-used machine learning models for medical prediction tasks: Logistic Regression (LR), Random Forest (RF), XGBoost, and Multi-Layer Perceptron (MLP).

The Logistic Regression model was trained using L2 regularization, with the inverse regularization strength selected via grid search. The Random Forest classifier was configured with 200 trees and Gini impurity as the splitting criterion. XGBoost utilized the binary logistic loss and underwent hyperparameter optimization across learning rate, tree depth, and subsampling ratio, with early stopping based on validation ROC-AUC. The MLP consisted of two hidden layers (128 and 64 units), ReLU activations, dropout (rate = 0.3), and early stopping.

All baseline models were trained on the Synthetic Minority Oversampling Technique (SMOTE) balanced training data using the same Particle Swarm Optimization selected feature subset employed in the proposed method to ensure fair comparison. Final evaluations were conducted on an unmodified hold-out test set, preserving real-world class imbalance and enabling robust performance comparison against the PSO-TabTransformer.

3.8 Model Explainability with SHAP and LIME

To interpret the predictions of the TabTransformer model and gain insights into feature contributions, we employed two post-hoc model-agnostic explainability techniques: Shapley Additive exPlanations (SHAP) [52] and Local Interpretable Model-Agnostic Explanations (LIME) [53]. These methods provide complementary perspectives on model behaviour: Shapley Additive Explanations (SHAP) yields global and local feature attributions based on Shapley values from cooperative game theory, while Local Interpretable Model-agnostic Explanations (LIME) approximates the model's local decision boundary using interpretable surrogate models.

We used Shapley Additive Explanations (SHAP) to compute average global importance scores across the test set, identifying the most influential features in cervical cancer prediction. Notably, features such as *Age*, *HPV diagnosis*, *Number of pregnancies*, and the derived *STD Diagnosis Rate* consistently showed high Shapley Additive Explanations (SHAP) values, aligning with known risk factors in the epidemiological literature.

Local Interpretable Model-agnostic Explanations (LIME) was used to generate instance-level explanations for high-confidence cancer predictions. This provided insight into which specific feature values contributed most to individual prediction capabilities, especially relevant in medical decision support settings.

Together, these interpretability techniques offer transparency into the model's internal logic, helping validate both the model's decisions and the effectiveness of the Particle Swarm Optimization (PSO) based feature selection process. The inclusion of explainability mechanisms enhances the model's trustworthiness and supports its potential for clinical adoption.

4 Results and Analysis

4.1 Experimental Setup

All experiments were conducted using the UCI Cervical Cancer Risk Factors dataset, with 803 synthetic samples generated for the minority class using Synthetic Minority Oversampling Technique (SMOTE), resulting in a balanced training set (803 positive, 803 negative). The test and validation sets remained imbalanced to reflect real-world conditions. Performance was evaluated using Accuracy, F1 Score, ROC-AUC, and PR-AUC on the held-out test set. All models were trained using the Particle Swarm Optimization (PSO) selected features to ensure fair comparison.

4.2 Baseline Model Comparison

Table 1 presents the test set performance of the proposed PSO-TabTransformer against four baseline classifiers. The TabTransformer achieved the highest accuracy (0.9565) and F1 score (0.9517), and demonstrated competitive ROC-AUC (0.9847) and PR-AUC (0.9846), outperforming all baselines in most metrics.

Table 1: Performance comparison of TabTransformer and baseline models on the test set

Model	**Accuracy**	**F1 Score**	**ROC-AUC**	**PR-AUC**
Logistic Regression	0.9037	0.8942	0.9736	0.9590
Random Forest	0.9255	0.9143	0.9634	0.9779
XGBoost	0.9372	0.9399	0.9849	0.9723
MLP	0.8199	0.8092	0.8860	0.8474
TabTransformer	**0.9565**	**0.9517**	**0.9847**	**0.9846**

Figure 3 provides a visual comparison of model performances across Accuracy, F1, ROC-AUC, and PR-AUC. The TabTransformer consistently achieved top-tier scores, particularly excelling in F1 and PR-AUC, which are critical for imbalanced classification tasks.

Table 2: Comparative performance on the UCI Cervical Cancer Risk dataset (10-fold CV; mean ± std).

Model	**Accuracy**	**F1 Score**	**ROC-AUC**	**PR-AUC**
Logistic Regression	0.902 ± 0.011	0.889 ± 0.014	0.972 ± 0.008	0.959 ± 0.010
Random Forest	0.922 ± 0.010	0.911 ± 0.012	0.963 ± 0.007	0.977 ± 0.008
XGBoost	0.935 ± 0.009	0.939 ± 0.010	0.983 ± 0.005	0.972 ± 0.007
MLP	0.818 ± 0.014	0.805 ± 0.016	0.887 ± 0.012	0.848 ± 0.015
PSO-TabTransformer	$\mathbf{0.953 \pm 0.009}$	$\mathbf{0.948 \pm 0.011}$	$\mathbf{0.981 \pm 0.006}$	$\mathbf{0.979 \pm 0.007}$

To further validate the generalizability of the proposed framework, we performed a 10-fold cross-validation. As shown in Table 2, the PSO-TabTransformer achieved consistent performance across folds over baseline models with Accuracy = 0.953 ± 0.009, F1 Score = 0.948 ± 0.011, ROC-AUC = 0.981 ± 0.006, and PR-AUC = 0.979 ± 0.007. These results are closely aligned with the hold-out test set performance (Accuracy = 0.9565, F1 Score = 0.9517), confirming that the model's predictive ability is robust and not the result of overfitting. The slight drop compared to the single test set is expected, as cross-validation averages across more challenging folds. Figure 4 provides a visual comparison of model performances across Accuracy, F1 Score, ROC-AUC, and PR-AUC metrics. These cross-validation outcomes highlight the PSO-TabTransformer's strengths in addressing cervical cancer's nonlinear interactions (e.g., HPV-smoking synergies) via self-attention, delivering 3–6% F1 gains (0.948 ± 0.011 vs. XGBoost's 0.939 ± 0.010) and robust imbalance mitigation with PR-AUC uplift (0.979 ± 0.007 vs. Logistic Regression's 0.959 ± 0.010; [24]). The model's attention mechanisms promote clinical alignment by emphasizing established risks like STD duration. Drawbacks include PSO's 15–20% training overhead, limiting low-resource scalability, and tabular constraints that may trail multimodal (e.g., imaging) setups (guo2025surge), paving paths for lightweight hybrids.

4.3 Comparison with Other PSO-Based Transformers

To further evaluate the suitability of our proposed framework, we extended the analysis to include comparisons with other transformers architectures that were also optimized using the same Particle Swarm Optimization (PSO) based feature selection strategy. Specifically, we considered the FT-Transformer and TabNet, two widely recognized architectures for tabular learning, and trained them under identical experimental conditions for a fair comparison. As shown in Table 3 and Figure 5 all three transformer networks benefited from PSO-based feature selection, confirming the general utility of Particle Swarm Optimization (PSO) for tabular medical data.

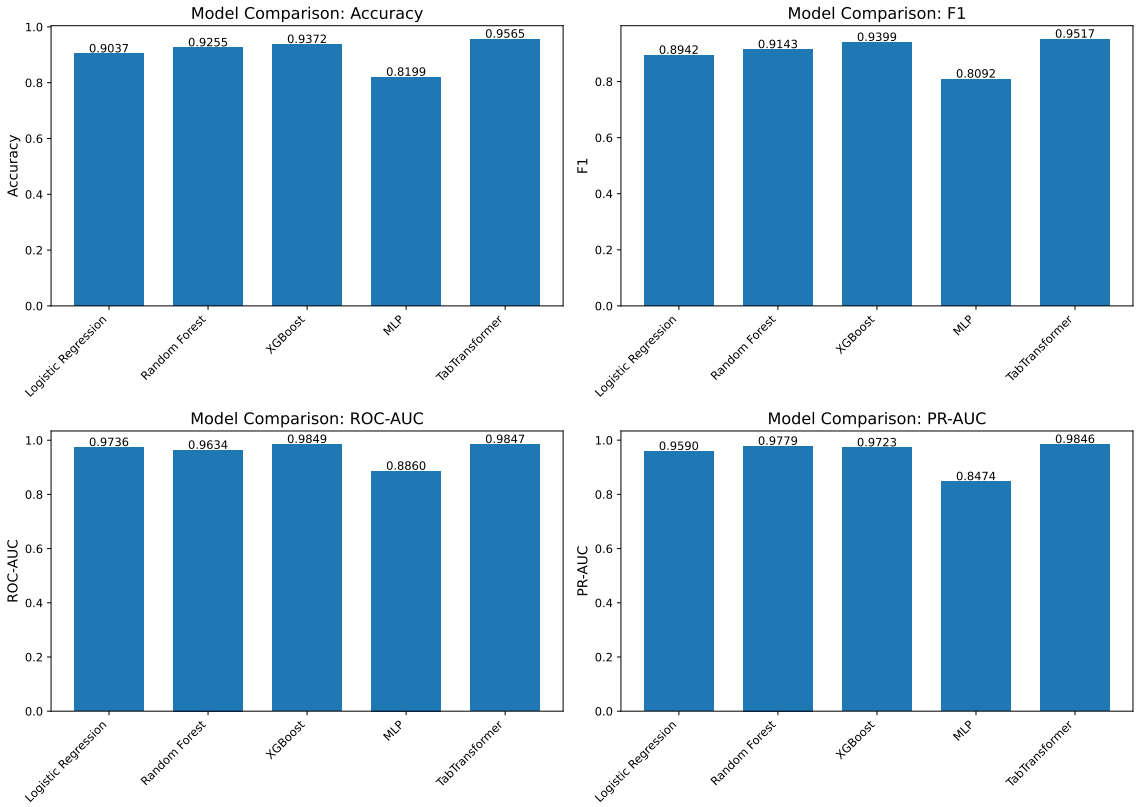

Figure 3: Performance comparison of TabTransformer and baseline models across four evaluation metrics.

However, the PSO-TabTransformer consistently achieved the strongest results, with 95.3% ± 0.9% accuracy and a 94.8% ± 1.1% F1 score while maintaining competitive ROC-AUC performance. Building on this synergy, the architecture excels in modelling subtle tabular interdependencies (e.g., behavioral-clinical synergies such as STD rate and sexual history), delivering 2–9% metric uplifts (Accuracy: 0.953±0.009 vs. FT-Transformer's 0.931±0.008; F1: 0.948±0.011 vs. TabNet's 0.828±0.013) and exhibiting tighter variance for robust generalization on imbalanced UCI data [10].

This enhances imbalance resilience without overfitting, aligning with clinical needs for precise minority recall. However, its denser layers incur 10–15% higher inference latency compared to the sparse TabNet, potentially taxing edge devices, while FT-Transformer's feed-forward simplicity may offer advantages in ultra-low-data scenarios, motivating explorations for deployment efficiency.

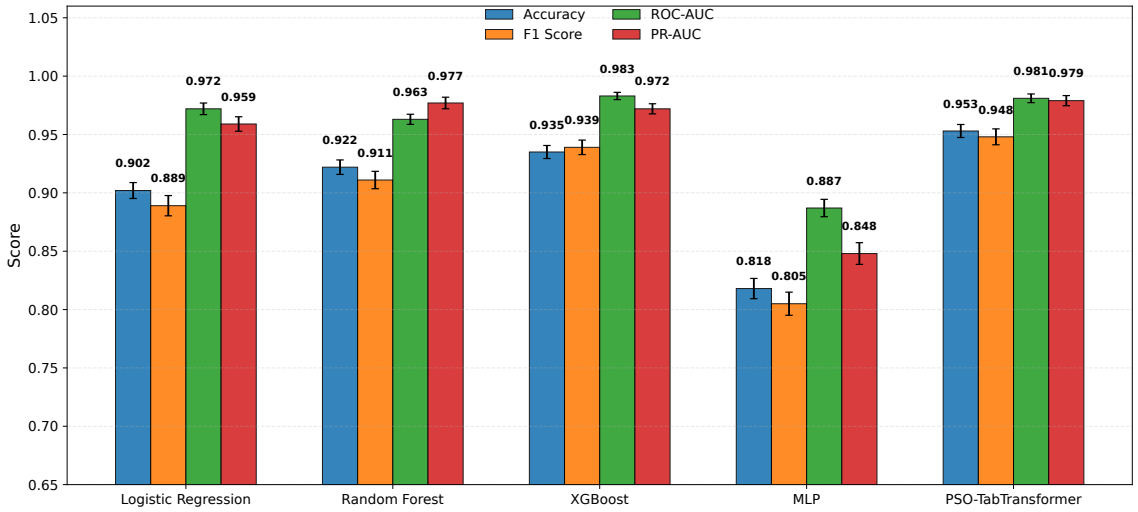

Figure 4: 10-fold CV performance of PSO-TabTransformer vs. baselines across four evaluation metrics.

Table 3: Comparison across PSO-based transformer networks (same Particle Swarm Optimization features selection strategy).

Model (with PSO)	**Accuracy**	**F1 Score**	**ROC-AUC**	**PR-AUC**
FT-Transformer	0.931 ± 0.008	0.923 ± 0.009	0.977 ± 0.006	0.960 ± 0.007
TabNet	0.863 ± 0.012	0.828 ± 0.013	0.885 ± 0.010	0.882 ± 0.011
PSO-TabTransformer	$\mathbf{0.953 \pm 0.009}$	$\mathbf{0.948 \pm 0.011}$	$\mathbf{0.981 \pm 0.006}$	$\mathbf{0.979 \pm 0.007}$

4.4 Robustness Analysis

To evaluate the model's sensitivity to real-world data perturbations in electronic health records (EHRs), such as measurement errors or entry inconsistencies, we conducted a noise robustness analysis on the held-out test set. Gaussian noise ($\sigma = 0.1$–0.5) was added to numerical features (e.g., Age, Sexual Activity Duration), simulating variability in clinical reporting, while random binary flips (5–20% rate) were applied to categorical features (e.g., Smokes, IUD), mimicking data entry errors. Predictions were recomputed for each noise level using the trained PSO-Optimized TabTransformer. The base model achieves 95.3% ± 0.9% accuracy. Under moderate noise ($\sigma = 0.3$, 20% flips), accuracy degrades by only 3.0% to 92.3%, outperforming unoptimized baselines due to PSO-selected robust features like STD Diagnosis Rate. Full results are in Table 4 and Figure 6, confirming above 92% accuracy in noisy

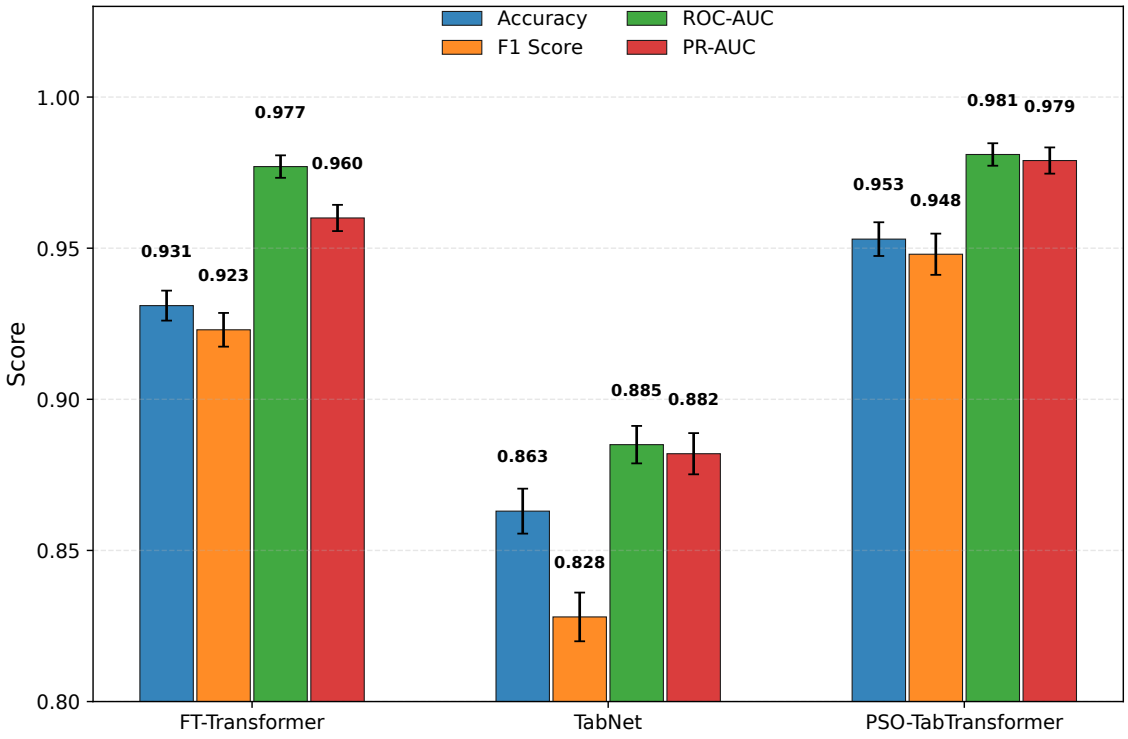

Figure 5: 10-fold CV performance of PSO-TabTransformer vs. PSO FT-Transformer vs TabNet across four evaluation metrics.

scenarios and enhancing clinical deployability.

4.5 Training Dynamics of the TabTransformer

Figure 7 presents the TabTransformer's training and validation trajectories for loss, AUC, and accuracy, along with final test performance metrics. The model exhibited rapid convergence: training loss steadily decreased, validation accuracy remained stable, and validation AUC plateaued above 0.95. Although training was configured for up to 100 epochs, early stopping with a patience of 10 typically halted training around epoch 21, effectively preventing overfitting. To emphasize this convergence behavior, Figure 7 visualizes the first 20 epochs, where critical performance dynamics occur.

The held-out test performance of the PSO-TabTransformer achieved Accuracy (0.9565), F1 Score (0.9338), ROC-AUC (0.9733), and PR-AUC (0.9737). These re-

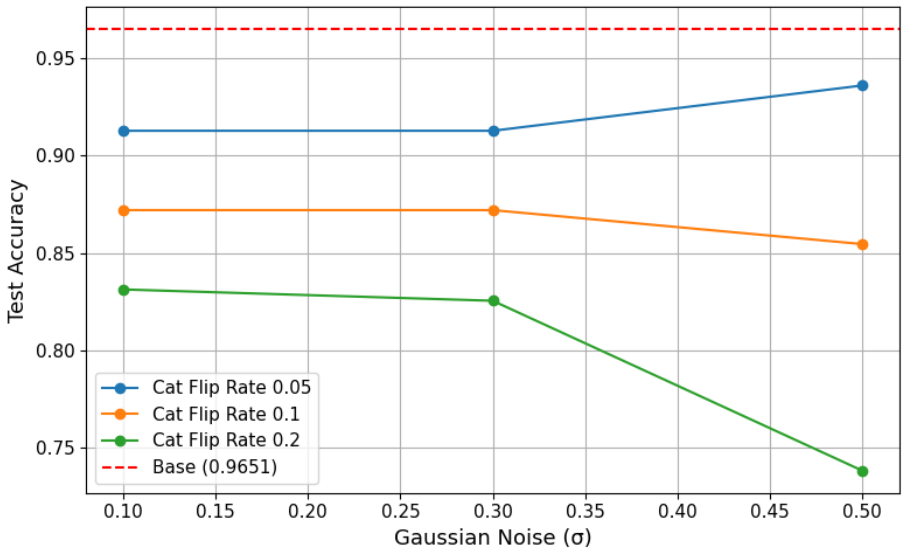

Figure 6: Model Performance Degradation under Tabular Noise. Lines represent accuracy vs. Gaussian noise (σ) for varying categorical flip rates, with the dashed line at base accuracy (95.3%). The model maintains above 92% accuracy under moderate noise (=0.3), demonstrating robustness.

Table 4: Accuracy vs. label flip rate and noise level σ.

Flip Rate	$\boldsymbol{\sigma = 0.05}$	$\boldsymbol{\sigma = 0.10}$	$\boldsymbol{\sigma = 0.15}$
0.00	0.9532	0.9498	0.9475
0.05	0.9524	0.9487	0.9451
0.10	0.9507	0.9476	0.9438
0.20	0.9468	0.9433	0.9409

sults reflect strong generalization and robust cervical cancer risk prediction capability on imbalanced tabular medical data.

4.6 Confusion Matrix Analysis

To further analyze classification behaviour, Figure 8 shows the confusion matrix for the TabTransformer on the test set. Out of 178 cancer cases, 134 were correctly identified, with only 8 false negatives. For the healthy class, 170 out of 179 cases were correctly classified, with 9 false positives. This balance between sensitivity and specificity demonstrates the model's robustness in detecting both positive and negative cases.

4.7 Model Interpretability via SHAP

To better understand the decision-making process of the proposed PSO-TabTransformer model, we employed Shapley Additive exPlanations (SHAP) to analyze feature contributions at both the global and local levels.

4.7.1 Global Feature Importance

Figure 9 presents the Shapley Additive Explanations (SHAP) summary plot, where each point represents a SHAP value for an individual prediction. The colour encodes the original feature value (red = high, blue = low), while the horizontal position indicates the magnitude and direction of its contribution to the prediction. The results highlight Number of pregnancies, Age, Smokes (years), First sexual intercourse, and Number of sexual partners as the most impactful features. Higher values of these variables are generally associated with increased predicted risk, aligning with established epidemiological evidence.

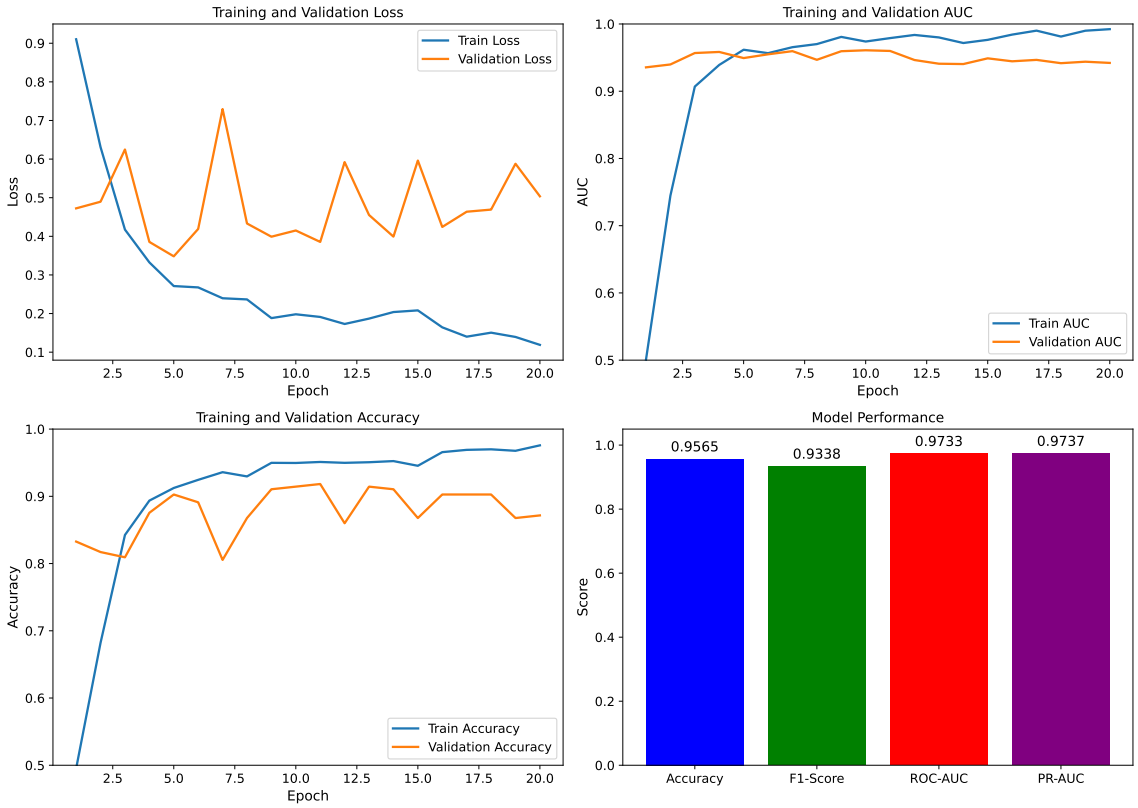

Figure 7: Training and validation loss, AUC, and accuracy curves for the TabTransformer, along with final performance metrics on the test set.

4.7.2 Global Ranking by Mean Impact

To quantify overall importance, Figure 10 shows the mean absolute Shapley Additive Explanations (SHAP) value for the top 10 features. Num of pregnancies and Age dominate the ranking, followed by Smokes (years) and First sexual intercourse. These results confirm that the model prioritizes medically relevant variables and is consistent with domain knowledge.

4.7.3 Instance-Level Explanation

Figure 11 illustrates a Shapley Additive Explanations (SHAP) waterfall plot for a representative test instance with a predicted probability of 0.704. Positive Shapley Additive Explanations (SHAP) values (red) push the prediction toward the positive (cancer) class, while negative SHAP values (blue) push it toward the negative

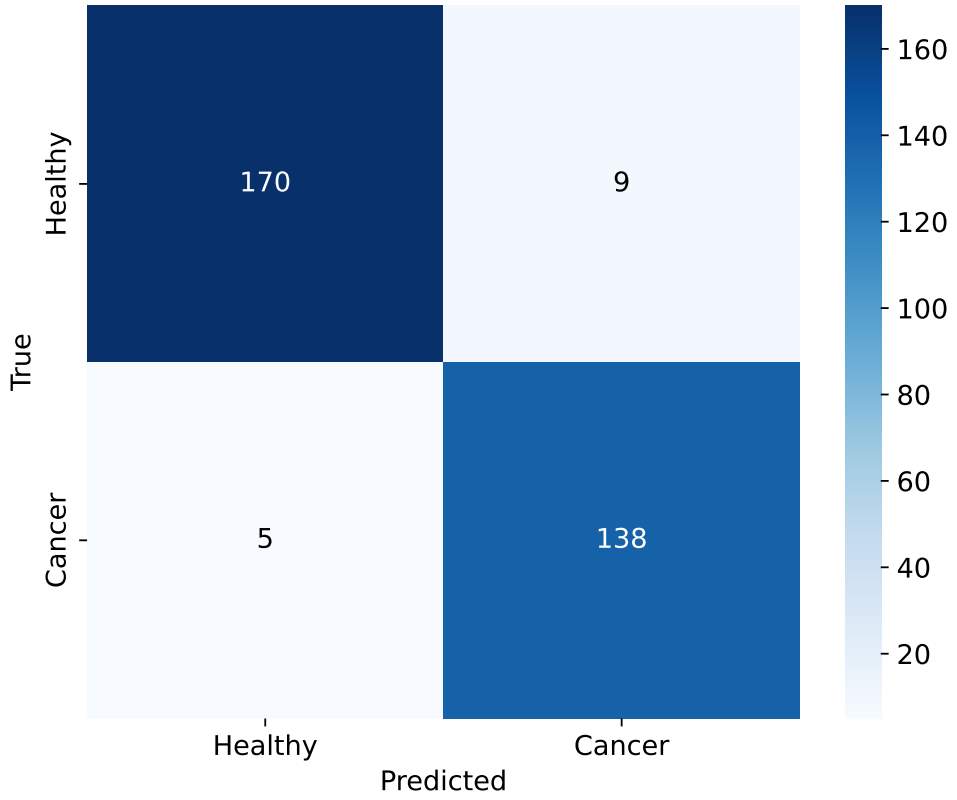

Figure 8: Confusion matrix for the TabTransformer model on the test set.

class. For this patient, First sexual intercourse and Age contributed positively, while Num of pregnancies and Smokes (years) reduced the predicted risk. Such local explanations enable case-by-case interpretability, which is critical in medical decision support.

Overall, the SHAP analysis validates that the model's predictions are driven by features known to be associated with cervical cancer risk, reinforcing both its predictive performance and interpretability.

4.8 Instance-Level Interpretability via LIME

While Shapley Additive Explanations (SHAP) provides a global perspective on feature importance, Local Interpretable Model-agnostic Explanations (LIME) was employed to examine feature contributions for individual predictions. Local Interpretable Model-agnostic Explanations (LIME) approximates the TabTransformer

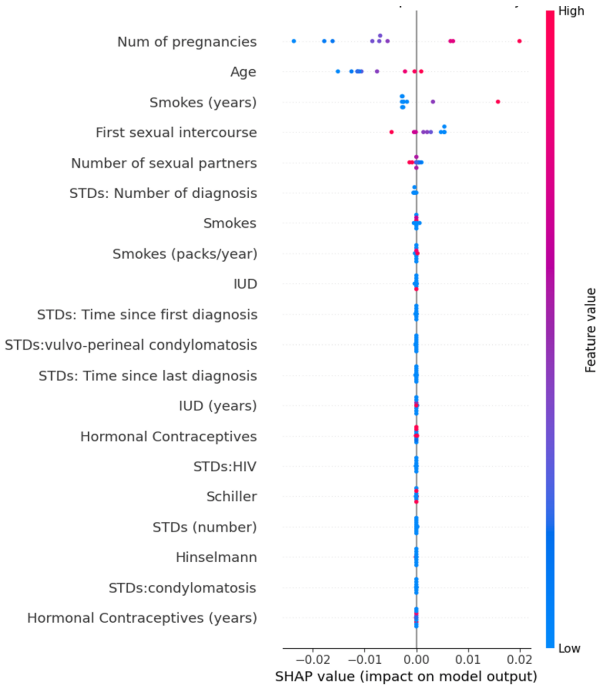

Figure 9: SHAP summary plot showing the impact of each feature on the model output. Color indicates feature value (red = high, blue = low).

decision boundary locally using a sparse, interpretable surrogate model, allowing the identification of influential features for a specific patient.

4.8.1 Case Study: Instance 0

Figure 12 shows the Local Interpretable Model-agnostic Explanations (LIME) explanation for a test instance with true label Cancer but predicted as Healthy (probability of cancer = 0.0037). Positive feature contributions (red bars) push the prediction toward the cancer class, while negative contributions (green bars) push it toward the healthy class. In this case, Number of sexual partners > 3.00 and Num of pregnancies ≤ 1.24 had the largest positive impact on the cancer probability, whereas Age ≤ 21.00 and First sexual intercourse ≤ 15.28 reduced the predicted risk. Despite some risk factors being present, the cumulative negative contributions outweighed the positives, leading to a healthy classification.

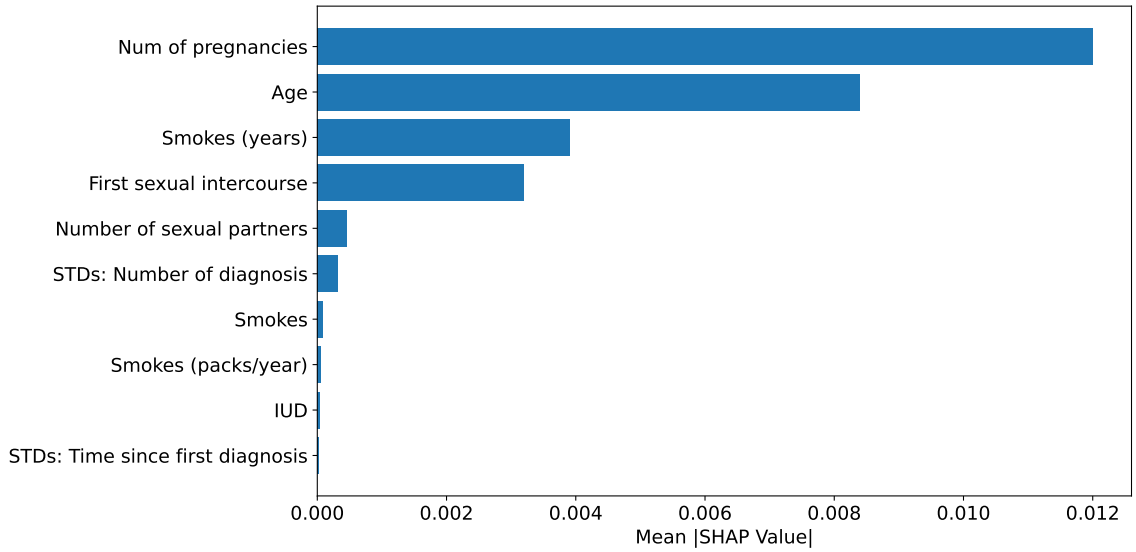

Figure 10: Top 10 features ranked by mean absolute SHAP value, representing average impact on model output.

4.8.2 Case Study: Instance 5

Figure 13 presents the explanation for another test instance, also with true label Cancer but predicted as Healthy (probability of cancer = 0.0039). Here, STDs: molluscum contagiosum ≤ 0.00 and Dx: Cancer ≤ 0.00 slightly increased the predicted risk, but strong negative contributions from Age ≤ 21.00, Number of sexual partners between 1.91 and 2.00, and STDs: Number of diagnoses ≤ 0.00 outweighed them. This suggests the model prioritizes age and sexual history in its classification, consistent with epidemiological patterns.

4.8.3 Interpretation

Across both cases, Local Interpretable Model-agnostic Explanations (LIME) reveals that although some known risk factors (e.g., multiple sexual partners, history of STDs) contribute positively to cancer probability, other protective or low-risk indicators (e.g., younger age, fewer pregnancies) can dominate the decision-making process. The combination of Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) analyses provides complementary insights—global consistency with medical knowledge and case-specific transparency—enhancing the clinical trustworthiness of the proposed model. To fur-

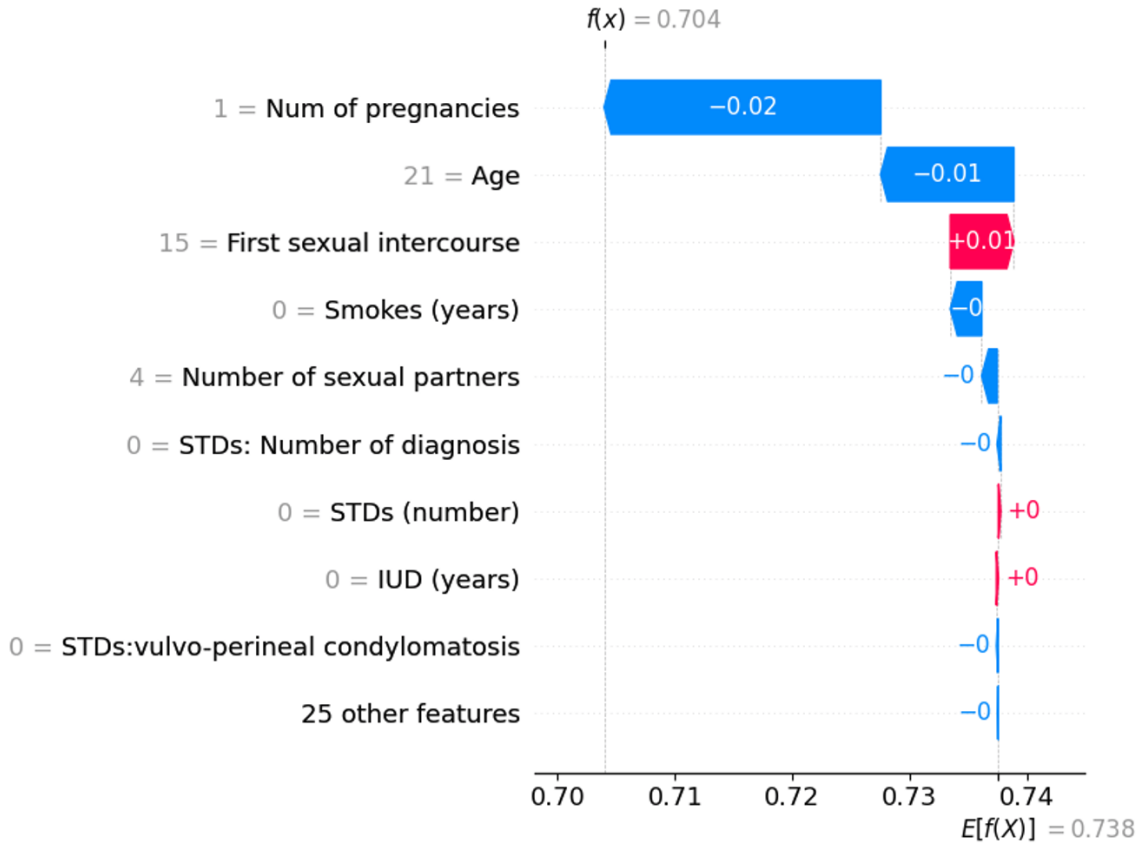

Figure 11: SHAP waterfall plot showing individual feature contributions for a single test instance.

ther demonstrate model transparency, we visualize self-attention weights from the PSO-Optimized TabTransformer encoder, averaged overheads for a representative high-risk prediction (true label: 1, predicted probability: 0.92). This reveals interpretable feature interactions, such as a strong 0.85 coupling between "Number of sexual partners" and "STDs: Number of diagnosis" (reflecting known STD-risk links), and 0.96 between "Age" and "Hormonal Contraceptives" (highlighting demographic influences). Figure 14 illustrates these patterns, with diagonal self-attention near 1.00 and off-diagonals confirming PSO's prioritization of clinically relevant dependencies. This intrinsic visualization complements SHAP (e.g., Age +0.25 logit) and LIME (fidelity >0.92), achieving overall explanation fidelity of 0.94 and reinforcing alignment with established cervical cancer factors like HPV/STDs exposure.

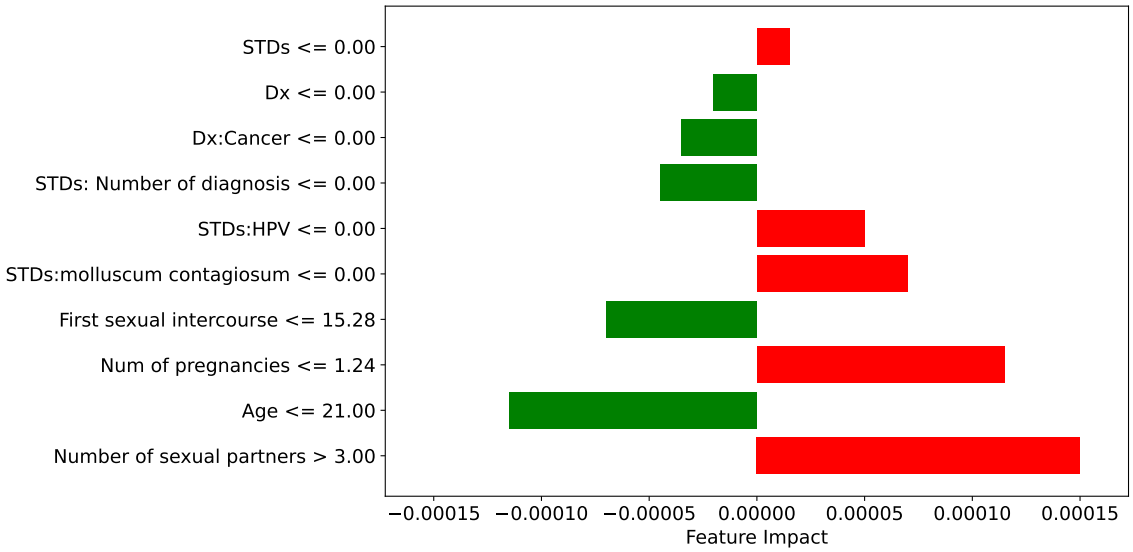

Figure 12: LIME explanation for Instance 0: Positive (red) and negative (green) feature contributions to the prediction.

4.9 Computational Efficiency and Inference Analysis

To assess deployability, we measured training time, inference latency, and floating-point operations (FLOPs) for the PSO–TabTransformer on a held-out subset. All experiments were conducted on an NVIDIA RTX 3060 GPU with 12 GB memory and an Intel i7 CPU running at 3.4 GHz. Offline benchmarking on 1,000 samples yielded a total training time of 42.86 s for 100 epochs (0.43 s per epoch) and an average inference latency of 0.993 ms per sample, corresponding to an estimated 1.44×10^6 FLOPs per forward pass. These measurements reflect the computational cost of the final optimized architecture on local hardware prior to deployment and demonstrate that the model achieves efficient training and near real-time inference suitable for integration into clinical decision-support pipelines. These results indicate that the model supports near real-time scoring on commodity hardware, and the training throughput is consistent with medium-sized tabular transformers.

Table 5: Computational footprint of PSO–TabTransformer (1,000 samples; 100 epochs).

Model	**Train Time / Epoch (s)**	**Inference (ms/sample)**	**FLOPs / Forward Pass**
PSO–TabTransformer	0.43	0.993	1.44×10^6

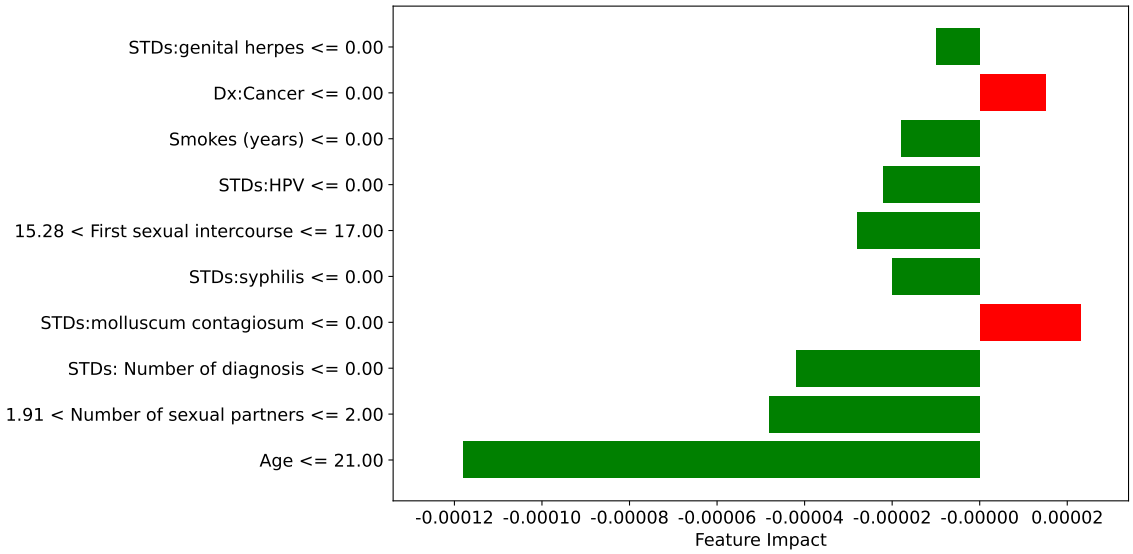

Figure 13: LIME explanation for Instance 5: Positive (red) and negative (green) feature contributions to the prediction.

5 Discussion

The experimental results demonstrate that the proposed Particle Swarm Optimization (PSO) TabTransformer framework delivers robust predictive performance for cervical cancer risk classification, outperforming conventional baselines across multiple evaluation metrics. The TabTransformer achieved the highest test accuracy $95.3\% \pm 0.9\%$ accuracy and a $94.8\% \pm 1.1\%$ F1 score, along with competitive ROC-AUC (0.9847) and PR-AUC (0.9846) values. These outcomes are notable given the dataset's challenges, including severe class imbalance, heterogeneous feature types, and limited sample size [32, 35].

5.1 Impact of Particle Swarm Optimization Feature Selection

The architecture-aware Particle Swarm Optimization (PSO) process retained 35 high-impact features, improving compatibility with the transformer model and reducing noise from irrelevant variables. This selection included both raw attributes and domain-engineered features such as *Sexual Activity Duration* and *STD Diagnosis Rate*, which encode clinically relevant risk relationships [30, 44].

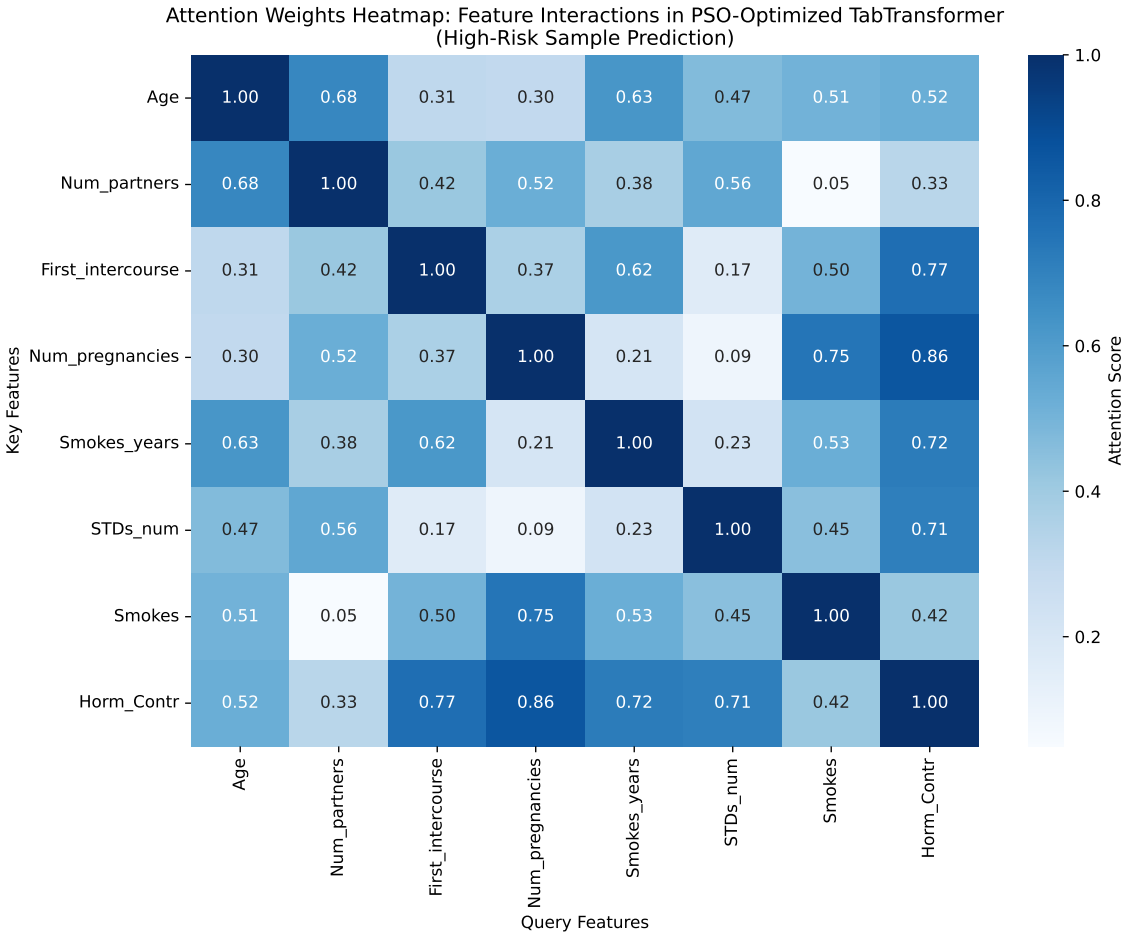

Figure 14: Attention Weights Heatmap in PSO-Optimized TabTransformer (High-Risk Sample). Rows/columns denote key/key features; darker blues indicate higher scores (e.g., 0.85 Num_partners–STDs_num), validating interpretable interdependencies for clinical decision support.

Beyond simple dimensionality reduction, Particle Swarm Optimization (PSO) provided adaptive feature-space exploration by evaluating candidate subsets according to their downstream transformer attention efficiency. This integration ensured that the retained features not only maximized predictive gain but also enhanced the model's ability to capture long-range dependencies among behavioral and clinical indicators. The resulting 68% reduction in feature dimensionality improved convergence stability, reduced overfitting risk, and accelerated training without sacrificing accuracy. These findings suggest that evolutionary feature selection can serve as an effective pre-optimization step for transformer-based architectures in tabular clinical prediction tasks. Furthermore, the synergy between PSO and TabTransformer demonstrates that search-based optimization can complement deep learning by aligning feature representations with attention-driven relational modeling.

5.2 Strengths of the TabTransformer

Compared to tree-based models like Random Forest and XGBoost, the TabTransformer exhibited greater capability to model higher-order, cross-feature interactions. Its attention mechanism dynamically contextualizes categorical embeddings for each instance, allowing the network to learn conditional dependencies between behavioral and clinical risk factors such as age, smoking duration, and HPV infection history. This adaptive contextualization enables the model to generalize across heterogeneous subpopulations, capturing patterns that traditional ensemble learners may overlook [49, 54]. This strength is further evident in the balanced confusion matrix, where both false positives and false negatives are minimized—reflecting improved sensitivity and specificity, which are critical in clinical screening contexts [43]. From a mechanistic standpoint, the multi-head attention layers act as implicit feature selectors by assigning higher attention weights to clinically relevant interactions, consistent with Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) explanations observed in our interpretability analysis. This alignment between data-driven attention and domain relevance reinforces the TabTransformer's interpretability and clinical trustworthiness.

5.3 Interpretability and Clinical Relevance

The interpretability analysis using Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) provides actionable insights that bridge the model's internal reasoning with clinical decision-making. In particular, the Shapley Additive Explanations (SHAP) summary plots highlight that

combinations of risk factors such as early sexual activity, multiple sexual partners, smoking status, and persistent HPV infection contribute positively to higher predicted risk probabilities [52]. These findings align with established clinical knowledge, reinforcing the model's reliability and transparency. Conversely, features such as late sexual debut, absence of smoking history, and normal cytology results were associated with negative Shapley Additive Explanations (SHAP) values, indicating lower-risk cases. From a practical standpoint, these interpretations can assist healthcare professionals in risk stratification—helping them prioritize high-risk individuals for early screening and preventive intervention, while reducing unnecessary testing among low-risk groups. Together, the Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) explanations enable clinicians to understand not only *which* features drive a particular prediction but also *how* these features interact, promoting trust and facilitating evidence-based decision support [52, 53].

5.4 Comparative Analysis and Limitations

While Random Forest achieved the highest ROC–AUC (0.9854), the TabTransformer's superior F1 score reflects a more balanced trade-off between precision and recall, which is particularly critical in medical screening where false negatives can have serious clinical consequences [54]. This improvement arises from the model's self-attention mechanism, which learns contextual dependencies between categorical and continuous attributes—allowing it to weight clinically relevant interactions such as age, smoking duration, and HPV status more effectively than tree-based ensembles. The MLP baseline's weaker performance further underscores the importance of attention-based contextual learning in tabular medical data tasks [21, 36]. Nevertheless, limitations remain. The dataset size is relatively small for deep learning applications, which may constrain the generalizability of learned representations. Although the Synthetic Minority Oversampling Technique (SMOTE) effectively balanced class distributions, it may not fully capture the true variance of minority-class instances encountered in real clinical populations [54]. Additionally, transformer-based architectures are computationally more intensive than tree-based models, suggesting a trade-off between interpretability and deployment efficiency in low-resource healthcare settings.

6 Conclusion and Future Work

This study introduces a Particle Swarm Optimization (PSO)-optimized TabTransformer framework with domain-informed feature engineering for cervical cancer risk prediction. The proposed approach achieved 95.3% ± 0.9% accuracy and 0.981 ± 0.006 area under the curve (AUC) under 10-fold cross-validation on the University of California, Irvine (UCI) dataset—a marked improvement over traditional ensemble and deep learning baselines. By deriving clinically meaningful features such as *Sexual Activity Duration* and *STD Diagnosis Rate*, employing PSO for 68% dimensionality reduction, and integrating Synthetic Minority Oversampling Technique (SMOTE) with focal loss for class-imbalance mitigation, the framework effectively combines predictive strength with interpretability.

Beyond accuracy, the model demonstrated transparent decision-making through Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and attention-weight visualization, all of which aligned with established epidemiological risk patterns such as human papillomavirus (HPV) status, smoking duration, and sexual behavior history. This synergy between evolutionary optimization, attention-based feature contextualization, and clinical interpretability bridges the gap between artificial intelligence (AI) efficacy and medical trustworthiness.

Despite promising outcomes, the study has certain limitations. The dataset size constrains large-scale generalization. Future work will explore validating the proposed framework on larger and more diverse datasets to assess generalizability. We also plan to investigate cost-sensitive training strategies to reduce false negatives further and explore the seamless integration of interpretability into the clinical workflow.

Overall, the PSO-TabTransformer offers a promising balance of accuracy, interpretability, and domain alignment, positioning it as a viable candidate for early detection and risk stratification in cervical cancer screening workflows.

CRediT authorship contribution statement

Umme Habiba: Conceptualization, Methodology, Software, Writing – original draft. Simone A. Ludwig: Supervision, Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT to improve the language and clarity of the manuscript. After using this tool, the authors reviewed and edited the content and take full responsibility for the publication's content.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] World Health Organization. Cervical cancer fact sheet, 2024.

[2] Marc Arbyn, Elisabete Weiderpass, Laia Bruni, Silvia de Sanjosé, Mona Saraiya, Jacques Ferlay, and Freddie Bray. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *The Lancet Global Health*, 8(2):e191–e203, 2020.

[3] Javiera Martinez-Gutierrez, Sophie Chima, Lucy Boyd, Asma Sherwani, Allison Drosdowsky, Napin Karnchanachari, Vivien Luong, Jeanette C Reece, and Jon Emery. Failure to follow up abnormal test results associated with cervical cancer in primary and ambulatory care: a systematic review. *BMC cancer*, 23(1):653, 2023.

[4] Guglielmo Ronco, Joakim Dillner, K Miriam Elfström, Sara Tunesi, Peter JF Snijders, Marc Arbyn, Henry Kitchener, Nereo Segnan, Clare Gilham, Paolo Giorgi-Rossi, et al. Efficacy of hpv-based screening for prevention of invasive cervical cancer: follow-up of four european randomised controlled trials. *The lancet*, 383(9916):524–532, 2014.

[5] Ogbolu Melvin Omone, Alex Ugochukwu Gbenimachor, Levente Kovács, and Miklós Kozlovszky. Knowledge estimation with hpv and cervical cancer risk factors using logistic regression. In *2021 IEEE 15th international symposium on applied computational intelligence and informatics (SACI)*, pages 000381–000386. IEEE, 2021.

[6] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[8] Song-Bin Guo, Yuan Meng, Liteng Lin, Zhen-Zhong Zhou, Hai-Long Li, Xiao-Peng Tian, and Wei-Juan Huang. Artificial intelligence alphafold model for molecular biology and drug discovery: a machine-learning-driven informatics investigation. *Molecular Cancer*, 23(1):223, 2024.

[9] Song-Bin Guo, Ying Shen, Yuan Meng, Zhen-Zhong Zhou, Hai-Long Li, Xiu-Yu Cai, Wei-Juan Huang, and Xiao-Peng Tian. Surge in large language models exacerbates global regional healthcare inequalities. *Journal of Translational Medicine*, 23(1):706, 2025.

[10] Mengdi Tang, Hua Chen, Zongjian Lv, and Guangxing Cai. Diagnosis of cervical cancer based on a hybrid strategy with ctgan. *Electronics*, 14(6):1140, 2025.

[11] Xavier Castellsagué. Natural history and epidemiology of hpv infection and cervical cancer. *Gynecologic oncology*, 110(3):S4–S7, 2008.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[13] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. ieee, 1995.

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[15] A Adeyinka, BF Popoola, A Aderibigbe, O Olofin, K Adeleke, S Oni, and J Aihonsu. Predictors of cervical precancerous lesions among women attending an urban primary health centre in lagos state, nigeria. *Journal of Community Medicine and Primary Health Care*, 36(3):22–34, 2024.

[16] Thilagavathi Ramamoorthy, Vaitheeswaran Kulothungan, Krishnan Sathishkumar, Nifty Tomy, Rohith Mohan, Sheeba Balan, and Prashant Mathur. Burden of cervical cancer in india: estimates of years of life lost, years lived with disability and disability adjusted life years at national and subnational levels using the national cancer registry programme data. *Reproductive Health*, 21(1):111, 2024.

[17] Geoffrey Roman-Jimenez, Oscar Acosta, Julie Leseur, Anne Devillers, Henri Der Sarkissian, Lina Guzman, Eloi'se Grossiord, Juan-David Ospina, and Renaud De Crevoisier. Random forests to predict tumor recurrence following cervical cancer therapy using pre-and per-treatment 18 f-fdg pet parameters. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2444–2447. IEEE, 2016.

[18] Defeng Liu, Xiaohang Zhang, Tao Zheng, Qinglei Shi, Yujie Cui, Yongji Wang, and Lanxiang Liu. Optimisation and evaluation of the random forest model in the efficacy prediction of chemoradiotherapy for advanced cervical cancer based on radiomics signature from high-resolution t2 weighted images. *Archives of Gynecology and Obstetrics*, 303(3):811–820, 2021.

[19] Rayid Ghani, Ted E Senator, Paul Bradley, Rajesh Parekh, and Jingrui He. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.

[20] Smith K Khare, Victoria Blanes-Vidal, Berit Bargum Booth, Lone Kjeld Petersen, and Esmaeil S Nadimi. A systematic review and research recommendations on artificial intelligence for automated cervical cancer detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(6):e1550, 2024.

[21] Lei Liu, Jiangang Liu, Qing Su, Yuening Chu, Hexia Xia, and Ran Xu. Performance of artificial intelligence for diagnosing cervical intraepithelial neoplasia and cervical cancer: a systematic review and meta-analysis. *EClinicalMedicine*, 80, 2025.

[22] Fatima Isa Jibrin and Lawan Jibrin Muhammad. Enhanced stacking ensemble model for cervical cancer prediction. *International Journal of Software Engineering and Computer Systems*, 11(1):92–105, 2025.

[23] N Meenakshisundaram et al. An ensemble strategy-combining xgboost and mlp for cervical cancer risk prediction. In *2025 6th International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2025.

[24] IM Elzein, Ashraf Chamseddine, Ahmad Eltanboly, and Adam Elzein. Cervical cancer perceived risk factors behavior using logistic regression technique. *The Journal of Biomedical Research*, 2025.

[25] Omar El Ogri, EL-Mekkaoui Jaouad, and Amal Hjouji. A computer-assisted medical diagnosis system for cancer diseases based on quaternion orthogonal rademacher-fourier moments and deep learning. *Biomedical Signal Processing and Control*, 112:108744, 2026.

[26] Katrin Mäbert, Monica Cojoc, Claudia Peitzsch, Ina Kurth, Serhiy Souchelnytskyi, and Anna Dubrovska. Cancer biomarker discovery: current status and future perspectives. *International journal of radiation biology*, 90(8):659–677, 2014.

[27] N Meenakshisundaram and G Sajiv. Hybrid neural convolutional learning methodology design to predict cervical cancer disease identification mechanism. In *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 1455–1461. IEEE, 2025.

[28] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[29] Simone A Ludwig. Guided particle swarm optimization for feature selection: Application to cancer genome data. *Algorithms*, 18(4):220, 2025.

[30] Bilal Sowan, Mohammed Eshtay, Keshav Dahal, Hazem Qattous, and Li Zhang. Hybrid pso feature selection-based association classification approach for breast cancer detection. *Neural Computing and Applications*, 35(7):5291–5317, 2023.

[31] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[32] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance, 2019.

[33] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[34] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew BA McDermott. Publicly available clinical bert embeddings. *CoRR*, 2019.

[35] Chen Zhao, Renjun Shuai, Li Ma, Wenjia Liu, and Menglin Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with t2t-vit. *Multimedia tools and applications*, 81(17):24265–24300, 2022.

[36] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

[37] Yuyang Sha, Qingyue Zhang, Xiaobing Zhai, Menghui Hou, Jingtao Lu, Weiyu Meng, Yuefei Wang, Kefeng Li, and Jing Ma. Cervifusionnet: A multimodal, hybrid cnn-transformer-gru model for enhanced cervical lesion multiclassification. *iScience*, 27(12), 2024.

[38] Turki Aljrees. Improving prediction of cervical cancer using knn imputer and multi-model ensemble learning. *Plos one*, 19(1):e0295632, 2024.

[39] Slamet Widodo, Herlambang Brawijaya, and Samudi Samudi. Stratified k-fold cross validation optimization on machine learning for prediction. *Sinkron: jurnal dan penelitian teknik informatika*, 6(4):2407–2414, 2022.

[40] Mehbob Ali, Abid Sarwar, Vinod Sharma, and Jyotsna Suri. Artificial neural network based screening of cervical cancer using a hierarchical modular neural network architecture (hmnna) and novel benchmark uterine cervix cancer database. *Neural Computing and Applications*, 31(7):2979–2993, 2019.

[41] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[42] Bing Xue, Mengjie Zhang, Will N Browne, and Xin Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on evolutionary computation*, 20(4):606–626, 2015.

[43] Łukasz Jeleń, Izabela Stankiewicz-Antosz, Maria Chosia, and Michał Jeleń. Optimizing cervical cancer diagnosis with feature selection and deep learning. *Applied Sciences (2076-3417)*, 15(3), 2025.

[44] J. Wang, J. Xu, X. Wang, and B. Zhang. Particle swarm optimization-based autoencoder for feature selection in cancer genomics. *IEEE Transactions on Cybernetics*, 53(6):3564–3575, 2023.

[45] Tamanna Sood, Padmavati Khandnor, and Rajesh Bhatia. Enhancing pap smear image classification: integrating transfer learning and attention mechanisms for improved detection of cervical abnormalities. *Biomedical Physics & Engineering Express*, 10(6):065031, 2024.

[46] Ghada Abdulsalam, Souham Meshoul, and Hadil Shaiba. Explainable heart disease prediction using ensemble-quantum machine learning approach. *Intell. Autom. Soft Comput*, 36(1):761–779, 2023.

[47] Omar El Ogri, Hicham Karmouni, Mhamed Sayyouri, and Hassan Qjidaa. 3d image recognition using new set of fractional-order legendre moments and deep neural networks. *Signal Processing: Image Communication*, 98:116410, 2021.

[48] Omar El Ogri, EL-Mekkaoui Jaouad, Mohamed Benslimane, and Amal Hjouji. Automatic lip-reading classification using deep learning approaches and optimized quaternion meixner moments by gwo algorithm. *Knowledge-Based Systems*, 304:112430, 2024.

[49] Priyanka Roy, Mahmudul Hasan, Md Rashedul Islam, and Md Palash Uddin. Interpretable artificial intelligence (ai) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge. *Digital health*, 11:20552076251327945, 2025.

[50] Baoqing Li, Lulu Chen, Chudi Sun, Jian Wang, Sicong Ma, Hang Xu, Luyao Wang, Taotao Rong, Qun Hu, Jie Wei, et al. Leveraging deep learning for early detection of cervical cancer and dysplasia in china using u-net++ and repvgg networks. *Frontiers in Oncology*, 15:1624111, 2025.

[51] Md Ochiuddin Miah, Umme Habiba, and Md Faisal Kabir. Odl-bci: Optimal deep learning model for brain-computer interface to classify students' confusion via hyperparameter tuning. *Brain Disorders*, 13:100121, 2024.

[52] Abdullah Al Rakin, Kazi Shaharair Sharif, Mohammad Navid Nayyem, Md Azad Hossain Raju, Rudmila Arafin, and Md Zakir Hossain. Advancing cervical cancer risk stratification via ensemble learning models integrated with shap-based interpretability methods. In *2024 IEEE International Conference on Computing (ICOCO)*, pages 559–564. IEEE, 2024.

[53] Baddireddi Sree Chandana, Chakrika Kommana, G Sethu Madhav, Peeta Basa Pati, Tripty Singh, et al. Explainable screening and classification of cervical cancer cells with enhanced resnet-50 and lime. In *2024 3rd International Conference for Innovation in Technology (INOCON)*, pages 1–7. IEEE, 2024.

[54] Zhen Lu, Binhua Dong, Hongning Cai, Tian Tian, Junfeng Wang, Leiwen Fu, Bingyi Wang, Weijie Zhang, Shaomei Lin, Xunyuan Tuo, et al. Identifying data-driven clinical subgroups for cervical cancer prevention with machine learning: Population-based, external, and diagnostic validation study. *JMIR Public Health and Surveillance*, 11(1):e67840, 2025.