GMR
The Original

DESIGNING A MULTI-TIER MODEL FOR CERVICAL CANCER PREDICTION USING LEARNING APPROACHES

R. Kiruthika[1], Dr. B. Shanmugham[2]

[1] Department of Electronics and Communications Engineering, School of Engineering And Technology, Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu, India. kiruthikar.1981@rediffmail.com
[2] Department of Electronics and Communication Engineering, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu, India. cyberphd85@gmail.com

ABSTRACT
Cervical cancer poses a significant threat to global health, disproportionately affecting low- and middle-income nations. Lifestyle decisions can play a role in increasing the likelihood of cervical cancer. The primary cause of most cervical cancers is Human PapillomaVirus (HPV), an infection transmitted through intimate contact. HPV Pathogens must persist over time to increase the risk of progressing to pre-cancer and cancer. Several factors contribute to the persistence of HPV infection, including age, STIs, sexual history, reproductive factors, tobacco use, and more. Risk assessment algorithms enable the identification of high-risk women, facilitating prioritized screening for cervical cancer. This research presents a lifestyle-based Multi-tier SVM ($mt - SVM$) classifier model for predicting cervical cancer risk. The $mt - SVM$ classifier was implemented to detect the most crucial characteristics. Oversampled data is then used to train and evaluate the performance of the pattern. An accuracy rate of 98.9% was accomplished using the $mt - SVM$ pattern. This model has the potential to establish a link between risk factors and cervical cancer prediction, facilitating Strategies for prevention and effective management strategies.

KEYWORDS- cancer, prediction, learning, multi-tier, validation

1. INTRODUCTION

Uterine cervical cancer begins with abnormal cell growth in the cervical region. The cervix is the confined, the lowermost section of the uterus, serving as a conduit to the vagina. Cervical cancer starts to develop from abnormal cell growth in the cervix, often progressing slowly over several years before symptoms become apparent. Symptoms of cervical cancer may include unusual vaginal bleeding, including bleeding between menstrual periods, following physical intimacy, or after the onset of menopause [1]. Additional female health indicators may be likely to experience include lower abdominal pain, painful intercourse, or abnormal vaginal bleeding or discharge. Globally, Cervical cancer falls under the category of the major contributor to the number of death rates among the female demographic, with a disproportionate impact on low-resource communities. India reports a significant burden of cervical cancer, with 123,907 new cases diagnosed annually and 77,348 deaths occurring each year, contributing to nearly a quarter of global cervical cancer mortality [2]. The prevention and potential elimination of cervical cancer are within reach, thanks to advances in medical science and public health initiatives. In response, the World Health Organization (WHO) has issued a request for proposals to prevent and control cervical cancer as a population health concern priority threat by 2030 through HPV vaccination, screening, and treatment efforts [3]. The ongoing high incidence of cervical cancer, despite existing screening and preventive measures, highlights the need for an improved risk prediction model to identify at-risk individuals better.

Recent advances in algorithms utilizing risk-based predictions have provided a new and advantageous method for enhancing the effectiveness of cervical health checkups and diagnostic techniques [4]. The analytical frameworks incorporate several variables, including healthcare and sociodemographic factors, such as age, sexual history, HPV status, immune deficiency status, and tobacco use, to predict the likelihood of cervical cancer occurrence in an individual [5]. Researchers have utilized statistical learning techniques to develop and optimize algorithmic models that accurately predict cervical cancer risk, maximizing the accuracy of positive and negative predictions. The accumulation of more comprehensive data, including cytology results and epigenetic landscape information, will likely refine risk-based prediction algorithms, enabling healthcare providers to identify high-risk women earlier and provide more targeted treatment. Developing and refining risk-based prediction algorithms typically involve using predictive modeling techniques, Such as deep learning models, Decision Tree classifiers, and logistic regression [6]. The developed algorithms aid clinicians in pinpointing patients at elevated risk, facilitating timely diagnostic evaluation, such as invasive and non-invasive diagnostic tests, including colposcopy and biopsy, and personalized diagnostic scanning recommendations. Early detection of cervical cancer is crucial for effective treatment and improved outcomes, enabled by diagnostic tools like Pap tests and HPV screening [7]. Unfortunately, the lack of access to screening methods disproportionately affects women in socioeconomically challenged communities, exacerbating existing health disparities. Additionally, women encounter various barriers when attempting to access these services [8].

This necessity prompted the development of a comprehensive cervical cancer risk assessment framework, which makes available a personalized probability of occurrence assessment influenced by cultural and socioeconomic factors, aiming to enhance prompt identification, tailor vulnerability evaluation, effective resource management, and promote informed decision-making. Many investigations into cervical cancer risk prediction have emphasized the development of increased precision predictive artificial intelligence-driven modeling approaches [9]. Nevertheless, the choice of preferred outcome variables in these predictive systems is equally vital for bolstering the validity of the data in a clinical setting. Previous investigations [10] utilized established target variables, including those used in the Hinselmann model, which considers demographic and behavioral factors like age, ethnicity, sexual behavior, and smoking habits [11]. However, the current research has employed a broad perspective, neglecting to adequately weigh the importance of key factors in cervical cancer risk prediction [12] – [15]. Our suggested methodology introduces a unique strategy by utilizing 'Dx: Cancer' and 'Dx: CIN' as the outcome of interest while incorporating established cervical cancer susceptibility factors. The application of this framework could lead to the development of a more" thorough Comprehension of cervical cancer genetic and environmental risk factors leading to improved early detection and prevention strategies. By focusing on cancer diagnosis as the target variable, the developed $mt - SVM$ cervical cancer susceptibility prediction model may provide higher accuracy and clinically meaningful predictions. The ultimate goal of this approach is to improve cervical cancer screening and prevention, with the potential to reduce incidence rates and alleviate the burden on healthcare systems.

The work is organized as: section 2 provides wider review on other related approaches. The methodology is drafted in section 3 with discussion in section 4. The work summary is given in section 5.

2. RELATED WORKS

The author in [16] investigated sociodemographic characteristics influencing cervical cancer risk in India, pinpointing specific characteristics that contribute to an elevated risk. The research found that certain lifestyle factors, such as having multiple sexual partners, practicing poor menstrual hygiene, entering early marriage, and making unhealthy dietary choices, increase the risk. The study reveals a strong association between HPV and cervical cancer, suggesting that HPV infection creates an environment conducive to the development of cervical cancer. The author in [17] conducted a comprehensive analysis of multiple machine learning frameworks on a given collection of data points to examine their ability to predict outcomes and identify areas for improvement. Their results showed an accuracy of 93.33% for both Multilayer Perceptron (MLP) and alternative machine learning models, including Decision Tree Classification (DTC), Random Forest Classifier (RFC), K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM) [18] – [20]. These exact results were obtained after optimizing the Model's Algorithmic parameters.

Additionally, the author et al. [21] employed a chi-square evaluation for attribute identification, selecting the top 10 features with the highest scores and eliminating irrelevant columns. The highest-ranked attribute was the annual number of cigarettes smoked; among the top 10 features, HPV diagnosis and cancer diagnosis received the lowest scores [22]. The researchers applied DBSCAN and iForest to detect and exclude outliers from the dataset resulting in a cleaner and more reliable dataset [23]. These outliers were subsequently removed from the dataset. The dataset was balanced with the application of SMOTE and SMOTETomek, which are intended to mitigate group imbalance through minority class Artificial inflation of the underrepresented group. SMOTE and SMOTETomek are two essential methods in predictive modeling for handling class imbalance and ensuring that models are trained on balanced datasets [24] – [25]. To determine their accuracy and reliability, the researchers evaluated the performance of four cervical cancer examination techniques. Random forest classification methodology was identified in the capacity of the best-performing mathematical procedure, exhibiting the highest-performing parameters for statistical performance measures. It was, therefore, selected as the model algorithm. A cervical cancer risk assessment app was created, using user-inputted data to provide personalized predictions based on the Model's analysis. The author [26] recently utilized Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) [27] to reduce pattern identification and dimension reduction key attributes for classifying Pap smear image classification into five predefined groups.

The author et al. [28] presented a complementary investigation demonstrating that employing a cross-validation framework with stratified sampling can enhance accuracy when combined with a commonly used machine learning algorithm. The author introduced a novel framework for addressing this difficulty by utilizing Boruta analysis, which identifies the most relevant feature subsets for classification tasks. The author introduced research on the design methodology and Construction of a cervical cancer risk prediction algorithm utilizing algorithmic learning approaches. The study involved collecting a study population consisting of 280 Cervical tumor patients and 350 participants with no medical conditions to identify predictive indicators that contribute to the development of cervical cancer. The researchers employed a collection of four algorithmic training strategies- LR, DT, SVM and RF to generate and assess the predictive model. A study's findings revealed that age and analysis showed that pregnancy history, smoking status, and reproductive health measures were all substantial contributors to the threat contributing to the development of cervical neoplasm [29]. The threat assessment tool built employing the Random Forest (RF) rational framework demonstrated exceptional performance, achieving a precision of 0.904 and an Area Under the Receiver Operating Characteristic Curve of 0.966. The author introduced a groundbreaking accuracy medicine approach to cervical cancer screening, offering a new approach to prevention and early

detection. The investigation was based on electronic medical record information from a wide demographic of women (n=58,000, aged 30-64) who underwent screening patterns for cervical neoplasm. The researchers developed a binary regression analysis-based probability estimation framework for developing cervical cancer in women over 5 years, facilitating targeted prevention and early detection. The Model considered multiple factors that influence cervical cancer risk, including age, racial/ethnic background, smoking habits, and the previous diagnosis of cervical dysplasia. The study revealed that the customized screening pattern demonstrated an improved detection rate for cervical cancer detection, surpassing the performance of current guidelines. The study's results imply that a tailored strategy for cervical cancer detection, considering unique risk factor profiles, could provide enhanced detection and prevention capabilities compared to the existing standardized approach [30].

The study presented by [13] explored the application of extremely developed machine learning techniques, including robust ensemble methods and interpretable black box models, for the development of a cervical cancer risk assessment tool. The information from a cohort between 2015 and 2018, a total of 858 women in the Tuscany region of Italy who participated in cervical cancer screening, were analyzed to inform the study's findings. The researchers used a combination of machine-based learning and prediction frameworks, including Random Forest, Gradient Boosting, Support Vector Machine, and Artificial Neural Networks, to create a predictive pattern for cervical cancer risk. The researchers used LIME and SHAP techniques to explain the predictions made by complex machine learning models, increasing model transparency. The study found the merged predictive framework, combining Random Forest, Gradient Boosting, and Artificial Neural Networks, demonstrated excellent predictive performance, with a precision of 0.903 and an AUC of 0.934 for the threat of Cervical Cancer Risk Prediction Modeling [14]. The research illustrated the importance of transparent and interpretable machine learning models in enabling the interpretation of machine learning models and their predictions, which is essential for building trust in AI systems. The research concludes that the combination of reliable and resilient ensemble learning approach models and explainable artificial intelligence methods are crucial for understanding complex models' promise for improving the precision, reliability, and transparency of predictive analytics for cervical cancer susceptibility. The study's results have significant consequences for enhancing Cervical cancer early detection and prevention initiatives.

Research conducted by [15] showed the application of machine learning in developing a reliable predictive model for lung adenocarcinoma outcomes. The researchers analyzed information from a substantial patient cohort using machine learning techniques, revealing essential features associated with prognostic outcomes, including tumor measurements, histological category, and lymph node condition. The researchers successfully developed a prognosis prediction model using the identified features, achieving outstanding accuracy in predicting patient outcomes, including long-term and recurrence-free survival outcomes. The research emphasizes the benefits of machine learning in optimizing cancer prognosis prediction precision and enabling personalized treatment planning, ultimately leading to better patient outcomes. The research by [20] explored the capabilities of machine learning algorithms, such as artificial neural networks, support vector machines, decision trees, and random forests, in improving cancer forecasting and outcome prediction, with a focus on their application in various cancer types. The study also emphasizes the significance of feature identification and dimensionality reduction in machine learning cancer diagnosis and the benefits of integrating machine learning in conjunction with other cutting-edge technologies, such as multimodal analysis of genomic and imaging data, to enhance precision. In their conclusion, the authors provide a forward-looking perspective on the difficulties and possibilities that lie ahead. Researchers proposed an algorithm powered by deep learning technology for predicting the risk of developing lung cancer in patients after undergoing low-dose CT screening. The method combined clinical and imaging characteristics; the Model was trained using a Comprehensive dataset of low-dose CT images [30]. The algorithm showed highly accurate predictions of lung cancer risk, decreasing unnecessary follow-up screenings.

3. METHODOLOGY

The technique involves a structured process for dataset extraction and interpretation of insights, as described in Fig 1, ultimately yielding a robust and reliable machine-learning pattern. The process begins with data pre-processing, converting information to a numerical structure, handling missing numbers, and generating a newly defined response variable. To reduce dimensionality and improve model interpretability, Feature Extraction is performed, which involves selecting the most impactful attributes and eliminating unnecessary or immaterial features. The training consists of building a classification model pipeline, increasing the sample size to mitigate addressing skewed class distribution, and utilizing cross-validation techniques to assess classification performance. To identify the most accurate model, assessing model quality entails evaluating the performance of different models by comparing them with multiple evaluation metrics, including ROC-AUC, accuracy, actual positive rate, and F1-measure.

3.1. Dataset

First, we import the dataset (https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors) under consideration and transform the information into a quantitative format to facilitate analysis. The next step involves verifying the dataset for any null or missing values. Given numerous null values in the dataset, it is essential to impute them to enhance model accuracy, reduce data bias, and boost overall data integrity. We have completed the categorization of predictor variables and response variables, which will guide the imputation of missing values.

Subsequently, the input features comprised categories founded on their attribute types, separating nominal from measurable variables. The data's midpoint value with no missing values was used to impute missing quantitative values in features containing continuous information. Similarly, for features with discrete data, the most frequent value among non-missing data was obtained to impute the missing numbers. To provide a comprehensive view of cancer status, we generated a new column by aggregating the values from 'Dx: Cancer' and 'Dx: CIN,' acknowledging the precancerous nature of CIN. This approach took the dominant value among complete data points. Positive cases in the 'Dx: Cancer' column were relatively low, with only 18 instances, and the 'Dx: CIN' column had nine positive cases. Combining the 'Dx: Cancer' and 'Dx: CIN' columns resulted in 27 patients diagnosed with cancer. Consequently, this would lead to a more precise evaluation of the disease's associated risk factors. The 'Cancer status' column is ultimately selected as the target variable, serving as the focal point for our investigation. The dataset has been meticulously cleaned, and as a result, it is now prepared for additional processing, including information preprocessing and feature selection. Through rigorous data cleaning and validation, we have guaranteed the statistics' accuracy, completeness, and reliability, providing a groundwork for advanced analysis and interpretation. The dataset has been thoroughly refined through the elimination of null values, enabling the application of sophisticated data analysis techniques, feature engineering methods, and predictive modeling approaches to uncover hidden insights and relationships.

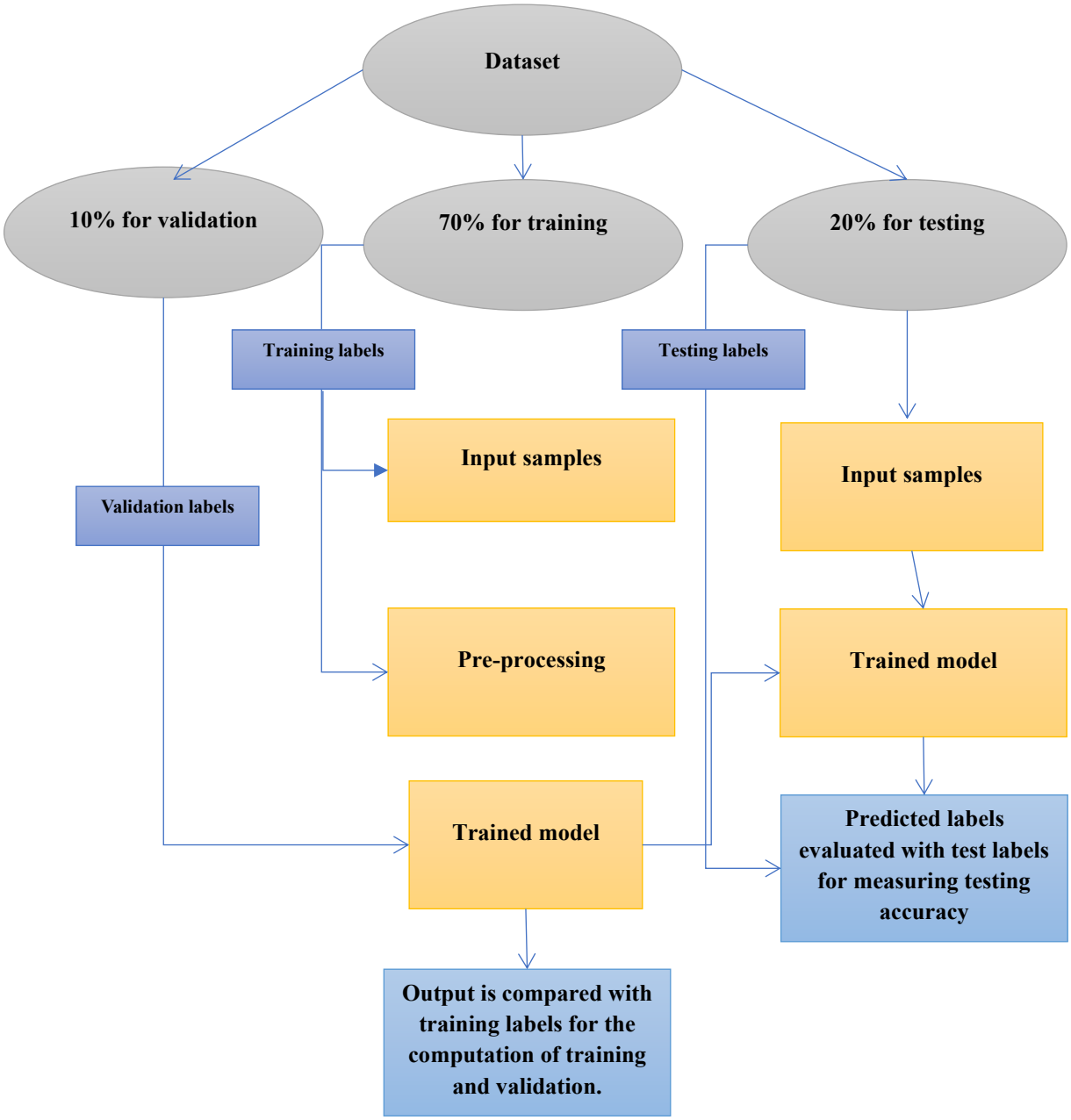

Fig 1: Experimentation workflow

3.2. Feature extraction

Constructing a Gray-Level Co-Occurrence Matrix (GLCM) for samples with a single channel is possible. The GLCM is a square matrix where the dimensions of the matrix are comparable to the number of distinct tonal values present in the original samples. The component at position (i, j) delivers a quantifiable indicator of the co-occurrence frequency of pixels with gray levels i and j, respectively, offering valuable information about the texture and structural properties. The neighbor relationship between pixels, characterized by magnitude and orientation, must be determined beforehand for GLCM construction to represent spatial dependencies accurately. A scaling method was used to normalize the GLCMs, dividing each matrix element by the total sum and yielding a normalized matrix with a unity sum. This normalization step yields the relative frequency $p(i, j)$ for each pair of gray levels, enabling the calculation of Second-order statistical texture features derived from the gray level co-occurrence matrix. These characteristics serve as descriptive metrics for the surface patterns within the source data and were further expanded as a set of 14 distinct features. Given that these features are not entirely orthogonal, a common approach is to select a subset of features for analysis, as discussed. This study utilized four specific characteristics obtained from the GLCM: Texture attributes including contrast, correlation, energy, and homogeneity. The contrast feature K is first introduced, quantifying the gray-level difference between neighboring pixels.

$$K = \sum_{i,j} (i-j)^2 p(i,j) \tag{1}$$

The contrast value ranges from 0, indicating a constant to a maximum value of $(G-1)^2$, where G represents the entire spectrum of gray levels within the samples. Additionally, the definition of correlation R is:

$$R = \sum_{i,j} \frac{(i-j)^2 p(i,j)}{\sigma_i \sigma_j} \tag{2}$$

Here, R is a statistical metric that measures the correlation between adjacent pixel intensity values to detect data patterns and relationships. The variables μ_i and μ_j represent the expected values of the rows and columns, correspondingly, in the GLCM. The variables σ_i and σ_j represent the dispersion measures of the elements in i and j, respectively, offering insight into the variability of GLCM. The calculation of energy E, alternatively known as angular second moment, involves summing the squares across all entries in the GLCM, resulting in a measure that characterizes the texture's uniformity and homogeneity.

$$E = \sum_{i,j} p(i,j)^2 \tag{3}$$

The energy feature measures the uniformity of the data, spanning from 0 to 1 in value, where a value of 1 represents a perfectly uniform sample. Fourth, Homogeneity H is defined as in Eq. (4):

$$H = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{4}$$

Here, H evaluates the proximity of the elements in the GLCM to the main diagonal. Homogeneity measures H Varies between $0\ and\ 1$, where a score of 1 signifies complete homogeneity, which occurs when the GLCM is diagonal, meaning all pixels neighboring the original samples have uniform values. The GLCM with extracted characteristics was identified and computed from the concatenated generating 2D data using Matlab in conjunction with the Image Processing Toolbox. The co-occurrence analysis was restricted to the tumor region, excluding any interactions between the tumor and the surrounding background. Spatial analysis in three dimensions was not conducted as a result of significant disparity relative to the high resolution within the plane (0.78 mm) and the relatively coarse data resolution per slice (6 mm).

3.3. Feature selection

Variables were ranked and selected based on their contribution to the dataset's overall variance, following the method described below:

1) calculating the total variance (T_0) of the entire set of available variables and
2) specifying the desired level ($L \leq 100\%$) of variance explained for the optimal subset of variables.
3) Select the remaining variables with the highest variance as the following variable to include.
4) Map the remaining data unselected attributes onto the orthogonal dimension complement of the reduced feature space formed through the action of determined variable(s).
5) Evaluate the entire data dispersion (T_r) of the remaining projected variables and compare it with the original total variance (T_0). If the variance has decreased ($T_r < T_0$), return to step 1 to choose an additional variable.
6) The concluding set of the identified variables was utilized as predictors in the succeeding phase of categorization. To evaluate the impact of explained variance, four different levels ($L = 100\%, L = 99\%, L = 95\%, and\ L = 90\%$) were examined using the aforementioned Variable selection technique. Feature selection

technique was applied to identify the most discriminative attributes from 21 statistical features and 16 GLCM texture features. However, it was not used for the four isotropic GLCM features or two clinical factors.

3.4. Classification with Multi-tier SVM ($mt - SVM$) classifier

The $mt - SVM$ classification problem is a multi-class classification task. The model uses a hyperplane to segregate the cervical cancer dataset is marked as $\{+1, -1\}$, by identifying the optimal boundary that maximizes the separation across positive and negative sample categories. The pattern effectively distinguishes the classification boundary in the cervical cancer feature realm, aligning with Cervical cancer input features data on a curved classification boundary. Further details concerning the $mt - SVM$ method is can be found. Given a set of N distinct cervical cancer data points are denoted as (fe_i, y_i), fe_i is a vector with D attributes and y_i is a d-dimensional response variable.

$$\sum_i a_i K\,(fe, fe_i) we_i + b;\;\; i \in [1, N] \tag{5}$$

Here, $mt - SVM$ model is formulated such that the kernel function $K(fe, fe_i)$ is utilized and the parameters α and b are introduced, where α corresponds to the coefficient a, b represents the decision boundary of the $mt - SVM$ weighting approach based on distance is a widely recognized and established interpolation technique. The distance-based weighting approach utilizes the similarity of points to calculate the weights used in interpolation. This approach enables the prediction of unknown cervical cancer locations through weighted consideration of nearby positions of existing data points. The distance weighing model calculates the interval between the unknown cervical cancer site and established coordinates.

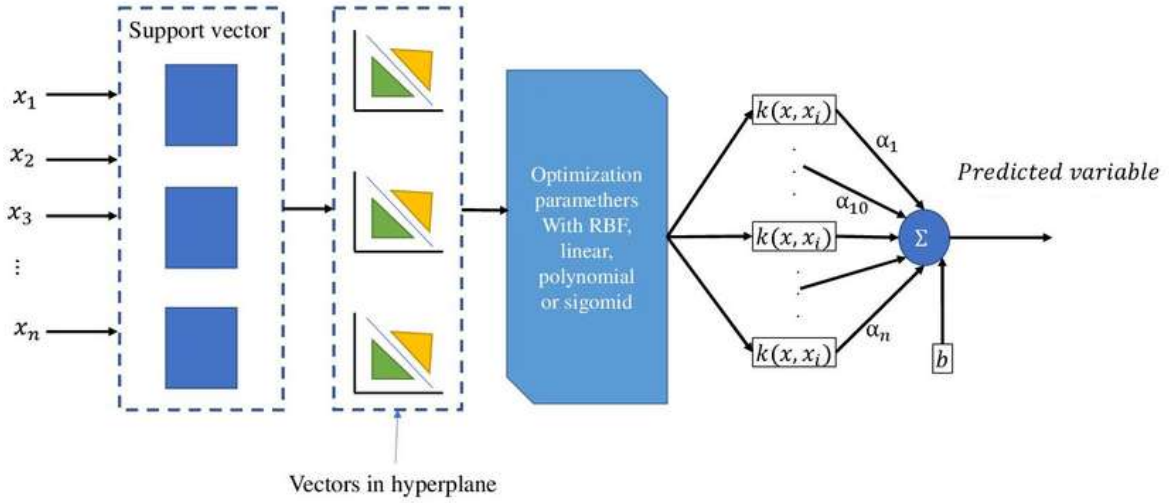

Fig 2: Proposed model

Additionally, the semi-variance value is utilized to compute the weight value. The effectiveness within a collective framework predominantly depends on the heterogeneity or individual models capable of comprising it. In our suggested framework, we utilize a large share of the voting methods to combine the results from various classification models. This approach aggregates the outcomes and selects the optimal decision through voting. In our proposed $mt - SVM$ model, we employ a majority voting approach. The outcome method merges the predicted outcomes from numerous classifiers and uses a poll outcome system to determine the optimal decision. This approach allows the model to leverage the strengths of individual classifiers and produce a more accurate outcome. In essence, the voting method aggregates the results from different classifiers and selects the most popular choice through a voting process. The objective is to divide the cancer samples into uniform segments, ensuring that the smaller class is adequately represented. This process is iterated until the desired sample size is achieved, ensuring equal representation. Although the process ensures equal-sized samples, the original records remain distinct, with the possibility of duplicated records across different subsets.

3.5. Training process

The training set, which accounts for 75% of the overall data is employed to train the Model. By evaluating the importance of each feature, we were able to select the most critical ones and remove the redundant features from the dataset, thereby enhancing data quality and model performance. We employ a model pipeline to identify the most suitable Model, which is then trained to achieve high precision and provide accurate predictions along with risk percentages. Our proposed $mt - SVM$ model explicitly targets the data/class imbalance issue, commonly referred to as the Over-representation problem. Class imbalance arises in binary classification problems when one class has a substantially smaller number of instances than the other. Under these circumstances, a model's performance may suffer, as it tends to favor the dominant class, leading to biased results. Over-representation addresses class imbalance by creating synthetic samples that augment the minority class, promoting a more balanced distribution. Synthetic samples can be made using techniques including random oversampling. Due to the dataset's small sample size and high dimensionality (over 15 features), the model's performance may need to be improved leading to lower accuracy. Model performance is assessed through CV techniques on an independent

information set to address this issue by ensuring a more reliable evaluation. This process involves dividing the available data into multiple subsets are known as folds, training the Model on a subset of these folds, and then evaluating its performance on each of the remaining folds. This approach helps to mitigate Model overfitting on the training data, allows for hyperparameter tuning, and provides a more accurate assessment of the Model's performance. In the derived method, we employ stratified CV, a widely obtained technique for assessing a model's predictive capability on a specific information set. This technique is primarily designed to preserve the class distribution of the original dataset when dividing it into training and validation datasets, thereby ensuring that throughout every cross-validation run, it is representative of the overall dataset. Next, we developed a pipeline of various popular predictive categorization algorithms, including SVM, Random Forest classifier, Naive Bayes classifier, K-nearest neighbor classifier, LightGBM, and Adaboost. To enhance the models' performance, we tuned their model configuration settings, allowing them to become more responsive and adaptable to the distinctive features of the data.

3.5. Evaluation

Following this, a cross-validation evaluation was undertaken, in which the validation dataset, accounting for 25% of the statistics, was used to evaluate and compare the performance of each machine learning pattern, enabling us to identify the ideal predictive Model for risk assessment. To compare model effectiveness, we utilized a range of predictive analytics evaluation metrics, including ROC curves, AUC scores, F1 scores, precision, and recall. The ROC as quantified by the AUC, serves as a holistic standard for measuring model effectiveness across the entire range of thresholds. This performance measure condenses the Model's sensitivity to differences between favorable and unfavorable classes into a single, easily interpretable value. The ROC-AUC score provides a clear benchmark: a score of 0.5 indicates that the Model is performing at a chance level, while a score of 1 represents a flawless model. Accuracy is a performance key metric that assesses the correctness of a model's predictions by evaluating the proportion of correct optimistic predictions. Precision reflects the Model's effectiveness in preventing incorrectly predicted instances, thereby minimizing false alarms. A high precision score of this magnitude indicates that the Model effectively mitigates erroneous positive analysis, ensuring that most of its optimistic predictions are accurate.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The recall is a performance metric that assesses the proportion of correct Successful prediction of positive results relative to the overall count of actual positive instances. Recall Assesses the Model's proficiency in detecting actual instances with positive outcomes, providing insight into its ability to recognize relevant data points. An excellent recall performance signifies that the pattern is highly effective in detecting and identifying the most positive data instances, resulting in few false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

The recall is a performance metric that assesses the percentage of correct the number of correctly predicted positive instances as a fraction of all actual positive instances. It determines the Model's effectiveness in accurately identifying positive instances. A high recall score indicates that the framework accurately identifies the majority of positive cases.

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

Here, F1-score represents a balance between exactness and comprehensiveness, calculated as their harmonic mean. The F1-score provides a concise, single metric that encapsulates a model's performance in terms of accuracy and coverage. A high F1 score suggests that the Model successfully balanced a strong balance between exactness and comprehensiveness, demonstrating excellent overall performance. In summary, ROC AUC, accuracy, sensitivity, and F1-score are crucial evaluation performance indicators that present a complete and nuanced understanding of a machine learning pattern performance. These metrics thoroughly understand the model's discrimination power between positive and negative classes, minimize false positives, and accurately identify true positives.

4. RESULTS AND DISCUSSION

This segment of the analysis discusses the outcomes pertaining to the feature extraction techniques process, random sampling with replacement techniques, and model analysis of results, providing a comprehensive overview of the results. The methodology of extracting informative attributes yielded noteworthy outcomes, highlighting its effectiveness in identifying relevant features. Fig 3 displays the ranking of important features, listed in order of their relevance, from highest to lowest. As evident from the figure, one feature dominates the others in importance, indicating its substantial impact on the Model. Notably, the 'HPV' feature emerges as the most critical factor, aligning with real-world knowledge that individuals with HPV are at a significantly elevated risk of cervical cancer incidence. In addition to HPV, other influential factors such as tobacco use, advancing age, multiple sexual relationships, and total pregnancies also contribute to the outcome, albeit to a lesser extent than

HPV. This finding is consistent with clinical expectations and further validated by a heat map illustration of feature correlations in Fig 4. Following removing non-essential feature optimization and oversampling methods with GLCM, the refined information set facilitates learning a suite of supervised learning algorithms for classification within a unified pipeline. A diverse range of machine learning models are employed. The performance of each Model is evaluated under various conditions, including feature extraction versus no feature extraction, K-folds resampling method for estimating model accuracy, and random replication of the minority class. The assessment outcomes of the models are presented in Table 1 to Table 3, while a visual representation of the results can be found in Fig 5. Although the patterns do not substantially improve accuracy, notable enhancements are observed in AUC, accuracy, sensitivity, and F1 scores. In previous training sessions, linear models like SVM classifiers achieved remarkably high accuracy without employing any techniques, whereas other performance metrics were subpar. This discrepancy was mainly due to the inherent class imbalance present in the dataset. Table 2 reveals that the NB model obtained a recall of 90, indicating excellent sensitivity, but its precision and accuracy were remarkably low, at 0.034 and 0.084, respectively. Following dimensionality reduction and class balancing techniques, the NB model notably improved accuracy, reaching 0.732 and precision, achieving 0.837, although sensitivity decreased to 0.587. Likewise, gradient boosting's accuracy did not substantially increase, but notable improvements were observed in AUC, accuracy, sensitivity, and F1 score, which is beneficial for assessing the probability of cervical cancer occurrence. The following calculation can be employed to determine a patient-specific risk probability. A quantified risk assessment score is calculated utilizing the proposed $mt - SVM$ model, as it has been identified as the top-performing Model integrated within the workflow.

$$Probability\_score = \frac{M + 1}{N + K} \tag{8}$$

The following notations are used: M represents the frequency of instances belonging to a particular label present in the training dataset, N denotes the overall count number of data points in the training collection, and K signifies the number of classes. For instance, a female with five sexual partners, a smoking habit of 37 packs per year, and three STD diagnoses but without HPV has a relatively low cervical cancer risk of 19%. In stark difference emphasis, if identical conditions, an 18-year-old woman had received an HPV diagnosis, the risk of cervical cancer specific to her would have skyrocketed to 91.1%. Likewise, consider a 45-year-old woman with a single intimate partner and a relatively late initiation of sexual activity at age 30. Based solely on these specific data points, the risk of cervical cancer is extremely low, at 0.02%; however, if the same woman had tested positive for HPV, her risk would surge to 80%. The examples above clearly illustrate that the presence or absence of HPV has a profound impact on an individual's risk percentage. Notwithstanding the dominant influence of HPV, the Model also incorporates and weighs various other factors, including age, number of intimate partners, smoking habits, and more. Consequently, the model offers a balanced perspective on clinical findings and exhibits a high degree of concordance with the pathophysiology of the infections.

As presented in Table 2, the various models employed in this study have demonstrated outstanding performance across the board. Notably, the Gradient Boosting and XGBoost models achieved an accuracy of 98.6%, closely followed by the Random Forest model, which attained an accuracy of 98.7%. A comparative analysis of the results reveals that the derived patterns have surpassed the alternative models' performance, demonstrating superior accuracy, precision, and recall. The author presented a comparative study of machine learning models. Our patterns demonstrated superior performance, surpassing others in accuracy and precision. The author proposed an ensemble model that attained an impressive accuracy of 95% and a precision of 100%, indicating exceptional performance in identifying true positives. Notably, the pattern's sensitivity was 67%, suggesting it failed to identify a significant proportion of positive instances. The author in [25] reported an accuracy of 96.38% using Bayes Net and SVM models needs to improve the performance of the proposed models in this work. The author reported a notably lower accuracy of 83.16% with their ensemble model, which pales compared to the superior performance of the proposed work models. The author proposed a Random Forest model that achieved perfect precision (100%) but had a relatively lower sensitivity of 75%. Similarly, the XGBoost and Decision Tree models proposed by the authors achieved a perfect recall of 100%, but their precision was notably low. A comparative analysis of the various studies, including ours, reveals several commonalities and shared themes.

Table 1: Performance evaluation without feature selection

Methods	Accuracy	ROC	Precision	Recall	F1-score
Gradient model	88	69	88	84	86
XGBoost	86	70	85	86	85
NB	72	27	83	75	67
Adaboost	78	82	77	78	77
DT	53	55	47	60	54
Light-GBM	61	87	54	70	62
RF	87	79	89	84	87
SVM	23	86	81	67	72
Linear	63	52	82	69	74

Polynomial model	49	43	90	68	73
MLP	60	76	83	81	89
K-NN	81	56	79	89	82
Logistic Regression	81	36	75	68	72
Proposed $mt-SVM$	**98**	**99**	**98**	**98.5**	**98.6**

Table 2: Performance evaluation with feature selection

Methods	**Accuracy**	**ROC**	**Precision**	**Recall**	**F1-score**
Gradient model	90	71	90	85	86
XGBoost	88	72	86	87	85
NB	74	30	85	76	67
Adaboost	80	85	79	80	77
DT	55	58	50	62	54
Light-GBM	63	89	55	75	62
RF	89	80	90	88	87
SVM	25	88	82	69	72
Linear	65	55	83	72	74
Polynomial model	52	45	92	75	73
MLP	62	75	85	85	89
K-NN	83	58	80	90	82
Logistic Regression	85	40	76	70	72
Proposed $mt-SVM$	**99.2**	**99.5**	**99**	**99.2**	**99.4**

Table 3: Performance evaluation with feature selection

Ref.	**Model**	**Accuracy**	**Precision**	**Recall**
[13]	BayesNet	63	70	90
	Naïve Bayes	66	71	80
	RF	33	80	45
[18]	Ensemble	59	90	67
[25]	DT	45	69	64
	SVM	61	76	61
	C3.5	69	51	62
[26]	Ensemble SVM	36	48	35
	MLP	45	45	61
	LR	34	73	45
[27]	XGBoost	68	80	80
	DT	69	75	70
	MRF	70	80	84
[29]	XGB	98.7	99.9	98.5
	XDT	98.8	99.6	99.7
	RF	98.9	98.6	98.5
Proposed model	**$mt-SVM$**	**99**	**99.2**	**99.4**

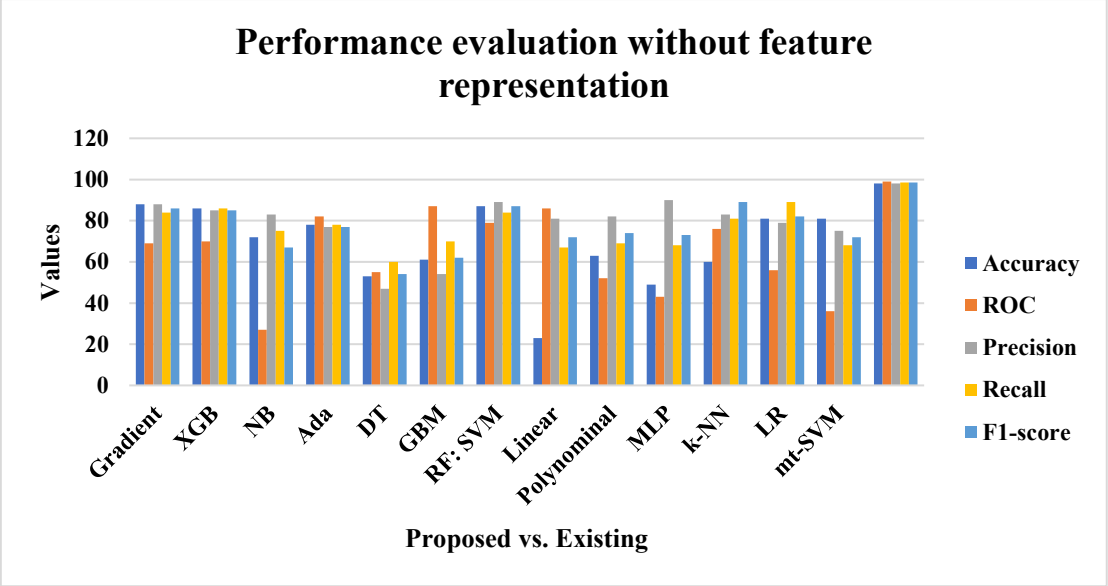

Fig 3: Performance evaluation without feature representation

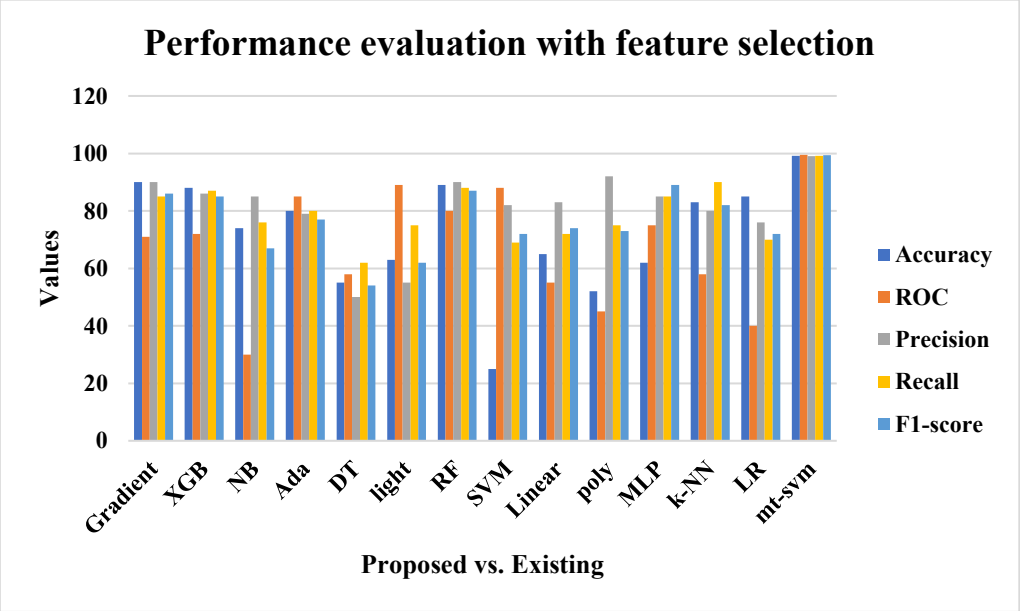

Fig 4: Performance evaluation without feature representation

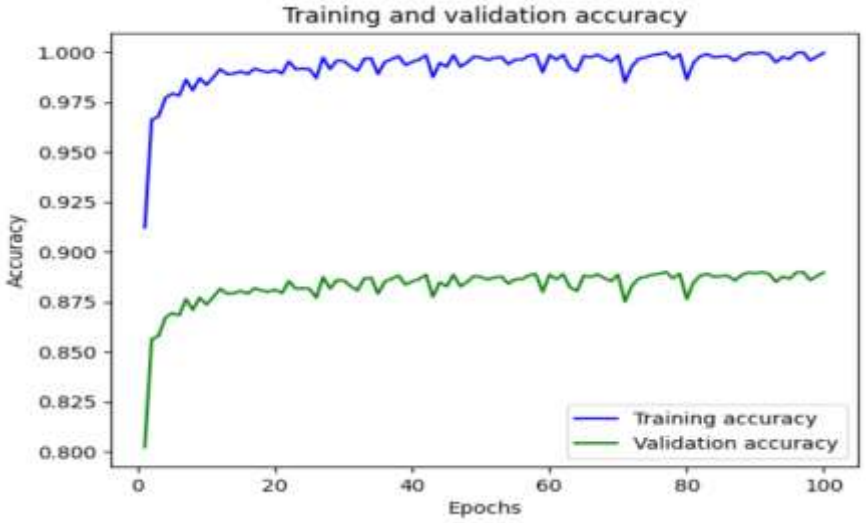

Fig 5: Accuracy analysis

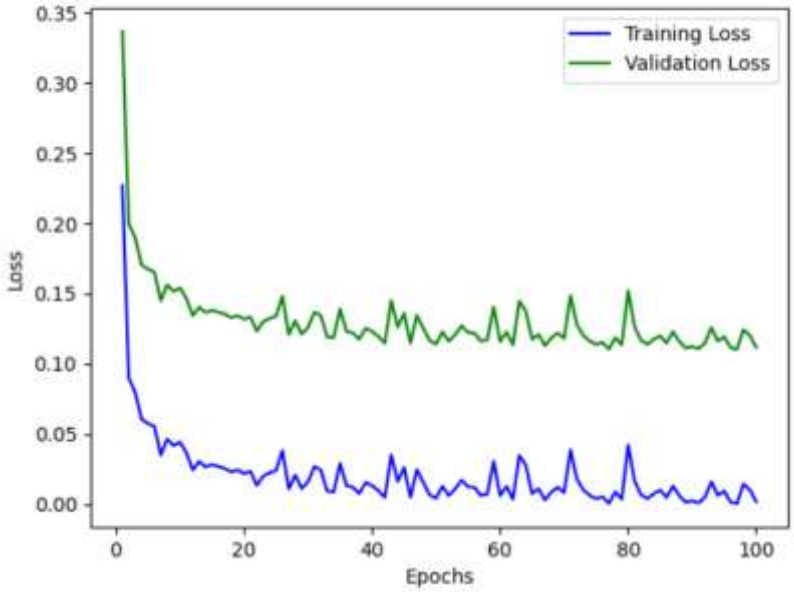

Fig 6: Loss analysis

Notably, research efforts have also focused on mitigating the problem of over-fitting, utilizing K-fold CV to ensure the generalizability of their findings. Consistent with the methodology used, we employed strategies for model hyperparameter optimization to systematically rely on the parameters of our predictive analytics tools and enhance their predictive accuracy. Notably, our study exhibits several unique characteristics that differentiate it from other existing works in the field. In contrast to the ensemble learning and interpretability techniques employed, our approach utilized a more straightforward method, leveraging a correlation heatmap to uncover relationships within the data. Notably, several studies need to pay more attention to employing over-sampling techniques, which can result in overfitting and subsequently lead to biased and unreliable predictive models. Notably, the study incorporated a novel genetic support system, leveraging genetic code mapping data to bolster the accuracy and robustness of their predictions, a distinct approach that diverges from our methodology. Our models demonstrate superior performance across all evaluation metrics, mainly due to our comprehensive approach to addressing potential issues, including overfitting, which can arise from limited and imbalanced training data. Our models achieved a notable balance between accuracy and efficiency by carefully selecting and incorporating only the most relevant features. Our predictive model transcends the limitations of a binary outcome by considering a comprehensive range of risk factors, providing a more nuanced and informative assessment. Integrating these variables enables our model to deliver a more thorough and accurate risk evaluation, quantified as a percentage score. By expressing risk as a percentage, our Model provides a more nuanced and actionable understanding of risk levels are empowering individuals to make informed decisions and adopt targeted lifestyle modifications. Although our research reveals positive results, it is essential to recognize the study's limitations, which may affect the accuracy and reliability of our results. One notable limitation of our research is the limited size of our information set, which may restrict the external validity of our research conclusions. The restricted sample size could have also reduced the quantitative analysis robustness of our findings, making it more challenging to identify meaningful correlations or implications between variables.

5. CONCLUSION

Our proposed
$mt - SVM$ model yields results that facilitate the identification of women at elevated Susceptibility to cervical malignancy, thereby empowering healthcare providers to deliver precision-based early intervention and diagnostic strategies, including Pap smears and HPV screening. Early detection using tailored screening enables timely intervention when the disease is most amenable to treatment and cure. Women identified as high-risk can receive personalized counseling on lifestyle changes, such as smoking cessation and practicing safe sex, to mitigate their risk. Our
$mt - SVM$ Model enables healthcare providers to craft tailored screening and prevention plans catering to each patient's risk profile. Our
$mt - SVM$ model can help reduce unnecessary examinations, streamline the screening process, and enhance its overall effectiveness by optimizing screening strategies. While our
$mt - SVM$ findings have shed new light on the variables linked to cervical carcinoma development, additional studies are warranted to explore further and refine our understanding of this multifaceted disease. Future research endeavors will focus on validating an in-depth exploration of our research findings and more heterogeneous datasets to enhance the broader applicability of our results. A dataset is deemed substantial when it comprises at least 1000 data points or more, contingent upon the specific research context and requirements. Securing a high-

quality dataset is crucial for developing a robust model and providing meaningful insights into the underlying factors and relationships.

REFERENCES

[1] Mao, X. Qin, L. Li, J. Zhou, M. Zhou, X. Li, Y. Xu, L. Yuan, Q.-N. Liu, and H. Xing, "A 15-long non-coding RNA signature to improve prognosis prediction of cervical squamous cell carcinoma," Gynecol. Oncol., vol. 149, no. 1, pp. 181–187, Apr. 2018

[2] Hossain and G. Muhammad, "Deep learning based pathology detection for smart connected healthcare," IEEE Netw., vol. 34, no. 6, pp. 120–125, Nov. 2020.

[3] Wang, P. Xia, L. Wan, L. Zhang, and L. Yu, "Identification of a five-lncRNA signature for predicting the risk of tumor recurrence in breast cancer patients," Int. J. Cancer, vol. 143, pp. 2150–2160, Nov. 2018.

[4] Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," Multimedia Syst., vol. 25, no. 5, pp. 565–575, Oct. 2019.

[5] Amin, M. S. Hossain, G. Muhammad, M. Alhussein, and M. A. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," IEEE Access, vol. 7, pp. 10745–10753, 2019.

[6] Hossain, "Cloud-supported cyber–physical localization framework for patients monitoring," IEEE Syst. J., vol. 11, no. 1, pp. 118–127, Mar. 2017.

[7] Yuan, Y. Yao, B. Cheng, Y. Cheng, Y. Li, Y. Li, X. Liu, X. Cheng, X. Xie, J. Wu, X. Wang, and W. Lu, "The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images," Sci. Rep., vol. 10, no. 1, p. 11639, Dec. 2020

[8] Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," Future Gener. Comput. Syst., vol. 102, pp. 643–649, Jan. 2020.

[9] Geeitha and M. Thangamani, "Integrating HSICBFO and FWSMOTE algorithm-prediction through risk factors in cervical cancer," J. Ambient Intell. Humanized Comput., vol. 12, no. 3, pp. 3213–3225, Mar. 2021.

[10] Paik, M. C. Lim, M. H. Kim, Y. H. Kim, E. S. Song, S. J. Seong, D. H. Suh, J. M. Lee, C. Lee, and C. H. Choi, "Prognostic model for survival and recurrence in patients with early-stage cervical cancer: A Korean gynecologic oncology group study (KGOG 1028)," Cancer Res. Treat., Off. J. Korean Cancer Assoc., vol. 52, no. 1, p. 320, 2019

[11] Mao, L. Dong, Y. Zheng, J. Dong, and X. Li, "Prediction of recurrence in cervical cancer using a nine-lncRNA signature," Frontiers Genet., vol. 10, p. 284, Apr. 2019

[12] Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," IEEE Internet Things J., vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[13] Gangeh, H. Zarkoob, and A. Ghodsi, "Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 14, no. 1, pp. 167–181, Jan. 2017

[14] Pérez-Suay and G. Camps-Valls, "Sensitivity maps of the Hilbert– Schmidt independence criterion," Appl. Soft Comput., vol. 70, pp. 1054–1063, Sep. 2018.

[15] Sørensen and M. Nielsen, "Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination," J. Neurosci. Methods, vol. 302, pp. 66–74, May 2018.

[16] Zhao, D. Wang, R. Yan, K. Mao, F. Shen, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," IEEE Trans. Ind. Electron., vol. 65, no. 2, pp. 1539–1548, Feb. 2018.

[17] Quach, K. Luu, C. N. Duong, and M. Savvides, "Reformulating level sets as deep recurrent neural network approach to semantic segmentation," IEEE Trans. Image Process., vol. 27, no. 5, pp. 2393–2407, May 2018.

[18] Chen, J. Yang, L. Hu, M. S. Hossain, and G. Muhammad, "Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators," IEEE Commun. Mag., vol. 56, no. 11, pp. 14–20, Nov. 2018.

[19] Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," Inf. Fusion, vol. 49, pp. 69–78, Sep. 2019

[20] Lahoura, H. Singh, A. Aggarwal, B. Sharma, M. A. Mohammed, R. Damaševičius, S. Kadry, and K. Cengiz, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," Diagnostics, vol. 11, no. 2, p. 241, Feb. 2021.

[21] Hussein, M. A. Burhanuddin, M. A. Mohammed, M. Elhoseny, B. Garcia-Zapirain, M. S. Maashi, and M. S. Maashi, "Fully automatic segmentation of gynaecological abnormality using a new violajones model," Comput., Mater. Continua, vol. 66, no. 3, pp. 3161–3182, 2021.

[22] Husham, A. Mustapha, S. A. Mostafa, M. K. Al-Obaidi, M. A. Mohammed, A. I. Abdulmaged, and S. T. George, "Comparative analysis between active contour and otsu thresholding segmentation algorithms in segmenting brain tumor magnetic resonance imaging," J. Inf. Technol. Manage., vol. 12, pp. 48–61, Dec. 2020

[23] Ghani, M. A. Mohammed, N. Arunkumar, S. A. Mostafa, D. A. Ibrahim, M. K. Abdullah, M. M. Jaber, E. Abdulhay, G. Ramirez-Gonzalez, and M. A. Burhanuddin, "Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques," Neural Comput. Appl., vol. 32, no. 3, pp. 625–638, Feb. 2020.

[24] Al Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," Sensors, vol. 22, no. 11, p. 4132, May 2022

[25] Adem, S. Kiliçarslan, and O. Cömert, ‘‘Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification,’’ Expert Syst. Appl., vol. 115, pp. 557–564, Jan. 2019.
[26] Ali, K. Ahmed, F. M. Bui, B. K. Paul, S. M. Ibrahim, J. M. W. Quinn, and M. A. Moni, ‘‘Machine learning-based statistical analysis for early stage detection of cervical cancer,’’ Comput. Biol. Med., vol. 139, Dec. 2021, Art. no. 104985.
[27] Ijaz, M. Attique, and Y. Son, ‘‘Data-driven cervical cancer prediction model with outlier detection and over-sampling methods,’’ Sensors, vol. 20, no. 10, p. 2809, May 2020.
[28] Rothberg, B. Hu, L. Lipold, S. Schramm, X. W. Jin, A. Sikon, and G. B. Taksler, ‘‘A risk prediction model to allow personalized screening for cervical cancer,’’ Cancer Causes Control, vol. 29, no. 3, pp. 297–304, Mar. 2018.
[29] Curia, ‘‘Cervical cancer risk prediction with robust ensemble and explainable black boxes method,’’ Health Technol., vol. 11, no. 4, pp. 875–885, Jul. 2021.
[30] Huang, C. T. Lin, Y. Li, M. C. Tammemagi, M. V. Brock, S. Atkar-Khattra, Y. Xu, P. Hu, J. R. Mayo, H. Schmidt, M. Gingras, S. Pasian, L. Stewart, S. Tsai, J. M. Seely, D. Manos, P. Burrowes, R. Bhatia, M.-S. Tsao, and S. Lam, ‘‘Prediction of lung cancer risk at follow-up screening with low-dose CT: A training and validation study of a deep learning method,’’ Lancet Digit. Health, vol. 1, no. 7, pp. e353–e362, Nov. 2019.