ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0365

ANALYSIS AND REVIEW ON FEATURE SELECTION AND CLASSIFICATION METHODS ON CERVICAL CANCER

Anjali Kuruvilla and B. Jayanthi
School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, India

Abstract

Cervical cancer is one of the most widely recognized gynecologic malignancies on the world and it is demanding since this malignant growth happens with no signs. As per World Health Organization (WHO), cervical cancer is the fourth most recurrent disease which is higher death rate that influenced women everywhere in the world. It has demonstrated that early discovery of any cancer when followed up with suitable diagnosis and treatment can expand the patient survival rate. But the existing techniques have problem with imbalanced dataset and feature selection-based classification accuracy. To conquer the previously mentioned issues, the existing strategies are analyzed different procedures of data mining and feature selection techniques which can be applied to bring out hidden information from the cervical cancer dataset. In this review, classification process and feature selection-based classification are performed to improve the given cervical cancer dataset accuracy significantly. In the classification process, the imbalanced data and redundant features are not handled effectively. Hence the feature selection-based classification is required to improve the cervical cancer classification accuracy. This survey is also analyzed the merits and shortcomings of each method applied to application. The comparative analysis is done using various classification techniques like Support Vector Machine (SVM), K Nearest Neighbor (KNN), Convolution Neural Network (CNN) and Synthetic Minority Oversampling Technique + Random Forest with Recursive Feature Elimination (SMOTE+RFE+RF) approach. The experimental result shows that the SMOTE+RFE+RF approach provides better performance in terms of higher accuracy, specificity, Positive Predicted Accuracy (PPA) and Negative Predicted Accuracy (NPA) and sensitivity rather than the other existing methods.

Keywords:

Cervical Cancer, Imbalanced Data, Classification, Early Detection, Machine Learning, Feature Selection

1. INTRODUCTION

Cervical cancer growth wherein strong cells on the outside of the cervix outgrow control shaping a mass of cells named a cancer, which would then be able to extend to different body parts. After breast malignancy, it is the second most frequent disease

Amongst woman around the world, and it is perhaps the most avoidable tumors with 90% of cases recognizable and treatable in its beginning phases [1]. Every year approximately 8.2 million individuals expire from disease that is 13% of absolute passings around the world. In 2017, just 26% of under non-industrial nations detailed having screening administrations accessible for open. In 90% developed nation treatment services are accessible contrasted with under 26% of non-developed nations. The normal malignant growth frequencies can attain up to 22 million over 2030 [2]. Lung and breast cancer cause a substantial number of early deaths in women, but cervical disease is the most dangerous since it is simply studied in women.

As the software engineering and data innovation yield development, explores on examining clinical datasets like diabetes, cervical malignancy and liver illness and so on likewise growth. According to the WHO, complete cervical disease control incorporates essential counteraction (immunization alongside HPV), optional anticipation (screening and therapy of pre-carcinogenic sores), tertiary prevention (diagnosis and therapy of intrusive cervical malignancy) and palliative consideration. Itis at the auxiliary screening stage that this investigation is to be utilized [3]. Cervical disease frequently has no side effects in the beginning phases, yet the normal manifestations that happen are uncommon vaginal dying, which happens in the wake of engaging in sexual relations among periods [4].

The significant factor of cervical malignancy is screening. Four screening strategies are there consist of cervical cytology likewise named as Pap smear test, biopsy, Schiller and Hinslemann [5]. The cytology screening procedure is a microscopic study of cells scraped from the cervix that is used to detect malignant or precancerous cervix conditions [6]. Biopsy procedure is a step-by-step process that begins with the identification of a living tissue test for analysis [7]. The Hinslemann test is a visual examination of the cervix that uses the resolution of iodine. Lugol iodine is utilised for visual examination of cervix subsequent to spreading Lugol iodine identification pace of suspect locale over the cervix, it is otherwise called Schiller test [8].

Many cervical cancers begin with the cells in the transformation region. Those cells don't unexpectedly transform into malignancy. Maybe, the common cells of the cervix first continuously make pre-malignant changes that change into disease. Experts use a couple of terms to portray these pre-cancerous transforms, consist of Cervical Intraepithelial Neoplasia (CIN), Squamous Intraepithelial Injury (SIL), and dysplasia. These movements is perceived via the Pap test and handled to hold development back from making. Despite the fact that cervical tumors start from cells with pre-cancerous changes (pre-diseases), simply a part of the woman with pre-malignant cells in the cervix make tumor. Commonly it considers an extended period of time for cervical pre-disease to change to cervical cancer, but it will happen in less than a year.

Classification is the procedure of predicts class labels. It is utilized to classify the information dependent on the training set and the attributes in a grouping feature. Cervical cancer classification is through data-based on predictable learning schemes, wherein only labeled data is used for learning. However, irrelevant and inappropriate features might confuse the classifier and lead to inaccurate results for the given cervical cancer dataset. Hence, the features selection is an essential step finding efficient features (more discriminant) and improving the quality of datasets (superior and quicker results). The accuracy of the cervical cancer classifier not only depends on the classification algorithm but also on the feature selection method. The feature selection with

classification-based solution is used for removing the irrelevant and redundant features from the original cervical cancer dataset.

The present review study is done in the state of art of different techniques on cervical cancer classification. There is several research and methodologies proposed but the cervical cancer classification dataset accuracy is not ensured considerably. Thus, this survey study suggests that the classification and feature selection-based classification algorithms. Therefore, the comprehensive survey study mainly focused on classification of cervical cancer dataset performance through efficient and effective methods.

The organization of survey study is arranged as; section 2 provides the literature review of existing methods, followed by research gap in section 3. Finally, section 4 defines the experimental result and section 5 provides conclusion of the survey study

2. REVIEW WORK

Classification method and feature selection methods are introduced by several authors to solve cervical cancer detection. This section discusses the current and most frequently used techniques of research published in recent times

2.1 REVIEW OF CLASSIFICATION METHODS

Khalilia et al. [9] utilized an Ensemble Learning (EL) method depends on rehashed random sub sampling. This procedure separates the trained information into different sub-examples, when guaranteeing every sub-example is completely adjusted. It analyzed the evaluation of Support Vector Machine (SVM), boosting, bagging and Random Forest (RF) to foresee the possibility of eight ongoing sicknesses. It classified eight infection classes. Generally, the RF learning strategy executed SVM, boosting and bagging regarding the field over the Receiver Operating Characteristic (ROC) and Area under the Curve (AUC). Also, RF provides lower computational complexity for every factor in the classification cycle

Tseng et al. [10] proposed progressed Machine Learning (ML) methods, generally assumed that the best strategy to create goal to an inferential issue of intermittent cervical disease. Generally, medical analysis of intermittent cervical malignant growth depended on doctor medical involvement in different possibility aspects. Because the risk aspects are general classifications, long stretches of medical investigation and knowledge have attempted to distinguish key factors for repeat. In this work, three ML methods such as SVM and C5.0 are applied to discover significant variables to foresee the repeat inclination for cervical cancer. The outcome show that C5.0 approach is the most helpful way to deal with the disclosure of repeat inclination factors

Teeyapan et al. [11] investigated different SVM algorithms, specifically Twin Weighted SVM (TWSVM), and Twin-Hypersphere SVM (THSVM), and analysis their exhibition on cervical cancer cell grouping in 2-class and 4-class situations. The cervical malignancy cell dataset called the LCH dataset utilized in this work is gathered and extricated from Lampang Cancer Hospital in Thailand. The outcomes demonstrate that TWSVM is desirable over SVM and THSVM on the cervical disease cell categorization

Yu et al. [12] proposed multi-classes imbalanced information categorization method depends on example data. This calculation uses the example data estimation to multi-classes imbalanced dataset. Besides, a classifier is contrived to categorize the information. Examinations on IRIS, WINE, and GLASS datasets illustrate that the method provides better outcome to classify the multi-classes imbalanced information

Machmud et al. [13] introduced ML which is utilized as classifier to distinguish the likelihood of Cervix probability depends on activities and its determinant. Early discovery technique is actually open testing. Activities and its determinant are capable as cervix indicator and occasion as prior discovery. Two well-known ML, Naïve Bayes (NB) and Logistic Regression (LR) are utilized as classifier. From the exploratory outcome, both NB and LR are capable as classifier to identify cervix hazard dependent on activities and its determinant along with better AUC and accuracy

Purnami et al. [14] centered to detect cervical cancer endurance dependent on those elements. Different characterizations strategies: Classification and Regression Tree (CART), Smooth SVM (SSVM), three request spline SSVM (TSSVM) are utilized. Because the cervical cancer information is imbalanced, Synthetic Minority Oversampling Strategy (SMOTE) is utilized for taking care of imbalanced dataset. The SMOTE-SSVM technique gave preferable outcome over SMOTE-TSSVM and SMOTE-CART

Wu et al. [15] recommended two improved SVM strategies, SVM- RFE and backing SVM- Principal Component Analysis (PCA) information is addressed via 32 elements and 4 objective factors: Hinselmann, Schiller, Cytology, and Biopsy. Every one of the four targets is analyzed and predicted via the three SVM-based methodologies, individually. Thus, it builds the examination between these three techniques and evaluate about positioning consequence of variables via the ground truth. It is illustrated that SVM-PCA strategy is better than the other methods

Fatlawi et al. [16] improved a characterization method without applying the normal expense for every mistake case might prompt temperamental outcomes. The method relies upon a Decision Tree (DT) algorithm with an expense framework that diverse expense esteems. It contains a greater expense for mistake in cases that have a positive clinical trials as tainted patients yet categorized not contaminated patients. The method gives more exact outcome in both binary class and multi class order. It has a TP rate (0.429) contrasting and (0.160) for classic DT in binary class task

Sophea et al. [17] investigated an ML-based framework for unusual cell recognition just as cell type categorization. The framework depends on HoG attribute extraction technique and SVM algorithm. The outcome directed on Harlev dataset has illustrated an acknowledgment pace of 94.70% for ordinary and strange cell characterization while assuming about that core in every cell is entirely identified. While assuming a genuine situation application through utilizing the nucleus identification technique, the framework files 88.83% of detection rate

Alyafeai et al. [18] developed a completely mechanized pipeline for cervix cancer and cervical discovery from cervigram pictures. The pipeline comprises of two pre-prepared deep learning models for the programmed cervix discovery and cervical tumor arrangement. The principal model recognizes the

cervix area multiple times quicker than cutting edge information driven models while accomplishing an identification precision of 0.68 regarding crossing point of association assess. Self-extricated attributes are utilized continuously model to group the cervix tumors. These attributes are gotten the hang of utilizing two lightweight models dependent on Convolutional Neural Network (CNN). The DL classifier outflanks existing models as far as arrangement precision and speed. The classifier is described via AUC score of 0.82 while arranging every cervix district multiple times quicker. At long last, the pipeline precision, speed and lightweight design make it proper for cell phone arrangement. Such organization is required to radically improve the early discovery of cervical disease in non-developed nations

Gan et al. [19] proposed cost-sensitive characterization calculation to increase the detection performance. The calculation utilizes variable misclassification rate dictated through tests circulation likelihood to prepare classifier, at that point executes arrangement for imbalanced information in clinical analysis. The adequacy of approach is inspected on the Cleveland heart dataset (Heart), Indian Liver Patient Dataset (ILPD), Dermatology dataset and Cervical malignant growth hazard factors dataset from the UCI learning store. The exploratory outcomes show that the calculation performs better than other strategies

Ilyas et al. [20] introduced an ensemble classifier technique dependent on majority voting in favour of a precise conclusion tending to the patient ailments or manifestations. The examination analyzes a wide scope of accessible classifiers, to be specific DT, SVM, RF, K-Nearest Neighbor (KNN), NB, Multiple Perceptron (MP), J48 Trees, and LR algorithms. The examination documentations an improvement in detection precision of 94% that beats the forecast accuracy of single characterization techniques checked on the equivalent benchmarked datasets. Consequently, the method gives a second assessment to health specialists for sickness recognizable and convenient therapy

Yu et al. [21] proposed four grouping methods and the main method is a 10-layer CNN. The next model (CNN + SPP) is a development of the basic model with a Spatial Pyramid Pooling (SPP) layer to treat cells depending on their sizes. The commencement module replaced the CNN layers in the third model, which was depending on the primary model. In any case, the fourth model (CNN + commencement + SPP) combined the SPP layer and the beginning module addicted to the main model. The testing results exhibited that the fourth model yields the greater accuracy. The inferences of existing classification methods are discussed in Table.1.

Table.1. Inferences on existing classification techniques

Author	Methods	Merits	Demerits
Khalilia et al. [9]	EL, SVM, boosting, bagging and RF methods	It provides lower computational complexity for every factor in the classification cycle	However, it has issue with overfitting
Tseng et al. [10]	ML, SVM and C5.0 methods	This approach increases the classification accuracy	In few cases it has problem with missing attributes
Teeyapan et al. [11]	THSVM method	It provides higher specificity and accuracy	It does not execute well while the data set has more noise
Yu et al. [12]	Multi-classes imbalanced information categorization method	This method provides better outcome to classify the multi-classes imbalanced information	It has issue with number of iterations
Machmud et al. [13]	ML, NB and LR methods	It improves high dimensionality It provides better AUC and accuracy	But execution time is longer for larger dataset
Purnami et al. [14]	CART, SSVM AND SMOTE algorithms	It produces better solutions in terms of accuracy, sensitivity and specificity	However, the computation cost is very high
Wu et al. [15]	SVM, RFE and PCA algorithms	This method is used to increase the accuracy	This method has the noise issues.
Fatlawi et al. [16]	DT algorithm	It produces higher F-score function and classification accuracy	It takes longer calculation time
Sophea et al. [17]	ML based framework	It doesn't require as much training data. It is highly scalable	In few cases, it provides lower accuracy
Alyafeai et al. [18]	CNN algorithm	It requires only a small number of iterations It improves the early discovery of cervical disease	It has issue with redundant features
Gan et al. [19]	Cost-sensitive classification algorithm	It handles large volumes of data It provides better imbalanced data accuracy	But it has issue with error rates
Ilyas et al. [20]	DT, SVM, RF, KNN, NB and MP methods	It is used for evaluating the quality, speedup, and scalability dataset	However, it is expensive
Yu et al. [21]	CNN framework	It decreases the computation time It yields the greater precision	But it has issue with cost complexity

2.2 REVIEW OF FEATURE SELECTION-BASED CLASSIFICATION METHODS

Machine learning techniques are widely used to solve real world problems; they play an important role in the medical field and disease diagnosis. In machine learning, several numbers of features are available for knowing the cervical cancer risk. Thus, selection of features from dataset becomes difficult task. Feature selection methods have been introduced before selection of features and then classification methods have been introduced to increase classification accuracy.

Mahmoudi et al. [22] discussed about gene selection structure, using wrapper method with neuro-fuzzy methodology for cancer detection. Adaptive Neuro Fuzzy Inference System (ANFIS) isa classifier for chose quality of genes from Particle Swarm Optimization (PSO) or Genetic Algorithm (GA) strategies uses on six datasets of microarray quality articulation information for various malignant growths. This classifier gives the preeminent outcomes for single data of all the datasets and it gives satisfactory results when compared with others classifiers.

Harb et al. [23] introduced utilization of multivariate filters and wrapper method, that utilizes PSO with Correlation Feature Selection (CFS) and PSO with the classifier utilized in the arrangement cycle. The output of attributes is given as contribution to five classifiers. Wrapper element subsets are via the forecast execution of a classification on the offered subset. In contrast to channels, wrappers are utilized to look through all potential subsets of attributes and investigate the common data among attributes. Subsequently selecting classifier, wrapper assesses the grouping execution either through cross-approval or hypothetical execution limits.

Hamed et al. [24] discussed a novel attribute selection strategy for type embedded is introduced and proposed about for certain starter outcomes utilizing existing benchmark datasets. The novel technique is named RFE that operates in a forward style and depends on SVM. The novel strategy is used to five diverse benchmark datasets and it is demonstrated prevalent execution as far as precision and time when contrasted with Filter, Wrapper and other embedded techniques.

Tan et al. [25] recommended an integrative ML technique to deal with examine various gene expression data over cervical malignant growth to distinguish group of hereditary indicators which are related and might ultimately help in the determination of cervical tumors. The integrative examination is made out of three stages: to be specific, (i) gene data investigation of individual dataset; (ii) meta-examination of various datasets; and (iii) selection of features and ML examination. Thus, 21 gene data are distinguished via the integrative ML investigation that containsone unsupervised and seven supervised strategies. A useful examination with GSEA (Gene Set Enrichment Analysis) is implemented on the decide 21-feature selection and demonstrated improvement in a nine-potential gene data signature.

Al-Wesabi et al. [26] introduced assorted detection procedures and illustrates the benefit of attribute determination ways to deal with the best foreseeing of cervical cancer infection. 32 features with 800 are there and 58 examples. Furthermore, this information experiences missing qualities and unevenness information. Thus, under sampling, over sampling and embedded under sampling and over sampling are utilized. Besides, dimensionality decrease methods are needed for developing the precision of the classifier. Consequently, attribute determination strategies are concentrated as they separated into two particular classes, filters and wrappers. This Tree classifier is demonstrated to be valuable in taking care of characterization task with superior execution.

Abdoh et al. [27] focused on utilizing cervical cancer elements to assemble characterization model utilizing RF order strategy with the Synthetic Minority Oversampling Technique (SMOTE) and two elements decrease procedures RFE and PCA. Most clinical informational collections are frequently imbalanced on the grounds that the quantity of patients is considerably less than the quantity of non-patients. In view of the unevenness of the pre-owned informational index, SMOTE is utilized to tackle this issue. Subsequent to looking at the outcomes, it tracks down that the mixed RF method with SMOTE progress the performance of classification

Jain et al. [28] recommended two stage combined method for classification of cancer, incorporating CFS with improved-Binary Particle Swarm Optimization (iBPSO). In this work, this method chooses a lower dimensional group of prognostic attributes to categorize organic examples of double and multi class tumors utilizing NB classifier with delineated 10-crease cross-approval. The iBPSO additionally controls the issue of early combination to the neighborhood ideal of conventional BPSO. Exploratory outcomes are contrasted and seven other notable techniques and the method displayed better outcomes by means on detection accuracy and the quantity of attribute selection.

Ahmed et al. [29] focused the cervical cancer disease elements and track down a method which gives great execution in anticipating cervical cancer. RFE and EL dependent on RF method is utilized to distinguish the variables having a lot of importance in foreseeing cervical malignancy. SVM, MLP, and LR utilized to assess the exhibition. SVM with a gathering approach dependent on RF classifier has higher results than others with the respect to metrics like accuracy, precision, and AUC score for an autonomous test set.

Nithya et al. [30] intended for identifying cervical disease and the dataset utilized. In this work, missing variable is there, repetitive attribute and imbalanced objective classes. Henceforth this examination purposes to deal with these issues through coordinated component choice way to deal with accomplish an ideal element subset. ML-based characterization and wrapper strategies are presented. The subset achieved through this combined methodology is utilized in increased forecast measure. The optimal and an ideal element subset are chosen dependent on the exhibition effectiveness of the classifiers in anticipating the outcomes. Proposed algorithm achieves ideal component subset with exactness in clustering and to give computational complexity to cervical disease diagnosis.

Shukl et al. [31] proposed for acquiring the tiny subset of significant attributes, new filter type of feature selection technique, named ReCFS. The strategy is a blend of both attribute-attribute connection and closest neighbor weighted attributes to track down an ideal subset of attributes to limit relationship among attributes. The adequacy of the attributes via technique is accessed utilizing two classifiers, for example, NB and K-NN on genuine datasets. The inferences of existing feature selection-based classification methods are discussed in Table.2.

Table.2. Inferences on Existing Feature Selection-based Classification Methods

Reference	Methods	Merits	Demerits
Mahmoudi et al. [22]	Wrapper based ANFIS algorithm	It provides better correlation features	However, it has issue with computational expense
Harb et al. [23]	Multivariate filters, wrapper method, PSO and CFS methods	This approach increases the classification accuracy	But it has issue with incomplete dataset
Hamed et al. [24]	RFE, SVM, Filter, Wrapper and other embedded methods	It provides higher specificity, sensitivity and accuracy	Time complexity is still an issue
Tan et al. [24]	Integrative ML technique	This method ultimately help in the determination of cervical tumors	It has issue with imbalanced data
Al-Wesabi et al. [26]	Wrapper filter	It provides superior execution performance	In few cases, it has issue with precision measure
Abdoh et al. [27]	RFE, RF, PCA and SMOTE algorithms	It progresses the performance of classification	It increases the overlapping of classes and it produces additional noise
Jain et al. [28]	CFS with iBPSO algorithms	This method displayed better outcomes by means on detection accuracy and the quantity of attribute selection	This method has time complexity issues.
Ahmed et al. [29]	RFE, EL, SVM, RF, MLP and LR algorithms	It produces better AUC and precision	Sometimes it is slower process
Nithya et al. [30]	ML based framework and wrapper strategies	Computational efficiency	In few cases, it provides lower f-score
Shukl et al. [31]	ReCFS algorithm	It improves the early discovery of cervical disease It removes redundant features	It has issue with AUC

2.3 RESEARCH GAP

In the existing methods, feature selection-based classification is not performed efficiently hence the redundant features still exist on the dataset. Hence weighted-based feature selection with classification algorithm will be used to increase the cervical cancer diagnosis performance. Also, hybrid deep learning-based algorithms are needed to be considered for larger data issues to improve the cervical cancer classification performance more accurately.

3. CONTRIBUTION OF THE WORK

In the existing work, feature selection and classification methods are performed over the cervical cancer dataset. But the imbalanced data problem is not handled efficiently in the previous work. Identifying those features with imbalanced dataset and building a classification model to classify whether the cases are belongs to cervical cancer or not is a difficult task. When using machine learning methods, one class dominates the dataset, which implies with the purpose of the number of samples in one class far out numbers the number of samples in the other classes, ensuing in an imbalanced dataset. In this scenario, the dataset is named an imbalanced dataset, and it misleads the classification and has an impact on the classification outcomes. This problem is solved with SMOTE. SMOTE is an oversampling technique with the purpose of making use of k-nearest neighbours towards synthetically increase the minority class and balance the dataset. SMOTE be able to be explained with the subsequent steps.

Step 1: Selects the feature vector xi and recognize the KNNs x_{knn}.

Step 2: Computes the variation among the feature vector and KNN.

Step 3: Multiplies the variation with a random number among 0 and 1.

Step 4: Adds the output number towards feature vector in the direction of recognize a new point on the column part.

Step 5: Repeats the procedure from 1 to 4 towards discovering the feature vectors.

SMOTE+RFE+RF is introduced which handles imbalanced dataset effectively. Largely medical data sets are frequently imbalanced since the number of patients is much fewer than the number of non-patients. SMOTE is utilised to tackle this problem due to the imbalance of the used dataset. It demonstrates the potential of this technique to detect accurately make use of categorize cervical cancer patients. The Classification and Regression Tree (CART) technique is used to investigate the precise classification of some dependent variables (y) and independent variables (x) as well as the relationship between them. In Random Forest (RF), each tree constructs an independent decision tree by randomly selecting a subset of the dataset. RF continuously splits the particular random subset from the root node to a child node until each tree reaches a leaf node, pruning not included. RF also uses the Recursive Feature Elimination (RFE) algorithm for variable importance grouping. It was employed in gene microarrays with tens of thousands of characteristics. In a linear SVM, they used a back-word selection

strategy. It can also be used in conjunction with other linear classification algorithms.

4. EXPERIMENTAL RESULTS

This section is used to experiment the classifiers via a cervical cancer dataset from University of California at Irvine (UCI) [32]. The dataset includes historical medical records, habits and demographic information for 858 cases with 32 features for each case.

4.1 EVALUATION METRICS

In order to measure the each classifiers results, accuracy, sensitivity, specificity, Positive Predicted Accuracy (PPA) and Negative Predicted Accuracy (NPA) as shows in Eq.(1)-Eq.(5) to evaluate the performance of the classification.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Sensitivity} = TP/(TP+TN) \quad (2)$$

$$\text{Specificity} = TN/(TN+FP) \quad (3)$$

$$\text{PPA} = TP/(TP+FP) \quad (4)$$

$$\text{NPA} = TN/(TN+FN) \quad (5)$$

where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive and *FN* is False Negative

Table.3. Comparison Values for Cervical Cancer Dataset vs. Classifiers

Metrics (%)	PSO + ANFIS	Wrapper +SVM	RFE+ SVM	SMOTE+ RFE+RF
NPA	84.63	86.52	88.78	89.41
PPA	85.36	86.47	87.21	89.78
Sensitivity	82.27	84.92	87.79	89.22
Specificity	80.15	82.41	84.63	86.19
Accuracy	82.35	85.54	88.76	90.17

The Table.3 shows the performance comparison of NPA metric with respect to classifiers. Total number of classifiers is taken in x-axis and the NPA results achieved by classifiers are drawn in y-axis. The SMOTE+ RFE + RF algorithm provides higher NPA whereas PSO+ANFIS, Wrapper+SVM and RFE+SVM algorithms provide lower NPA for the given cervical cancer dataset. Thus, the result concludes that the SMOTE+RFE+RF approach gives higher NPA value of 89.41%, whereas other methods such as PSO+ANFIS, Wrapper+SVM and RFE+SVM gives NPA value of 84.63%,86.52% and 88.78% results.

The Table.3 also shows the performance comparison of PPA metric with respect to classifiers. Total number of classifiers is taken in x-axis and the PPA results achieved by classifiers are drawn in y-axis. The SMOTE+ RFE+ RF algorithm provides higher PPA whereas PSO+ANFIS, Wrapper+SVM and RFE+SVM algorithms provide lower PPA for the given cervical cancer dataset. Thus, the result concludes that the SMOTE+RFE+RF approach gives higher PPA value of 89.78%, whereas other methods such as PSO+ANFIS, Wrapper+SVM and RFE+SVM gives PPA value of 85.36%,86.47% and 87.21% results.

The Table.3 further shows the performance comparison of sensitivity metric with respect to classifiers. Total number of classifiers is taken in x-axis and the sensitivity results achieved by classifiers are drawn in y-axis. The SMOTE+RFE+RF algorithm provides higher sensitivity whereas PSO +ANFIS, Wrapper+SVM and RFE+SVM algorithms provide lower sensitivity for the given cervical cancer dataset. Thus, the result concludes that the SMOTE+RFE+RF approach gives higher sensitivity value of 89.22%, whereas other methods such as PSO+ANFIS, Wrapper+SVM and RFE+SVM gives sensitivity value of 82.27%, 84.92% and 87.79% results.

The Table.3 shows the performance comparison of specificity metric with respect to classifiers. Total number of classifiers is taken in x-axis and the specificity results achieved by classifiers are drawn in y-axis. The SMOTE+RFE+RF algorithm provides higher specificity whereas PSO+ANFIS, Wrapper+SVM and RFE+SVM algorithms provide lower specificityfor the given cervical cancer dataset. The proposed SMOTE+RFE+RF approach gives higher specificity value of 86.19%, whereas other methods such as PSO+ANFIS, Wrapper+SVM and RFE+SVM gives specificity value of 80.15%,82.41% and 84.63% results.

The Table.3 shows the performance comparison of accuracy metric with respect to classifiers. Total number of classifiers is taken in x-axis and the accuracy results achieved by classifiers are drawn in y-axis. The SMOTE+RFE+RF algorithm provides higher accuracy whereas PSO +ANFIS, Wrapper+SVM and RFE+SVM algorithms provide lower accuracy for the given cervical cancer dataset. The proposed SMOTE+RFE+RF approach gives higher accuracy value of 90.17%, whereas other methods such as PSO+ANFIS, Wrapper+SVM and RFE+SVM gives accuracy value of 82.35%,85.54% and 88.76% results.

5. CONCLUSION AND FUTURE WORK

This survey discussed the classification methods, feature selection-based classification algorithms and existing techniques for cervical cancer classification. Diverse feature selection methods and classification approaches are analyzed on the cervical cancer dataset to improve the classification performance. Feature selection is very significant stage to choose the important features especially in the cervical cancer dataset. Then these chosen attributes are carried out to the classification process.

Therefore, the improvement in classification accuracy is increased through the reduction of features can be seen. Several classification methods such as KNN, NB, SVM, CNN, wrapper-based ANFIS, wrapper method with SVM, RFE+SVM are executed on the data set. It also looked at the benefits and drawbacks of each method when applied to a data collection. As a result, this survey will provide a valuable look into present classification solutions, as well as their benefits and drawbacks. Existing algorithms, on the other hand, have difficulty detecting the strengths and shortcomings of their cervical cancer classification, as well as identifying relevant elements. To overcome this issue, SMOTE+RFE+RF approach is introduced for the better accuracy in all magnification factors. It demonstrates the potential of this technique to detect correctly use to classify cervical cancer patients. It identified the top most important features for classification of cervical cancer patients.

The result concludes that the SMOTE+RFE+RF approach provides better performance in terms of higher accuracy, specificity, PPA and NPA and sensitivity than the other existing methods.

In future work, hybrid feature selection with deep learning algorithm can be proposed to progress the cervical cancer classification using their best features.

REFERENCES

[1] N. Kamil and S. Kamil, "Global Cancer Incidences, Causes and Future Predictions for Subcontinent Region", *Systematic Reviews in Pharmacy*, Vol. 6, No. 1, pp. 13-17, 2015.

[2] R.L. Siegel, K. D. Miller and A. Jemal, "Cancer Statistics, 2017", *CA: A Cancer Journal for Clinicians*, vol. 67, pp. 7-30, 2017.

[3] F. Bray and A. Jemal, "Global Cancer Statistics 2018: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *CA: A Cancer Journal for Clinicians*, Vol. 68, pp. 394-424, 2018.

[4] R.A. Kerkar and Y.V. Kulkarni, "Screening for Cervical Cancer: An Overview", *Journal of Obstetrics and Gynecology of India*, Vol. 56, No. 2, pp. 115-122, 2006.

[5] G. Guvenc, A. Akyuz and C.H. Acikel, "Health Belief Model Scale for Cervical Cancer and Pap Smear Test: Psychometric Testing", *Journal of Advanced Nursing*, Vol. 67, No. 2, pp. 428-437, 2011.

[6] S.A Syed, K. Sheela Sobana Rani and G.B. Mohammad, "Design of Resources Allocation in 6G Cybertwin Technology using the Fuzzy Neuro Model in Healthcare Systems", *Journal of Healthcare Engineering*, Vol. 2022, pp. 1-8, 2022.

[7] H. Ramaraju, Y. Nagaveni and A. Khazi, "Use of Schiller Test Versus Pap Smear to Increase Detection Rate of Cervical Dysplasias", *International Journal of Reproduction, Contraception, Obstetrics and Gynecology*, Vol. 5, pp. 1446-1450, 2017.

[8] A. Ghoneim, G. Muhammad and M.S. Hossain, "Cervical Cancer Classification using Convolutional Neural Networks and Extreme Learning Machines", *Future Generation Computer Systems*, Vol 102, pp. 643-649, 2020.

[9] N. Arivazhagan, K. Somasundaram, D. Vijendra Babu and V. Prabhu Sundramurthy, "Cloud-Internet of Health Things (IOHT) Task Scheduling using Hybrid Moth Flame Optimization with Deep Neural Network Algorithm for E Healthcare Systems", *Scientific Programming*, Vol. 2022, pp. 1-9, 2022.

[10] C.J. Tseng, C.J. Lu, C.C. Chang and G.D. Chen, "Application of Machine Learning to Predict the Recurrence-Proneness for Cervical Cancer", *Neural Computing and Applications*, Vol. 24, No. 6, pp. 1311-1316, 2014.

[11] K. Teeyapan, N. Theera Umpon and S. Auephanwiriyakul, "Application of Support Vector-based Methods for Cervical Cancer Cell Classification", *Proceedings of IEEE International Conference on Control System, Computing and Engineering*, pp. 514-519, 2015.

[12] C. Yu, F. Li, G. Li and N. Yang, "Multi-Classes Imbalanced Dataset Classification-based on Sample Information", *Proceedings of IEEE International Conference on High Performance Computing and Communications*, pp. 1768-1773, 2015.

[13] R. Machmud and A. Wijaya, "Behavior Determinant-based Cervical Cancer Early Detection with Machine Learning Algorithm", *Advanced Science Letters*, Vol. 22, No. 10, pp. 3120-3123, 2016.

[14] S.W. Purnami, "Cervical Cancer Survival Prediction using Hybrid of SMOTE, CART and Smooth Support Vector Machine", *AIP Conference Proceedings*, Vol. 1723. No. 1, pp. 1-9, 2016.

[15] W. Wu and H. Zhou, "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches", *IEEE Access*, Vol. 5, pp. 25189-25195, 2017.

[16] H.K. Fatlawi, "Enhanced Classification Model for Cervical Cancer Dataset-based on Cost Sensitive Classifier", *International Journal of Computer Techniques*, Vol. 4, No. 4, pp. 115-120, 2017.

[17] P.R.U.M. Sophea, D.O.D. Handayani and P. Boursier, "Abnormal Cervical Cell Detection using Hog Descriptor and SVM Classifier", *Proceedings of 4th International Conference on Advances in Computing, Communication and Automation*, pp. 1-6, 2018.

[18] Z. Alyafeai and L. Ghouti, "A Fully-Automated Deep Learning Pipeline for Cervical Cancer Classification", *Expert Systems with Applications*, Vol. 141, pp. 1-40, 2020.

[19] D. Gan, J. Shen and N. Liu, "Integrating TANBN with Cost Sensitive Classification Algorithm for Imbalanced Data in Medical Diagnosis", *Computers and Industrial Engineering*, Vol. 140, pp. 1-9, 2020.

[20] Q.M. Ilyas and M. Ahmad, "An Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence Towards Sustainable Health", *IEEE Access*, Vol. 9, pp. 12374-12388, 2021.

[21] S. Yu, X. Feng and X. Huang, "Automatic Classification of Cervical Cells using Deep Learning Method", *IEEE Access*, Vol. 9, pp. 32559-32568, 2021.

[22] S. Mahmoudi and H.R. Kanan, "ANFIS-Based Wrapper Model Gene Selection for Cancer Classification on Microarray Gene Expression Data", *Proceedings of Iranian Conference on Fuzzy Systems*, pp. 1-6, 2013.

[23] H.M. Harb and A.S. Desuky, "Feature Selection on Classification of Medical Datasets-based on Particle Swarm Optimization", *International Journal of Computer Applications*, Vol. 104, No. 5, pp. 14-17, 2014.

[24] T. Hamed, R. Dara and S.C. Kremer, "An Accurate, Fast Embedded Feature Selection for SVMs", *Proceedings of 13th International Conference on Machine Learning and Applications*, pp. 135-140, 2014.

[25] M.S. Tan, S.W. Chang and H.J. Yap, "Integrative Machine Learning Analysis of Multiple Gene Expression Profiles in Cervical Cancer", *PeerJ*, Vol. 6, pp. 1-24, 2018.

[26] Y.M.S. Al-Wesabi, A. Choudhury and D. Won, "Classification of Cervical Cancer Dataset", *Proceedings of Annual Conference on IISE*, pp. 1456-1461, 2018.

[27] S.F. Abdoh, M.A. Rizka and F.A. Maghraby, "Cervical Cancer Diagnosis using Random Forest Classifier with Smote and Feature Reduction Techniques", *IEEE Access*, Vol. 6, pp. 59475-59485, 2018.

[28] I. Jain, V.K. Jain and R. Jain, "Correlation Feature Selection-based Improved-Binary Particle Swarm Optimization for Gene Selection and Cancer Classification", *Applied Soft Computing*, Vol. 62, pp. 203-215, 2018.

[29] M. Ahmed, M.M.J. Kabir and M.M. Hasan, "Identification of the Risk Factors of Cervical Cancer Applying Feature Selection Approaches", *Proceedings of International Conference on Electrical, Computer and Telecommunication Engineering*, pp. 201-204, 2019.

[30] B. Nithya and V. Ilango, "Machine Learning Aided Fused Feature Selection-based Classification Framework for Diagnosing Cervical Cancer", *Proceedings of 4th International Conference on Computing Methodologies and Communication*, pp. 61-66, 2020.

[31] A.K. Shukl and D. Tripathi, "Knowledge Discovery in Medical and Biological Datasets by Integration of Relief-F and Correlation Feature Selection Techniques", *Journal of Intelligent and Fuzzy Systems*, Vol. 20, pp. 1-12, 2020.

[32] K. Fernandes, J.S. Cardoso and J. Fernandes, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening", *Proceedings of International Conference on Pattern Recognition and Image Analysis*, pp. 243-250, 2017.