

Enhanced Precision in Cervical Cancer Diagnosis: A Multi-Stage Machine Learning Fusion Approach

Bhimsingh Bohara¹, Dr. Pramod Kumar²

Research Scholar, School of Science & Technology, The Glocal University, Saharanpur (U.P.)¹

Professor, School of Science & Technology, The Glocal University, Saharanpur (U.P.)²

Abstract: Cervical cancer remains a leading cause of female mortality, particularly in developing regions, underscoring the critical need for early detection and intervention guided by skilled medical professionals. While Pap smear images serve as valuable diagnostic tools, many available datasets for automated cervical cancer detection contain missing data, posing challenges for machine learning models' efficacy. To address these hurdles, this study presents an automated system adept at managing missing information using ADASYN characteristics, resulting in exceptional accuracy. The proposed methodology integrates a voting classifier model harnessing the predictive capacity of three distinct machine learning models. It further incorporates SVM Imputer and ADASYN up-sampled features to mitigate missing value concerns, while leveraging CNN-generated features to augment the model's capabilities. Notably, this model achieves remarkable performance metrics, boasting a 99.99% accuracy, precision, recall, and F1 score. A comprehensive comparative analysis evaluates the proposed model against various machine learning algorithms across four scenarios: original dataset usage, SVM imputation, ADASYN feature utilization, and CNN-generated features. Results indicate the superior efficacy of the proposed model over existing state-of-the-art techniques. This research not only introduces a novel approach but also offers actionable suggestions for refining automated cervical cancer detection systems. Its impact extends to benefiting medical practitioners by enabling earlier detection and improved patient care. Furthermore, the study's findings have substantial societal implications, potentially reducing the burden of cervical cancer through enhanced diagnostic accuracy and timely intervention.

I. INTRODUCTION

Cervical cancer is a cancer that begins in the cervix, an opening of the uterus that links to the vagina. Frequently, it arises from an enduring infection with the sexually transmitted virus called human papillomavirus (HPV). HPV has the ability to cause erratic changes in cervical cell structure. If these changes are not treated, they can eventually lead to cancer [1]. Cervical cancer ranks as the leading cause of mortality among women, following lung and breast cancer [2]. There is a widespread belief that cervical cancer is often considered untreatable when it reaches its advanced stages [3]. There have been notable recent advancements in enhancing this disease's detection through imaging techniques. Based on the information given by the World Health Organization (WHO), cervical cancer is the fourth most commonly diagnosed cancer worldwide. Around 570,000 new cases were recorded in 2018 alone, making up 7.5% of all mortality from female cancer [4].

Roughly 85% of the projected 311,000 yearly cervical cancer-related fatalities are thought to take place in nations with lower- and middle-income economies. Saving lives is greatly aided by early identification of cervical cancer. In comparison to women without HIV, women living with HIV have a six-fold increased chance of getting cervical cancer, with HIV thought to be a factor in 5% of cases overall. The availability of essential equipment, consistent screening techniques, proper supervision, and the quick diagnosis and treatment of discovered lesions are some of the aspects that affect screening efficacy [5].

Squamous cell carcinoma, which accounts for around 70–80% of cases, and adenocarcinoma, which starts from epithelial cells in the cervical canal that secrete mucus, are the two main kinds of cervical cancer [6]. Adenocarcinoma has increased recently, despite the fact that squamous cell carcinoma is more prevalent and currently accounts for 10 to 15% of uterine malignancies. Due to the fact that adenocarcinoma grows in the cervical canal, its screening process is challenging. Depending on the kind and stage of the disease at diagnosis, several treatments and prognoses are available for cervical cancer. Early detection through regular screenings and HPV vaccination can significantly improve outcomes for people at risk.

Cervical cancer encompasses several types, including the most common squamous cell carcinoma and the second most common adenocarcinoma [7]. Each type has unique characteristics, aggressiveness, and treatment considerations, underscoring the importance of regular screenings for early detection and appropriate management. Specifically high-risk strains of the human papillomavirus (HPV) are the main cause of cervical cancer [8]. In HPV-infected women, there are several risk factors that might raise the probability of cervical cancer [9].

Smoking, genital herpes, having several sexual partners, a weaker immune system, poorer socioeconomic position, insufficient genital cleanliness, and a larger number of childbirths are among these risk factors. Cervical cancer symptoms might vary based on the size of the tumour and the stage of the disease. The biggest issue, however, is in the early stage, which frequently has no visible signs and is generally identified inadvertently during normal yearly check-ups. Approximately 90% of patients have evident symptoms in the latter stages [10].

The key sign connected with cervical cancer is irregular vaginal bleeding. Early screening for cervical cancer is essential for detecting precancerous or cancerous changes in the cervix. Common screening methods include Pap smears and HPV tests [11]. However, these screenings have limitations, such as false positives and negatives, age-related factors, limited access to screening in some areas, and the focus on certain types of cervical cancer. Cervical cancer screening typically entails a gynaecological examination, which can cause discomfort and pain for patients [12]. The unease associated with this examination may lead to delays or avoidance, impeding the timely diagnosis of cervical cancer. Because of this, the death rate in these nations is substantially higher, with low-income countries accounting for nearly nine out of every ten cervical cancer-related fatalities [6]. It is crucial to increase cervical cancer screening rates since at an early stage it has relatively good 5-year survival rates, increasing to 90% [13]. Nevertheless, these screening rates vary among countries, with developed nations exhibiting higher rates, while developing countries face alarmingly low screening participation [14]. Cervical cancer prevention strategies differ, but depending on screening tests are inadequate. Early diagnosis is imperative to prevent fatalities from invasive forms of the disease.

Presently, ML, DL, and computer vision are emerging as valuable approaches for diagnosing medical conditions [15]. ML models have significantly improved the results of medical diagnosis [16]. Bhavani and Govardhan [17] proposed stacked ensemble model using SMOTE for predicting cervical cancer. Karamati et al. [18] also applied SMOTE and KNN features and applied multimodel approach for cervical cancer detection. Li et al. [19] used CNN for screening of cervical cancer patients. While recent studies have made significant strides in applying machine learning techniques, such as ensemble models and advanced sampling methods, to predict cervical cancer, there remains a need to comprehensively assess the combined impact of these approaches on improving accuracy, handling missing data, addressing class imbalances, and extracting complex features. Moreover, limited research has specifically evaluated the effectiveness of integrating Convolutional Neural Networks in conjunction with traditional ML methods for cervical cancer prediction. ML algorithms achieve accurate and reliable results by applying diverse preprocessing methods, like data cleansing and feature engineering, to the medical dataset. These discoveries can help medical practitioners diagnose illnesses quickly and give patients the best care possible. This paper employs ML techniques to develop a computer-aided diagnosis (CAD) system for the accurate and prompt identification of cancer. The following significant contributions are provided by this study:

- An ML-driven framework is proposed for predicting cervical cancer in patients. A voting classifier is incorporated into the suggested model to improve prediction accuracy.
- The SVM imputation method is employed to generate artificial missing values, aiming to mitigate the challenge posed by incomplete data.
- ADASYN (Adaptive Synthetic Sampling), an oversampling technique that generates synthetic minority class samples, is applied to address class imbalance issues.
- To handle complex features, this research employs a Convolutional Neural Network (CNN).
- The effectiveness of the suggested method is assessed using four scenarios: the original dataset, the dataset with SVM imputation alone, the dataset with SVM imputation followed by ADASYN, and with CNN generated features dataset.

The paper is organised as follows: An extensive analysis of the current classification methods for cervical cancer diagnosis is given in Section. In Section further detail on the dataset and the suggested cervical cancer detection methodology is provided which employs several classification algorithms and up-sampling methods.

The purpose of Section is to provide the findings of the research and promote discussion. The work is finally concluded in Section which also suggests possible lines of inquiry for further study.

II. RELATED WORK

During the recent years, there has been a surge in the development and application of ML models to accelerate research and innovation in various domains. Numerous investigations have been done in the categorization of cervical cancer [20, 21]. Various studies and their findings are summarized in this section.

ML and DL models are extensively being used in medical diagnostics, including breast cancer diagnosis [22], lung cancer detection [23], colorectal cancer [24], and numerous other healthcare applications [25–27]. Some research works applied DL approaches in various tasks like Alzheimer's diagnosis [28], medical imaging [29], and pathology image segmentation [30]. Medical diagnosis has been upgraded by different tools and techniques like surgical analysis [31], EEG encoding [32], CT imaging [33, 34], and surgical navigation [35].

The discrete wavelet and cosine transform were used by Kalbhor and colleagues in their research study [36] to extract characteristics. They used the fractional coefficient technique to effectively decrease the dimensionality of these characteristics. The reduced characteristics were then fed into seven different machine learning classifiers in an effort to discriminate between various cervical cancer subgroups. In another research investigation conducted by Devi and Thirumurugan [37], They used the C-means clustering technique to divide cervical cells. They then retrieved texture information and used Principal component analysis (PCA) for dimension reduction of the collected data. Following that, scientists used the K-nearest neighbours (KNN) method to categorise the cervical cells, attaining an excellent accuracy rate.

Alquran et al. [38] concentrated on cervical cancer classification. They integrated DL with a cascading SVM to attain precise outcomes. By combining methodologies, they effectively categorized cancer into seven groups, achieving a remarkable accuracy score. In another investigation, Kalbhor et al. [39] developed a novel hybrid approach that included DL and ML models, and a fuzzy network. Technique focused on feature engineering and Pap-smear picture categorization. They used transfer learning models such as AlexNet, GoogleNet, and ResNet. The experimental evaluation was done by utilising well-known datasets. Notably, the greatest classification accuracy was achieved by ResNet-50 architecture. Various other domains applied the advanced techniques [40, 41].

Radiation enteritis (RE) causes treatment intolerance or radiotherapy cessation, which severely lowers the patient's quality of life. The adverse effects of radiation therapy in patients with cervical cancer can be greatly decreased if the RE in patients can be anticipated in advance and focused therapeutic preventative treatment can be implemented. Additionally, the optimisation of the radiotherapy strategy and the choice of a customised radiation dose depend on the precise prediction of RE. Ma et al. [42] investigated the relationship of RE and dose volume in cervical cancer patients. Cancer diagnosis involves invasion [43] and migration of cancer cells [44–46]. The quality life of cervical cancer patient survivors has been analyzed in [47].

Tanimu et al. [48] conducted a research study with a primary focus on identifying risk factors linked to cervical cancer. They employed the decision tree (DT) classification algorithm and leveraged LASSO (least absolute shrinkage and selection operator) feature engineering methods and feature reduction techniques. These methods were used to pinpoint the critical characteristics to identify cervical cancer. The dataset utilized in their research presented challenges such as missing values and significant class imbalance. To tackle these issues, the research team adopted a technique known as SMOTETomek. The outcomes revealed that their proposed approach achieved outstanding accuracy. In a comparative analysis, Quinlan and colleagues [49] evaluated several ML classifiers for cervical cancer categorization. The dataset utilized in their analysis also showed signs of class imbalance, necessitating a method to solve this problem. To address the issue of class imbalance, the researchers used the SMOTE-Tomek in conjunction with a highly tailored RF. The findings showed that when combined with SMOTE-Tomek, the RF classifier attained an extraordinary accuracy rate.

Abdoh and colleagues [50] introduced a system for cervical cancer classification, employing the RF in conjunction with the SMOTE. They also incorporated two feature reduction techniques. Their experiment used a dataset with thirty characteristics.

The study looked at the effect of changing the feature size and discovered that utilizing SMOTE in conjunction with RF and other characteristics produced an excellent accuracy rate. Ijaz and colleagues [51] presented a data-driven approach for the detection of cervical cancer. The solution included both outlier identification and the SMOTE. The challenge was carried out utilizing the RF method with DBSCAN. According to their findings, when applied to a dataset of various features, their proposed method got a reasonable accuracy score.

III. PROPOSED APPROACH

The study made use of a dataset obtained from Kaggle, a recognised platform for publicly available datasets. A range of preprocessing processes were carried out to improve the efficacy of ML models and handle missing data. The SVM imputer was used to deal with missing data values. Following that, the dataset was divided into a 70:30 ratio, with 70% allotted as the train set and 30% as a test set. The suggested system used an ensemble technique known as RF + KNN + LR to identify cervical cancer. Voting classifiers are powerful strategies that aggregate results from many models to improve accuracy and durability. The models in the voting classifier have their own set of strengths and shortcomings, and their combined use results in greater overall performance. In this scenario, the suggested method for detecting cervical cancer incorporates three commonly used algorithms: RF, KNN, and LR. Fig 2 shows a process diagram explaining this strategy.

The voting classifier works by combining predictions from these three different ML methods. The typical strategy for building an ensemble/voting classifier is training numerous classifiers using the dataset and then combining their results. In this case, the RF, KNN, and LR models were all trained individually on the same dataset. Each model predicts the probability of each class inside the target variable. The estimated probabilities are then summed to get a final forecast for each instance in the dataset. A popular method for integrating the predictions is to compute a weighted average of the expected probabilities defined by the performance of each model on a validation dataset.

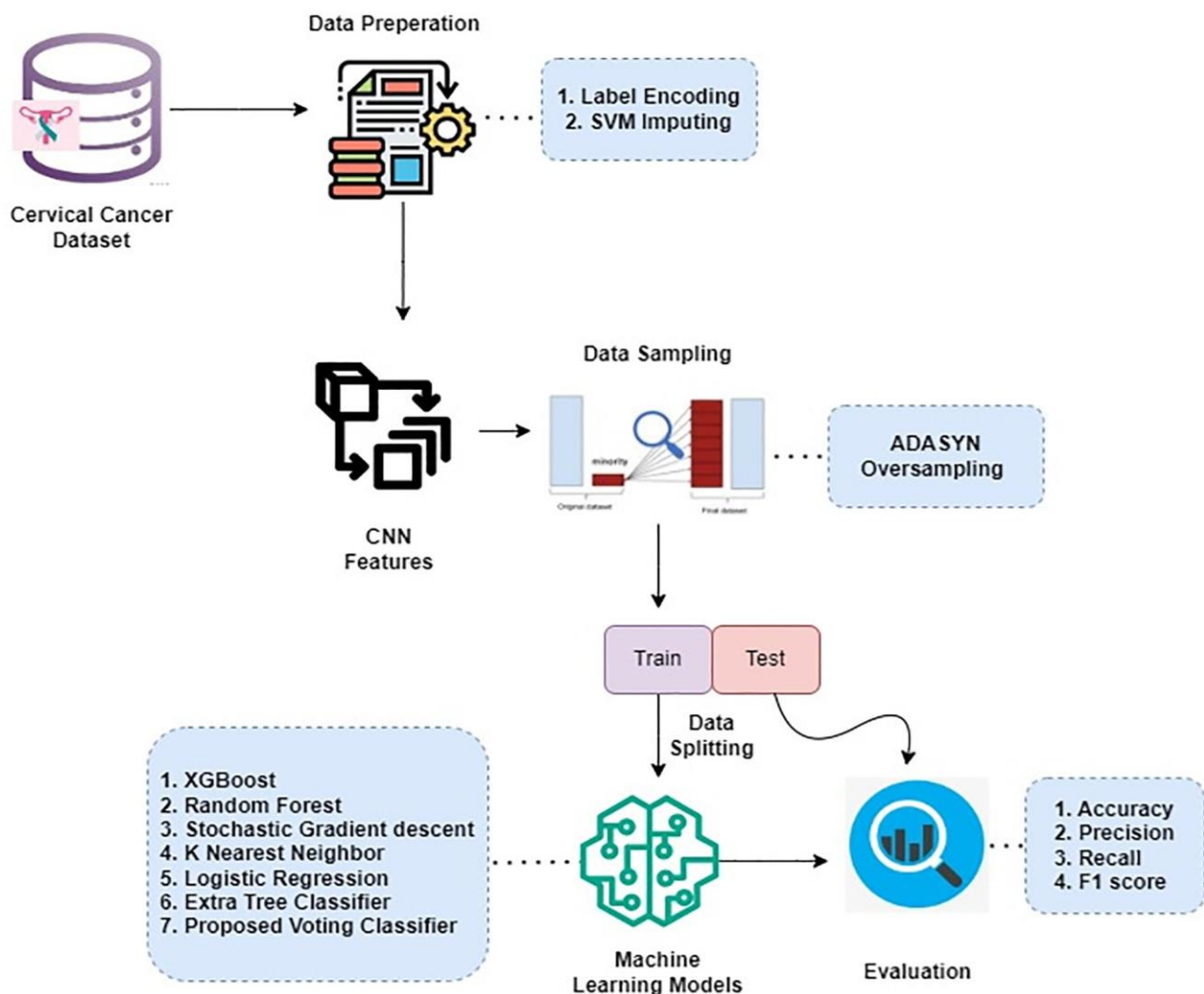


Fig 2. The workflow of the proposed methodology

Algorithm 1 Ensembling of SV-CNN model.

Input: input data δx ; $y \in \mathcal{N}$

MRF = Trained_RF

$i \leftarrow 1$

MKNN = Trained_KNN

MLR = Trained_LR

```
1: for  $i = 1$  to  $M$  do
2:   if MRF  $\neq \emptyset$  & MKNN  $\neq \emptyset$  & MLR  $\neq \emptyset$  & training_set  $\neq \emptyset$  then
3:     ProbRF-1 = MRF.probability(1 - class)
4:     ProbRF-2 = MRF.probability(2 - class)
5:     ProbCNN-1 = MKNN.probability(1 - class)
6:     ProbCNN-2 = MKNN.probability(2 - class)
7:     ProbCNN-1 = MLR.probability(1 - class)
8:     ProbCNN-2 = MLR.probability(2 - class)
9:     Decision function =  $\max_{i=1,2}$ 
Pclassifier
           Nclassifier
(Avg(ProbRF-1, ProbKNN-1, ProbLR-1)
, (Avg(ProbRF-2, ProbKNN-2, ProbLR-2)
10:   end if
11:   Return final label  $p^{\wedge}$ 
12: end for
```

Lines 3 to 6 of algorithm 1 indicate the probability scores of classes 1 and 2 from RF, KNN, and LR models, respectively. The probability score of a classifier, often used in classification tasks, represents the likelihood or confidence that a given data sample belongs to a particular class. It is calculated based on the output of the classifier model, such as in this case RF, KNN, and LR. This probability is not the final prediction of a specific target class.

It is like a raw score (prediction confidence). To convert raw scores into probabilities, some classifiers employ a probability calibration step. This step ensures that the calculated scores are well-calibrated and can be interpreted as probabilities. Based on the probability scores and the chosen threshold (0.5 in this case), the classifier assigns a final predicted class label to the data sample. The decision function of line 7 of algorithm 1, decides the final class based on the class which has more probability score than the assigned threshold. The working example of the decision function is added below for further clarification.

The working of this ensemble can be explained using an example. Each sample that undergoes processing by both the SVM and CNN is assigned a probability score. Consider a scenario where the RF model assigns a probability of 0.4 and 0.7 for class 1 and class 2, KNN model assigns a probability of 0.5 and 0.8 for class 1 and class 2, respectively, and the LR model assigns probability scores of 0.5 and 0.4 for the same two classes. Denoting the probability value of x as $P(x)$, where x ranges from 1 to 2, the final probability is calculated as follows:

$$P(1) = (0.4 + 0.5 + 0.5)/3 = 0.46$$

$$P(2) = (0.7 + 0.8 + 0.4)/3 = 0.63$$

This ensemble approach makes the final class label based on the probability scores for each class from both models used for voting. The final label is decided based on the highest average probability using line 7 of algorithm 1.

The proposed model uses the distinct capabilities of three separate ML models to provide results that are both accurate and resilient. To generalize the model while decreasing the over-fitting by training models on the cervical cancer dataset and fusing their results of predictions.

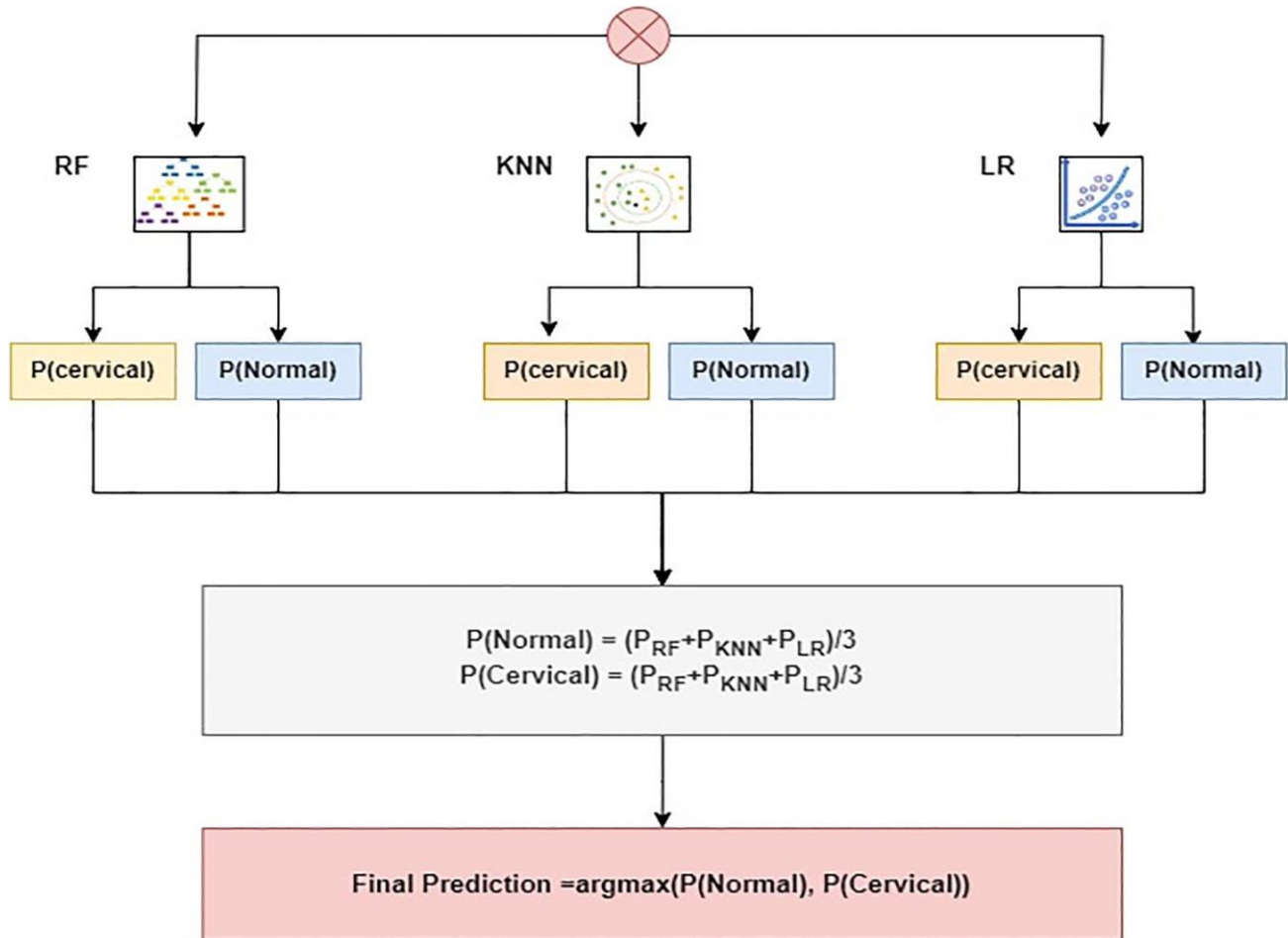


Fig 3 Structure of the proposed ensemble voting classifier

IV. RESULTS

This section discusses the results of experiments and their consequences, with an emphasis on determining the efficacy of the suggested methodology in comparison to existing methodologies. The assessment includes a variety of test parameters used for the cervical cancer dataset, and the results are contrasted with alternative ML approaches. The tests are conducted using the original dataset, the SVM Imputed dataset, the dataset upsampled using ADASYN and imputed with SVM, and the dataset containing CNN features.

A 5-fold cross-validation was undertaken to further confirm the efficacy of the suggested technique. The findings are shown in Table 6. Notably, the suggested model has an average accuracy of 99.27, as well as average precision, recall, and F1 score values of 99.96%, 99.96%, and 99.97%.

The comparison of classifiers with the original dataset, with SVM Imputed dataset, With ADASYN upsampled and SVM Imputed dataset, and with CNN generated features dataset is depicted in Fig 4. When applied to ADASYN-balanced data, the Voting classifier demonstrates its superiority over all other classifiers. RF, KNN and LR perform better than other individual classifiers in every scenario. These findings underscore the significance of selecting the right combination of ML approaches for the effective operation of an ensemble ML model. In the context of analyzing imbalanced text data, employing data balancing techniques like ADASYN significantly enhances classifier performance.

This highlights the importance of utilizing the statistical technique ADASYN to balance the data before training, as it plays a pivotal role in enhancing classifier performance. It becomes apparent that classifiers may not achieve their optimal performance when dealing with imbalanced classes within the original dataset. Consequently, when coupled with the SVM imputer and the ADASYN technique for cervical cancer detection, the proposed model demonstrates improved generalization and outperforms other models when appropriately configured.



ADASYN and SVM imputer are very useful techniques in improving the performance of the models with class imbalance problems. When experiments are performed using CNN features, the results of the classifier have shown significant improvement in results. The proposed voting classifier has outperformed with a 99.99% score of accuracy, precision, recall and F1 score. Ensembling RF, KNN, and LR models offer a strategic advantage by combining their

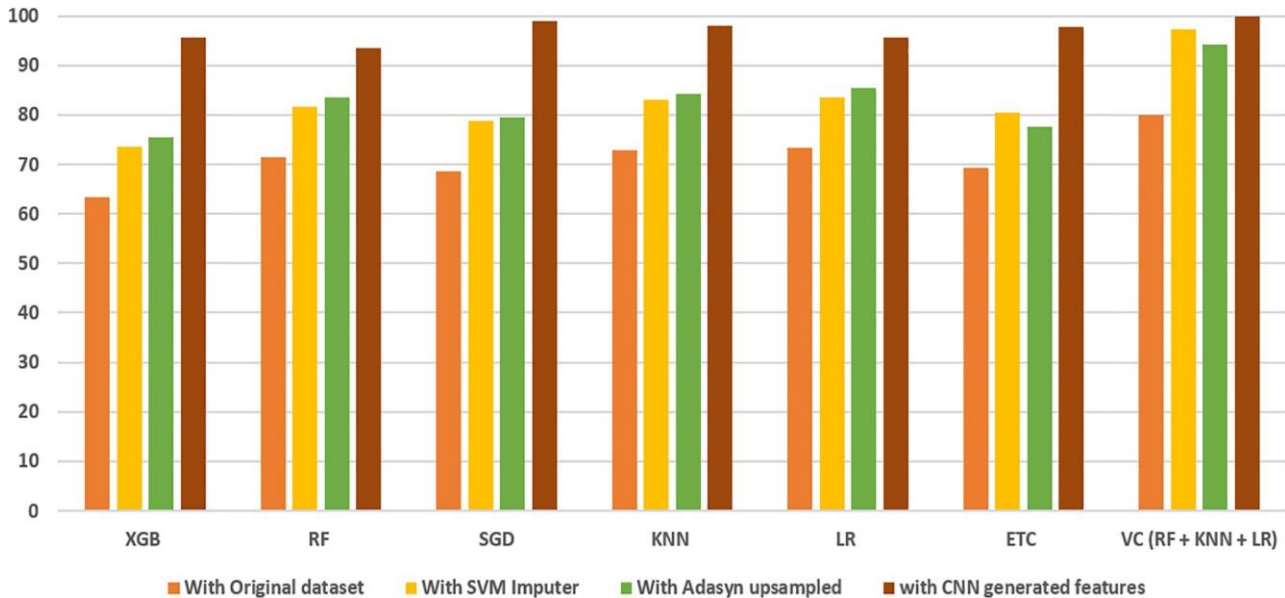


Fig 4. Comparison of model accuracies in all scenarios.

diverse learning approaches to improve predictive performance. This ensemble aims to leverage the strengths of each model: RF's robustness, KNN's pattern recognition, and LR's probabilistic interpretation, thereby enhancing overall accuracy, reducing overfitting, improving robustness against outliers, and providing a more comprehensive analysis of the cervical cancer dataset. The ensemble, facilitated by a voting classifier, fosters a collective decision-making process, resulting in a more robust and accurate predictive model for cervical cancer detection.

Comparative analysis with cutting-edge methods

To assess the efficacy of the suggested method, a performance comparison with existing models specialised in cervical cancer diagnosis is performed. This review includes a selection of current research from the existing studies that serve as comparative points. According to one research [48], a cancer detection model using Recursive Feature Elimination (RFE) and DT using SMOTETomek obtains an accuracy of 98.82%, precision of 87.53%, recall of 100%, and an F1 score of 93.333%. Another research [51] uses 10 features for the same job and obtains an accuracy score of 97.72%.

Furthermore, studies [52] applied MLP and [53] used SVM and revealed accuracy rates of 98.10% and 99%, respectively. Despite the great accuracy reported in previous research works, the suggested models outperform them, as shown in Fig 5. The suggested method outperforms previous approaches due to three main factors: managing missing data, employing an ensemble voting classifier, and adding CNN-generated features.

The novel mix of strategies, which includes correcting missing data, applying ensemble learning, and regulating class imbalance, is critical to the observed accuracy gains. While some earlier techniques may have failed to resolve the problem of missing data directly, this work uses an SVM imputation strategy in conjunction with ADASYN up-sampled features. In addition, the suggested technique makes use of a stacked ensemble voting classifier, which combines the results of three independent models. This collaborative approach is advantageous.

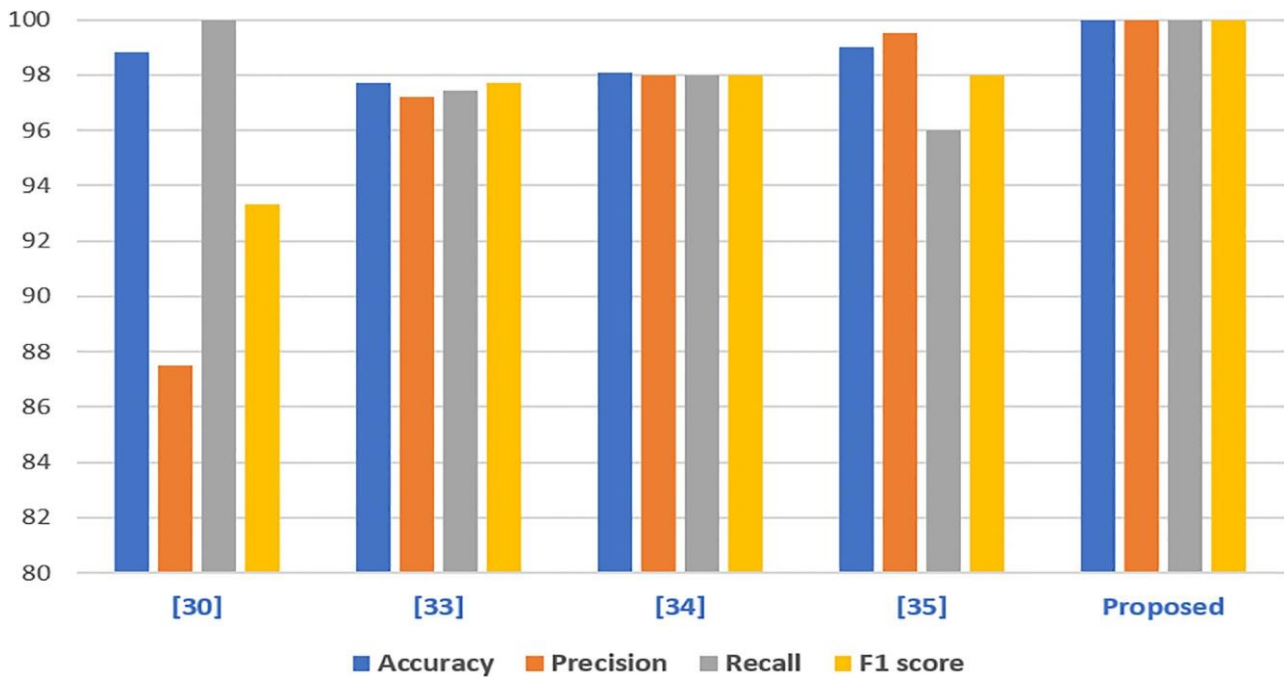


Fig 5. Comparison of model accuracies in all scenarios.

V. CONCLUSION

Cervical cancer poses a significant threat to women’s health, particularly in developing countries, where it ranks as a leading cause of mortality. Timely detection and treatment, guided by skilled medical professionals, are paramount in mitigating its devastating impact. Pap smear images have emerged as valuable diagnostic tools for identifying this form of cancer. However, numerous datasets designed for automated cervical cancer detection present a common challenge: missing values. These gaps in data can substantially hinder the performance of machine learning models, necessitating innovative solutions. In response to these challenges, this study introduces an automated system tailored for cervical cancer prediction. This system demonstrates remarkable proficiency in managing missing values through the utilization of ADASYN features, ultimately achieving exceptional levels of accuracy. The cornerstone of the proposed approach is a stacked ensemble voting classifier model, strategically combining the predictive capabilities of three distinct machine learning models. Furthermore, SVM Imputer and ADASYN up-sampled features are integrated into the proposed framework to effectively address concerns related to missing values. The inclusion of CNN-generated features further bolsters the model’s robustness.

Notably, the outcomes of this study reveal the exceptional performance of the proposed model, boasting remarkable metrics such as 99.99% accuracy, 99.99% precision, 99.99% recall, and a 99.99% F1 score. To comprehensively assess the proposed model, a comparative analysis is conducted against various machine learning algorithms under four distinct scenarios: using the original dataset, employing SVM imputation, incorporating ADASYN features, and harnessing CNN-generated features. These comparative evaluations underscore the superior efficacy of the proposed model when compared to existing state-of-the-art approaches. The research has the potential to significantly benefit medical practitioners by enabling earlier cervical cancer detection and improving patient care. Future work aims to develop stacked ensembles of machine and deep learning models for enhanced performance on higher-dimensional datasets.

REFERENCES

- [1]. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*. 2018; 68(6):394–424. PMID: 30207593
- [2]. Arbyn M, Weiderpass E, Bruni L, de Sanjose S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *The Lancet Global Health*. 2020; 8(2):e191– e203. [https://doi.org/10.1016/S2214-109X\(19\)30482-6](https://doi.org/10.1016/S2214-109X(19)30482-6) PMID: 31812369
- [3]. Pal A, Kundu R. Human papillomavirus E6 and E7: the cervical cancer hallmarks and targets for therapy. *Frontiers in microbiology*. 2020; 10:3116. <https://doi.org/10.3389/fmicb.2019.03116> PMID: 32038557

- [4]. Dong N, Zhao L, Wu CH, Chang JF. Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing*. 2020; 93:106311. <https://doi.org/10.1016/j.asoc.2020.106311>
- [5]. Zhang T, Luo Ym, Li P, Liu Pz, Du Yz, Sun P, et al. Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images. *Biomedical signal processing and control*. 2020; 55:101566. <https://doi.org/10.1016/j.bspc.2019.101566>
- [6]. Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: Epidemiology, risk factors and screening. *Chinese Journal of Cancer Research*. 2020; 32(6):720. <https://doi.org/10.21147/j.issn.1000-9604.2020.06.05> PMID: 33446995
- [7]. Bedell SL, Goldstein LS, Goldstein AR, Goldstein AT. Cervical cancer screening: past, present, and future. *Sexual medicine reviews*. 2020; 8(1):28–37. <https://doi.org/10.1016/j.sxmr.2019.09.005> PMID: 31791846
- [8]. Jalil AT, Karevskiy A. The cervical cancer (CC) epidemiology and human papillomavirus (HPV) in the middle east. *International Journal of Environment, Engineering and Education*. 2020; 2(2):7–12. <https://doi.org/10.55151/ijeedu.v2i2.29>
- [9]. Kashyap N, Krishnan N, Kaur S, Ghai S. Risk factors of cervical cancer: a case-control study. *Asia-Pacific journal of oncology nursing*. 2019; 6(3):308–314. https://doi.org/10.4103/apjon.apjon_73_18 PMID: 31259228
- [10]. Davies-Oliveira J, Smith M, Grover S, Canfell K, Crosbie E. Eliminating cervical cancer: progress and challenges for high-income countries. *Clinical Oncology*. 2021; 33(9):550–559. <https://doi.org/10.1016/j.clon.2021.06.013> PMID: 34315640
- [11]. Liang LA, Einzmann T, Franzen A, Schwarzer K, Schauburger G, Schriefer D, et al. Cervical cancer screening: comparison of conventional Pap smear test, liquid-based cytology, and human papillomavirus testing as stand-alone or cotesting strategies. *Cancer Epidemiology, Biomarkers & Prevention*. 2021; 30(3):474–484. <https://doi.org/10.1158/1055-9965.EPI-20-1003> PMID: 33187968
- [12]. O’Laughlin DJ, Strelow B, Fellows N, Kelsey E, Peters S, Stevens J, et al. Addressing anxiety and fear during the female pelvic examination. *Journal of Primary Care & Community Health*. 2021; 12:2150132721992195. <https://doi.org/10.1177/2150132721992195> PMID: 33525968
- [13]. Maver P, Poljak M. Primary HPV-based cervical cancer screening in Europe: implementation status, challenges, and future plans. *Clinical microbiology and infection*. 2020; 26(5):579–583. <https://doi.org/10.1016/j.cmi.2019.09.006> PMID: 31539637
- [14]. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*. 2020; 11(1):3923. <https://doi.org/10.1038/s41467-020-17419-7> PMID: 32782264
- [15]. Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Applied Sciences*. 2019; 1:1–16. <https://doi.org/10.1007/s42452-019-0645-7>
- [16]. Akter L, Islam MM, Al-Rakhami MS, Haque MR. Prediction of cervical cancer from behavior risk using machine learning techniques. *SN Computer Science*. 2021; 2:1–10. <https://doi.org/10.1007/s42979-021-00551-6>
- [17]. Islam MM, Haque MR, Iqbal H, Hasan MM, Hasan M, Kabir MN. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*. 2020; 1:1–14. <https://doi.org/10.1007/s42979-020-00305-w>
- [18]. Srinivasulu A, Ramanjaneyulu K, Neelaveni R, Karanam SR, Majji S, Jothilingam M, et al. Advanced lung cancer prediction based on blockchain material using extended CNN. *Appl Nanosci*. 2021; 13:1–13.
- [19]. Puente-Castro A, Fernandez-Blanco E, Pazos A, Munteanu CR. Automatic assessment of Alzheimer’s disease diagnosis based on deep learning techniques. *Computers in biology and medicine*. 2020; 120:103764. <https://doi.org/10.1016/j.compbiomed.2020.103764> PMID: 32421658
- [20]. Aggarwal R, Sounderajah V, Martin G, Ting DS, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*. 2021; 4(1):65. <https://doi.org/10.1038/s41746-021-00438-z> PMID: 33828217
- [21]. Wang S, Yang DM, Rong R, Zhan X, Xiao G. Pathology image analysis using segmentation deep learning algorithms. *The American journal of pathology*. 2019; 189(9):1686–1698. <https://doi.org/10.1016/j.ajpath.2019.05.007> PMID: 31199919
- [22]. Chang QQ, Chen CY, Chen Z, Chang S. LncRNA PVT1 promotes proliferation and invasion through enhancing Smad3 expression by sponging miR-140-5p in cervical cancer. *Radiology and Oncology*. 2019; 53(4):443. <https://doi.org/10.2478/raon-2019-0048> PMID: 31626590
- [23]. Xie X, Wang X, Liang Y, Yang J, Wu Y, Li L, et al. Evaluating cancer-related biomarkers based on pathological images: a systematic review. *Frontiers in Oncology*. 2021; 11:763527. <https://doi.org/10.3389/fonc.2021.763527> PMID: 34900711
- [24]. Chen S, Chen Y, Yu L, Hu X. Overexpression of SOCS4 inhibits proliferation and migration of cervical cancer cells by regulating JAK1/STAT3 signaling pathway. *European Journal of Gynaecological Oncology*. 2021; 42(3):554–560. <https://doi.org/10.31083/j.ejgo.2021.03.2416>
- [25]. Quinlan S, Afli H, O’Reilly R. A Comparative Analysis of Classification Techniques for Cervical Cancer

Utilising At Risk Factors and Screening Test Results. In: AICS; 2019. p. 400–411.

- [26]. Abdoh SF, Rizka MA, Maghraby FA. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*. 2018; 6:59475–59485. <https://doi.org/10.1109/ACCESS.2018.2874063>
- [27]. Ijaz MF, Attique M, Son Y. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors*. 2020; 20(10):2809. <https://doi.org/10.3390/s20102809> PMID: 32429090
- [28]. Jahan S, Islam MS, Islam L, Rashme TY, Prova AA, Paul BK, et al. Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. *SN Applied Sciences*. 2021; 3:1–17.
- [29]. Fernandes K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. In: *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*. Springer; 2017. p. 243–250.
- [30]. Mallinson H, Gammerman A. *Imputation using support vector machines*. University of London Egham, UK: Department of Computer Science Royal Holloway. 2003;.
- [31]. Rendon E, Alejo R, Castorena C, Isidro-Ortega FJ, Granda-Gutierrez EE. Data sampling methods to deal with the big data multi-class imbalance problem. *Applied Sciences*. 2020; 10(4):1276. <https://doi.org/10.3390/app10041276>
- [32]. Brandt J, Lanzen E. A comparative review of SMOTE and ADASYN in imbalanced data classification. *DIVA*. 2021;.
- [33]. Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. *Information Sciences*. 2021; 572:522–542. <https://doi.org/10.1016/j.ins.2021.05.055>
- [34]. Schonlau M, Zou RY. The random forest algorithm for statistical learning. *The Stata Journal*. 2020; 20 (1):3–29. <https://doi.org/10.1177/1536867X20909688>
- [35]. Liu Y, Gao Y, Yin W. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*. 2020; 33:18261–18271.
- [36]. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *Journal of thoracic disease*. 2019; 11(Suppl 4):S574. <https://doi.org/10.21037/jtd.2019.01.25> PMID: 31032076
- [37]. Sharaff A, Gupta H. Extra-tree classifier with metaheuristics approach for email classification. In: *Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018*. Springer; 2019. p. 189–197.