International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Cervical Cancer Prediction using NGBFA Feature Selection Algorithm and Hybrid Ensemble Classifier

[1] CH. Bhavani, [2] Dr. A. Govardhan
[1] Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India.
ch.bhavani@cvr.ac.in
[2] Professor, Rector & professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India
govardhan_cse@jntuh.ac.in

Abstract: Cervical Cancer (CC) is a substantial reason of death midst middle-aged women throughout the world, specifically in developing countries, with approximately 85% of deaths. CC patients can be healed if spotted in the early stages. As no symptoms appear in the initial stages, it has become a challenge for investigators to predict the disease in the early stages. Several machine learning algorithms have been used to predict CC since the last decade. Instead of using a single classifier for the prediction, ensemble methods give accurate results, creating and combining multiple models to produce improved results. In this study, we built a hybrid ensemble classifier, 'A Robust Model Stacking: A Hybrid Ensemble,' in which a homogenous ensemble will be performed on a pool of classifiers in the base level followed by a heterogenous ensemble using the majority voting (soft) algorithm to get the final prediction of the new data. The dataset used in this study contains 858 instances with 32 features built from the risk factors and four targets made from the CC diagnosis tests. We have solved the data imbalance problem using an oversampling technique called SMOTE. The model's efficiency was evaluated based on the accuracy, recall, f1-score, precision, and AUC-ROC curve metrics for all four target variables in the dataset. The proposed Biopsy method's accuracy is 98%, Hinselmann is 97%, Schiller is 96.09%, and Citology is 93%. We implement ensemble learning in this study to increase prediction accuracy and decrease bias and variance. We carried the experiments out using the Python language in Google Colab and Jupyter notebooks. The experimental results revealed that our proposed hybrid ensemble learning records a remarkable accuracy for all four target variables.

Keywords- Cervical cancer, homogenous ensemble, heterogenous ensemble, Feature Selection, prediction

I. INTRODUCTION

Cervical Cancer is the fourth most dangerous cancer and second most among women. This type of cancer is challenging because it has no symptoms in the initial stages. It can be curative if spotted in the early stages [1]. CC is a variant of cancer that arises at the cervix in women's uteri, where the malignant cells grow and divide abnormally in the body, invading other body organs [2, 3]. The primary cause of CC is infection with Human Papillomavirus (HPV). It is transmitted through sexual intercourse. Usually, every woman will be infected with HPV in their intercourse, and it goes away normally, but in a few cases, it leads to abnormal behavior of cells leading to pre-cancer and slowly converting into CC [4-6]. It is suggested that women aged between 30 and 49 years must undergo the screening test at least once a year to avoid the risk of CC. The usual symptoms of CC are abnormal or irregular menstruation, heavy vaginal discharge, inexplicable pain in the pelvis and spotting, back pain, tiredness or dullness, leg-torment, weight loss, and loss of hunger [7, 8]. Machine learning (ML), a branch of AI, uses a wide variety of probabilistic, statistical, and optimization methods that use historical data from the past to classify the current data. Especially in medical data, this ability is suited well. These ML algorithms are advantageous in improving accuracy, making decisions in emergency cases, and helpful, especially when medical assistants are in shortage. It is proven that ML applications are very successful in diagnosing any cancer [9, 10]. Predicting whether the patient is cancerous or non-cancerous based on a single ML model does not guarantee better accuracy, as no superior classifier can generate the best results at all times because it relies on the context used [11]. Indeed, every ML model has its own advantages and disadvantages in terms of classification. Hence, many researchers started working on improving classification efficiency by using a collection of classifiers instead of a single model [12]. The ensemble combines the advantages of multiple individual classifiers and lessens the weaknesses of individual classifiers. Ensemble classification techniques combine various classifiers based on a definite combination rule [12]. Over the past decade, ensemble techniques and methodologies have grabbed the attention of many researchers in the field of CC, so we get better diagnoses, prognoses, and treatment in a minimum amount of time. The main motive behind the design of the ensemble technique is to combine multiple single classifiers on specific association rules to produce a global model that provides a reliable solution or estimators or predictions for a given problem [13]. Experimental and theoretical verification provide better prediction performance when compared with individual models. Different ensemble learning methods have been proposed and used in classification and regression for solving real-world problems [14].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Usually, the ensemble techniques are broadly classified into two types. 1) Homogeneous ensemble techniques and 2) Heterogenous ensemble techniques. The homogenous ensemble technique is a process of combining multiple variants of the same type of classifier. The variants may be dividing the feature set, sampling the dataset, or training the models with various parameters or different parameter values. The best homogenous ensemble techniques are bagging and boosting. Bagging - a bootstrap aggregator that contains decision trees as the base learners. Each decision tree is constructed with a 'random sampling with replacement' technique on features and samples from the dataset. All the base learners will be trained simultaneously, and a majority voting algorithm will be applied to get the final prediction of the new unseen data [15, 16]. A heterogeneous ensemble learning is a process of grouping different classification models as the base learners and combining the output of these base models to get more accurate prediction results that could not be possible with individual classifiers. Stacking is an example of a heterogeneous ensemble technique, which combines the variety of base learners and forms a new dataset with the predictions of base learners at the meta-level, and a meta-level classifier is used to get the concluding prediction of the new data [17]. Building of heterogenous models can be broadly done in two ways. In the initial method, a static number of models are merged and whereas in the second method, models with various parameterizations are combined to get the final prediction.

II. RELATED WORK

The use of AI has gradually increased in the medical field for the diagnosis of diseases. Nevertheless, there is no such single classification algorithm, which will be performing accurately in all the scenarios.

In a study [18], the authors proposed a static heterogeneous ensemble that combines SVM, LR, DT, MLP, and KNN. The classifiers are resolute using 10-fold cross-validation. This tactic has shown better outcomes in the classification of lithofacies. Ref [19] proposes combining various carefully selected strong learners, for example, deep neural networks, and SVM, AdaBoost, and Gaussian processes, to form a strong learning model. In this study, the researchers used a fusion technique to generate the ensemble using the sum rule on diverse classifiers. In [20], from the 20 models, to select the optimum number of models, the authors proposed a genetic algorithm for ensemble pruning on homogenous ensembles. In a recent study [21], to create a heterogeneous combination that can be considered effective, the authors trim off the poor performers from base learners so that only optimal classifiers are kept in the ensemble. The efficacy of a classifier is recognized using the ROC-AUC measure. In [22], the researchers constructed a homogenous module, where various learning models will be trained on training data, to generate new training sets, they used the sum rule and majority voting for combining four different models.

In [23], proposed a CRISP-DM model which makes use of stacking with 3 algorithms KNN, SVM, and decision tree, and the efficacy of the model has been compared with individual classifiers, and identified the better accuracy. In [24], authors proposed a method that will assist doctors in predicting the survival rate of neuroblastoma-a pediatric cancer, with five heterogeneous base learners based on a genetic algorithm along with heterogenous feature selection for each classifier, interpretable rules used and was able to get extract more than 90% accuracy. In [25], it was shown that deep learning can enable accurate diagnosis of CC using risk factors, and 68% of AUC was achieved with supervised autoencoding. In [26], the authors proposed a hybrid model CervDetect model which combines a random forest classifier and shallow neural network for detecting CC and was able to achieve an accuracy of 93.6%. In [27], the author has done a relative analysis based on the mean value replacement along with the ensemble learning technique for predicting the risk of CC with incomplete data. In [28], the researchers used various meta-classification algorithms with a set of feature selection techniques and five classifiers had been selected as Meta classifier for evaluating the dataset. Based on the Attribute Selected they were able to get the lowest error rate with their proposal. In [29], the authors used RFFS and ETFS feature selection techniques along with a stacked ensemble algorithm to identify important risk factors that are responsible for the cause of CC. In [30], the authors employed a random forest algorithm with RFE for feature selection, and SMOTE for data balancing after comparing with various other data balancing techniques, for classification stacking algorithm was used to get a better classification performance with two-stage classifiers.

Most of the previous research work under CC prediction has not been focused on data balancing even though the CC data is imbalanced, only a certain number of studies discussed the selection of important features. Our study addressed the problem of data imputation which deals with missing data, data balancing, and feature selection and the proposed model shows better efficacy than the classification models used in the previous studies.

III. THE PROPOSED WORK

We have illustrated the architecture of the proposed model in Figure 1. In Phase 1, as part of data preprocessing, we eliminated the noise and performed data normalization and standardization. In phase 2, again, as part of preprocessing, we selected the required set of features using the Novel Genetic Inspired Binary Firefly Algorithm. In phase 3, we have projected the proposed hybrid model stacking ensemble classifier to predict whether the patient is cancerous or non-cancerous. In phase 4, we classify instances as malignant or benign.

Description of the Dataset

The CC dataset was taken from 'The University sitario de Caracas' hospital in Caracas, which consists of 858 records with 33 features shown in Figure 2. It consists of CC risk factors based on the patient's medical reports. It comprises four target or independent attributes (Biopsy, Hinselmann, Schiller, Cytology) and four diagnosis tests for CC. The basic objective of our proposed model is to predict the CC with better accuracy. As the dataset has been constructed through the survey of patients, only a few patients gave some data, as they were not

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

interested in revealing their personal data, which will be missing data in the dataset. The dataset consists of 803 non-cancerous patients' data and 55 cancerous patients' data, indicating the dataset's imbalances. The dataset has many features that could be more useful in predicting CC, suggesting the need for feature selection.

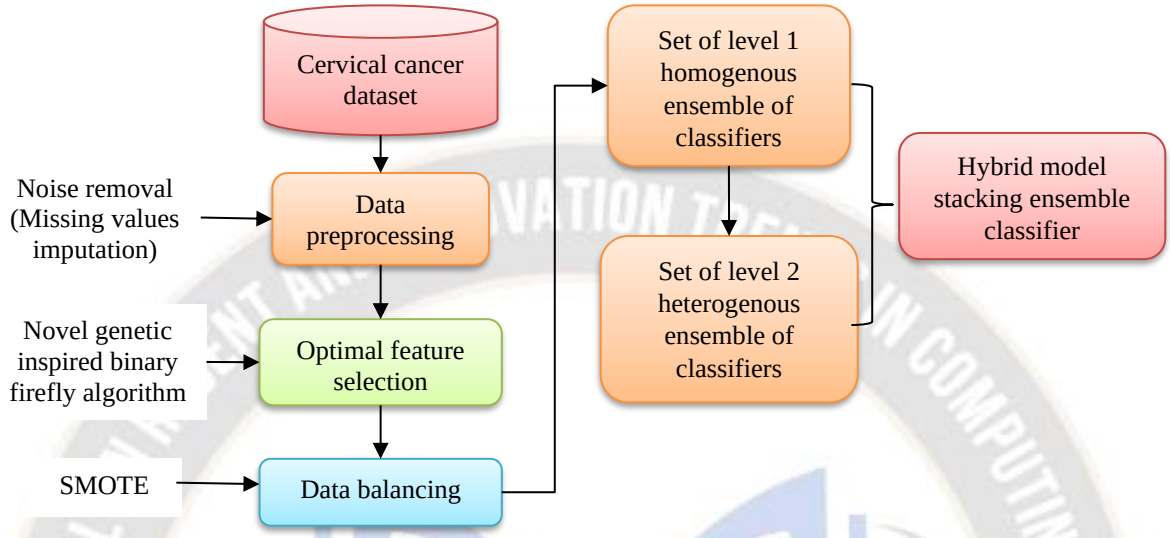

Figure 1: The overall structure flow of the methodology

Attribute	Type	Attribute	Type	Attribute	Type
Age	Integer	STDs	Bool	STDs:HIV	Bool
Number of sexual partners	Integer	STDs (number)	Integer	STDs:Hepatitis B	Bool
First sexual intercourse (age)	Integer	STDs:condylomatosis	Bool	STDs:HPV	Bool
Number of pregnancies	Integer	STDs:cervical condylomatosis	Bool	STDs: Number of diagnosis	Integer
Smokes	Bool	STDs:vaginal condylomatosis	Bool	STDs: Time since first diagnosis	Integer
Smokes (years)	Bool	STDs:vulvo-perineal condylomatosis	Bool	STDs: Time since last diagnosis	Integer
Smokes (packs/year)	Bool	STDs:syphilis	Bool	Dx:Cancer	Bool
Hormonal Contraceptives	Bool	STDs:pelvic inflammatory disease	Bool	Dx:CIN	Bool
Hormonal Contraceptives (years)	Integer	STDs:genital herpes	Bool	Dx:HPV	Bool
IUD	Bool	STDs:molluscum contagiosum	Bool	Dx	Bool
IUD (years)	Integer	STDs:AIDS	Bool		

Figure 2: Features and their data types

Medical history, habits, and demographics are in the dataset. Figure 3 shows the original dataset target variables result, yes for '1' and 'no' for '0' distribution. Due to privacy concerns, several patients did not answer highly private questions, resulting in too many missing values in the dataset, removing numerous risk factors and records. Data analysis led to many alternatives.

Data pre-processing

This transforms the dataset so the model can understand it. CC risk factors contain several missing values. Thus, we need an effective solution. Fill in or delete missing values. Remove the missing value strategy in large datasets if it is unimportant. We decreased rows from 858 to 737 by removing entries with missing data. In Equations (1) and (2), we replace missing values with the mean for numerical attributes and mode for categorical attributes to minimize the number of features, not the number of records in the dataset.

$$Mean\ \overline{a} = \frac{1}{m}\left(\sum_{l=1}^{m} a_l\right) \quad (1)$$

where, a_l denotes the lth variable, m denotes variables present in the dataset

$$Mode\ C = K + \left(\frac{v_n + v_{n-1}}{(v_n - v_{n-1}) + (v_n - v_{n+1})}\right) Z \quad (2)$$

where, Z denotes model class length, v_{m-1} denotes previous class, v_{m+1} denotes future class frequencies respectively

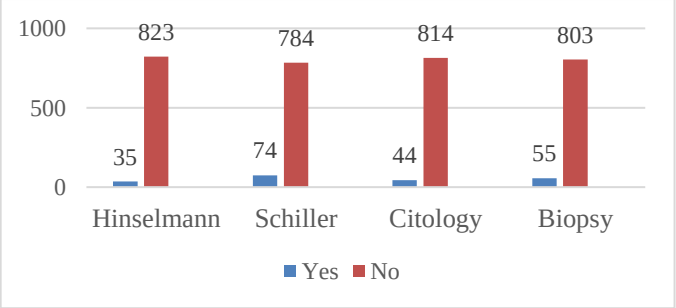

Figure 3: Target variables outcome distribution in the original dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

The dataset also removed the characteristics STDs_Time_since_first_diagnosis and STDs_Time_since_last_diagnosis that had above 60% missing values (787 of 858). However, the simple Imputer with the mean (Equation 3) has been used to insert the missing values in numerical features.

$$A = \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right) = (a_1 + a_2 + a_3 + \cdots + a_n)n \qquad (3)$$

Simple Imputer with mode has been used to insert the missing values among the categorical features. Now, these two parts were concatenated to form a single dataset, through which the missing values problem was resolved.

Data balancing

Since nearly 96% of the observations in the cervical cancer risk factors dataset are non-cancerous and just 4% are malignant cases, the distribution of positive and negative classes is severely skewed. Prior investigations in the cervical cancer risk factors dataset have given little attention to the issue of unbalanced datasets. ISMOTE is used to correct the large data imbalance in the CC dataset. ISMOTE is the resampling approach that integrates oversampling (SMOTE) with undersampling.

Feature Selection Algorithm: NBGFA

For feature reduction, the redundant features were identified using correlation analysis, and redundant features were deleted. As part of selecting the crucial features for the prediction of CC, recursive feature elimination with a random forest classifier using a 10-fold stratified cross-validation technique has been implemented. We identified eight optimal features for the biopsy target variable, 7 for cytology, 13 for Hinselmann, and 16 for Schiller. We standardized the new dataset with optimal features with the standard scalar method to avoid the model's biases towards the higher range of values features. As the dataset was not balanced, to avoid the biases of the classifier towards the majority class while predicting the class label, SMOTE–a synthetic minority over-sampling technique, which can overcome the over-fitting problem was employed by dividing the dataset with 80% of training data and 20% for testing. Because of this, we have increased the size of the dataset to 1282 records, which contain an equal number of records for both cancerous and non-cancerous patients. The NGBFA is a hybrid nature-inspired swarm intelligence feature selection algorithm that combines the genetic operation with a firefly binary version algorithm to extract the optimal number of features employing a random forest algorithm that presents an inference in the search space based on communal behaviors of the swarm [31].

Proposed Method – A Robust Model Stacking: A Hybrid Approach

There is no consent on any classification algorithm which can generate the best performance in all cases. Because of this, recently, ensemble learning has gained more attention as it will yield better accuracy than specific classifiers. In this study, a hybrid ensemble method, a robust model stacking, has been proposed, which contains the following steps.

Homogenous Ensembling

Step 1: Initially, the dataset *D* with *N* instances is separated into a training and testing data set.

Step 2: On the training data set, apply classifier algorithm 1 from the '*n*' classifiers set. Repeat this for 'm' times by varying classifier parameters to obtain an ensemble of size '*m*'.

Step 3: Apply a soft voting algorithm to obtain first-level predictions from the ensemble from Step 2.

Step 4: Repeat steps 2 and 3 by applying the remaining classifier algorithm 2 to *n*.

Heterogenous Ensembling

Step 5: Finally, apply the soft voting algorithm on the predictions obtained by step 3 for all '*n*' classifiers.

Step 6: Test the classifier accuracy from the predictions obtained from Step 5. We show the architecture of the projected model in Figure 4.

Pseudo Code

The pseudo code of the proposed model is presented as following:

Input: *D*: Training dataset

N: Number of instances

n: Number of classifiers

O**utput**: Final Predictions along with the classifier accuracy

Begin

Step 1: Homogenous Ensemble

for *i* := 1 to *n* do // for 'n' number of classifiers

for j := 1 to *m* do // for '*m*' different parameter sets

apply $\text{Classifier}_{ij}(D)$

Find ensemble for each classifier *i*

Step 2: Heterogenous Ensemble

Combine the output predictions of step 1 ensemble classifiers of i with soft voting algorithm to obtain final predictions.

Step 3: Final accuracy of the model

End

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

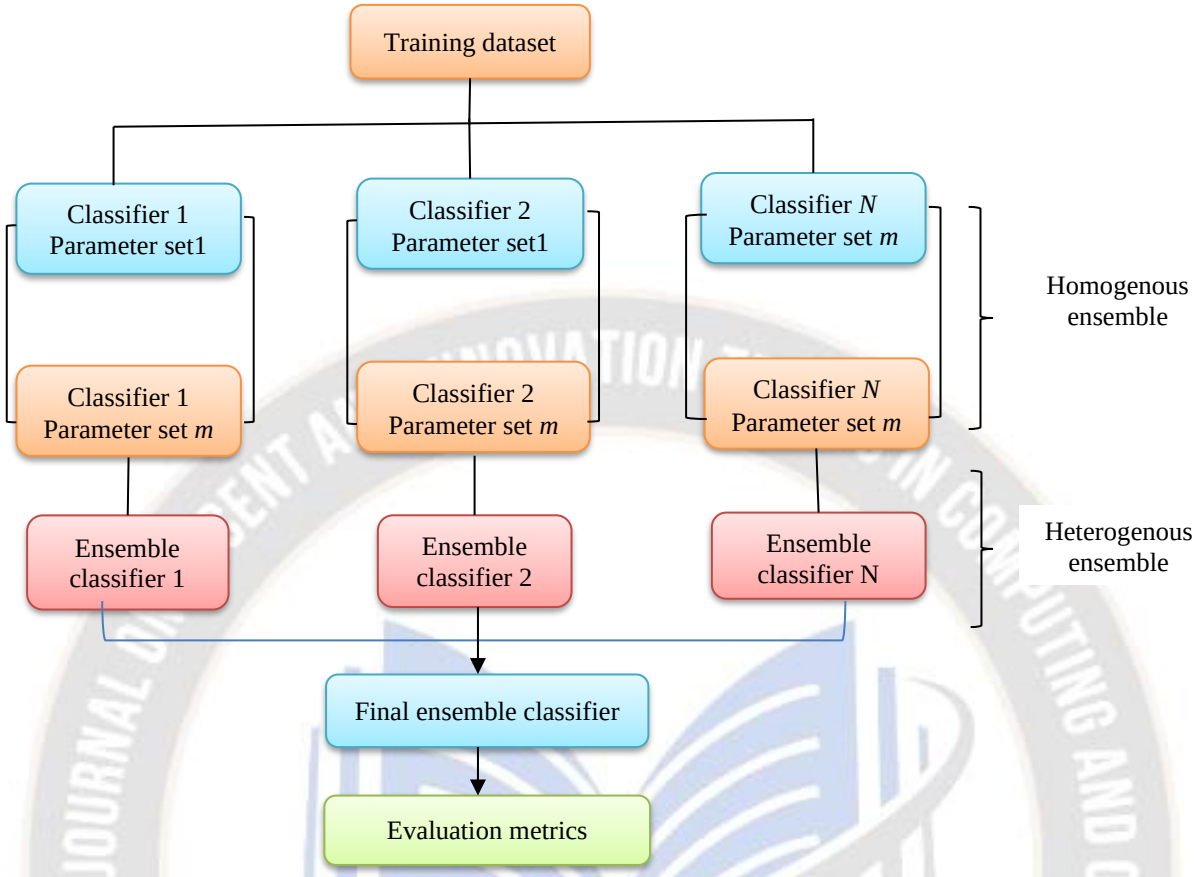

Figure 4: A Robust Model Stacking: A Hybrid Approach

IV. RESULTS AND DISCUSSIONS

Python was used to create the ML models. We execute the Python code on Google Colab, a collaborative platform. We run the application on an Intel Core i5 central processor unit (CPU) with 2.65 GHz. The machine contains 8 GB RAM. This section presents the results obtained for eight algorithms k nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), naïve Bayes (NB), random forest (RF), bagging classifier (BC), decision tree (DT) and multi-layer perception (MLP) classifiers were used in the base level. We implemented each of the algorithms five times with five different parameter sets. A homogenous Ensembling with a soft voting technique has been performed to get the final accuracy of the ensemble at level 1. Now, at level 2, a heterogenous ensemble was performed with the predictions generated for the test data by all the classifier ensembles at level 1. Using the soft voting algorithm, we generated the final accuracy. We trained 40 algorithms based on the voting rule. We found our robust model stacking: a hybrid ensemble performed better than the state-of-art classifiers. We have assessed the performance of the projected model using accuracy, precision, recall, f1-score, and ROC-AUC curve metrics (equations 4-8). The results of all four-target variables are tabled in Tables 1,2,3,4:

$$Accuarcy = \frac{\text{TP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

$$\frac{\text{Sensitivity}}{\text{Recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\frac{\text{Specificity}}{\text{Precision}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{F1} - \text{score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (7)$$

$$ROC - AUC = 1 - \frac{1}{p^{+}p^{-}} \sum_{a^{+}\epsilon P^{+}} \sum_{a^{-}\epsilon P^{-}} \left(\left(f(a^{+}) < f(a^{-}) \right) + \frac{1}{2} \left(f(a^{+}) = f(a^{-}) \right) \right) \quad (8)$$

The assessment of a model using solely ML metrics cannot adequately represent the model's scientific and impartiality. As a result, in this research, we employ three statistical parameters to measure the model's statistical efficiency. We describe each statistical indicator below:

Cohen's kappa (CK): This is a frequently used agreement metric, reflects the degree of agreement among the actual and projected outcomes on the classification issue and is computed following a formula [32]:

$$k = \frac{(q_i - q_j)}{(1 - q_j)} \quad (9)$$

where q_i denotes original ratio and q_j denotes theoretical ratio.

Matthews' correlation coefficient (MCC): In ML, MCC is an indication of the quality of the binary-class model, which is

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

effectively a correlation coefficient value with -1 and + 1. [33]: we compute MCC in this manner:

$$MCC = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (10)$$

Result of target feature: Biopsy

Because of NGBFA-RF, age, number_of_sexual_partners, first_sexual_intercourse, num_of_pregnancies, and IUD (years), we selected five optimal features as essential features for the biopsy target feature. We tabulated the experimental results in Table 1 for the biopsy target variable with an accuracy of 98% with the proposed model. Table 1 and 2, Figure 5 and 6, shows the outcomes of various specific, homo & heterogeneous classifiers in contrast with the proposed NGBFA - hybrid model stacking. Figure 7 shows the ROC-AUC curve values.

Table 1: Comparison of various Individual Classifiers for Biopsy

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	78.6	78.5	79	78.5
SVM	79.9	70	66	65
LR	57.7	59	58	57
NB	60.2	54	54	53
RF	93.4	84	84	84
DT	64.7	56	55	53
BC	88.29	84	74	74
MLP	88	76	75	76

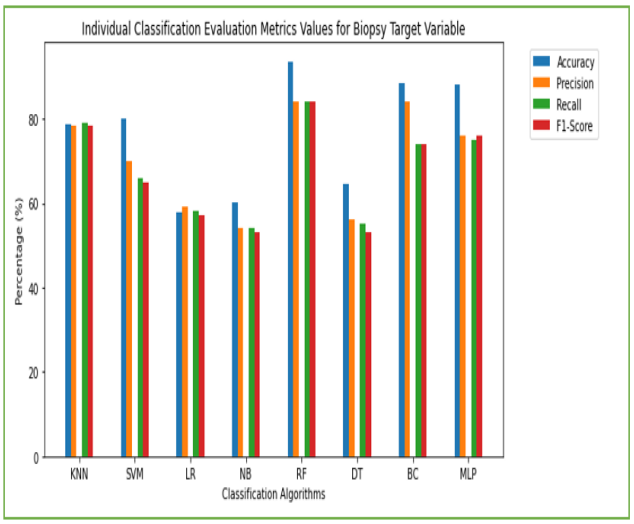

Figure 5: Comparison Chart of Individual Classifier Evaluation Metrics for Biopsy

Comparison of Proposed Hybrid Model Stacking with Traditional Ensemble Classifier Models

Table 2: Comparison of various Homo & Heterogenous Classifiers with Hybrid Model Stacking for Biopsy

Classifier	Accuracy	Precision	Recall	F1-Score
Ensemble KNN	86.77	89	87	86
Ensemble SVM	92.99	93	93	93
Ensemble LR	60.31	60	60	60
Ensemble NB	62.25	62	62	61
Ensemble RF	95.71	96	96	96
Ensemble DT	88.32	88	88	88
Ensemble BC	96.10	96	96	96
Ensemble MLP	93.77	94	94	94
NGBFA - Hybrid Model Stacking	**98**	**97.6**	**98**	**97.6**

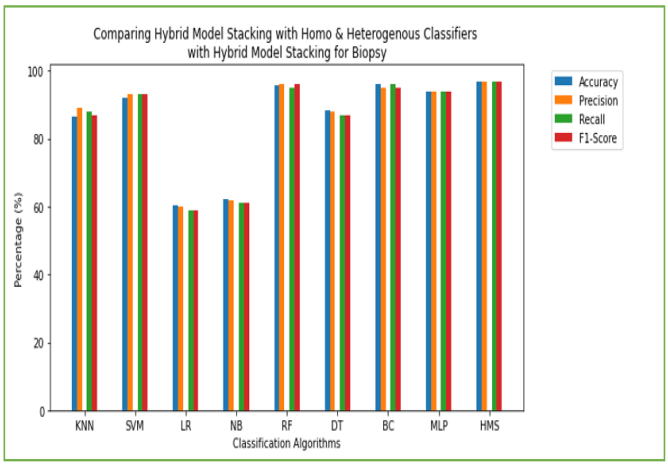

Figure 6: Graph Comparing Hybrid Model Stacking with Homo & Heterogenous Classifiers with Hybrid Model Stacking for Biopsy

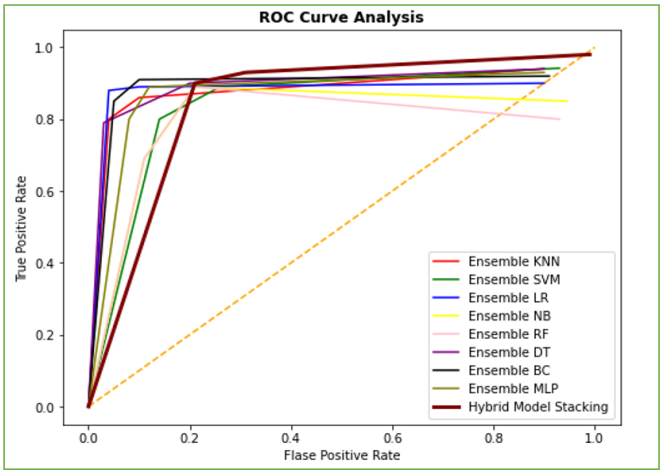

Figure 7: ROC Curve Analysis with Hybrid Model Stacking for Biopsy

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Result of target feature: Hinselmann Individual Classifier Performance Analysis

Because of NGBFA-RF, age, number_of_sexual_partners, first_sexual_intercourse, num_of_pregnancies, Smokes, smoking (packs/year), harmonal_contraceptives (years), Dx: Cancer, IUD 9 optimal features were selected as important features for the Hinselmann target feature. We tabulated the experimental results in Table 2 for the Hinselmann target variable with an accuracy of 97% with the proposed model. Tables 3 and 4 and Figures 8 and 9 show the outcomes of various specific homo & heterogeneous classifiers, contrasting with the proposed NGBFA - hybrid model stacking. Figure 10 shows the ROC-AUC curve values.

Table 3: Comparison Chart of Individual Classifier Evaluation Metrics for Hinselmann

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	83	79	79	78
SVM	85	79	78	78
LR	70	72	71	71
NB	68	71	65	64
RF	95	86	87	86
DT	76	76	74	73
BC	88	84	71	71
MLP	92	85	83	83

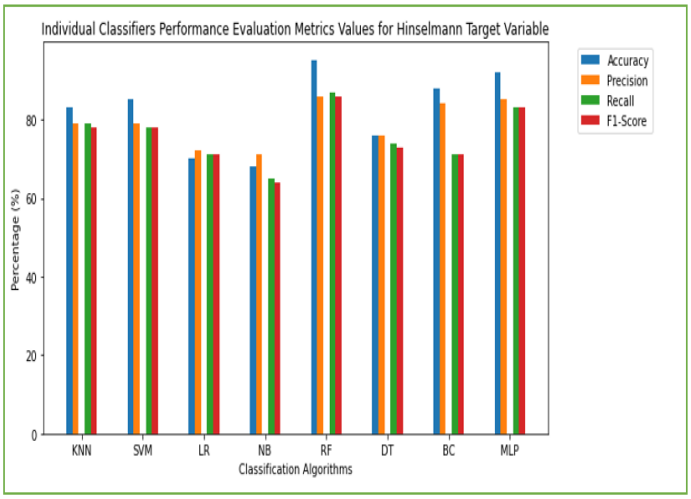

Figure 8: Comparison Chart of Individual Classifier Evaluation Metrics for Hinselmann

Comparison of Proposed Hybrid Model Stacking with Traditional Ensemble Classifier Models

Table 4: Comparison of various Homo & Heterogenous Classifiers with Hybrid Model Stacking for Hinselmann

Classifier	Accuracy	Precision	Recall	F1-Score
Ensemble KNN	90.53	92	91	91
Ensemble SVM	94.69	95	95	95
Ensemble LR	73.10	75	73	73
Ensemble NB	73.48	74	73	74
Ensemble RF	94.69	95	95	95
Ensemble DT	93.56	94	94	94
Ensemble BC	96.21	96	96	96
Ensemble MLP	95.07	96	95	95
NGBFA - Hybrid Model Stacking	**97**	**96**	**97**	**97**

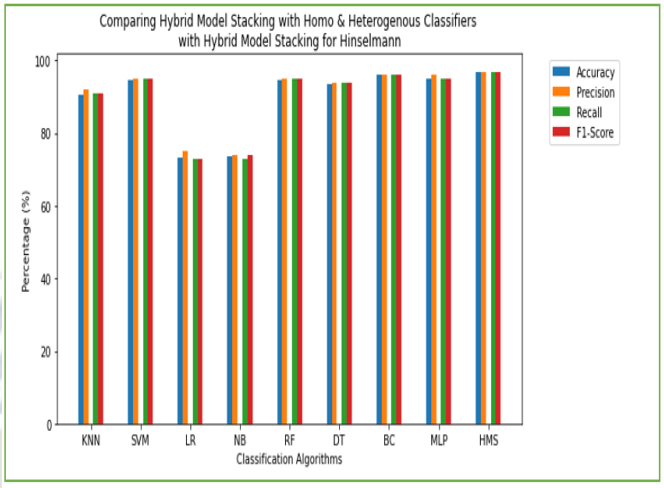

Figure 9: Graph Comparing Hybrid Model Stacking with Homo & Heterogenous Classifiers with Hybrid Model Stacking for Hinselmann

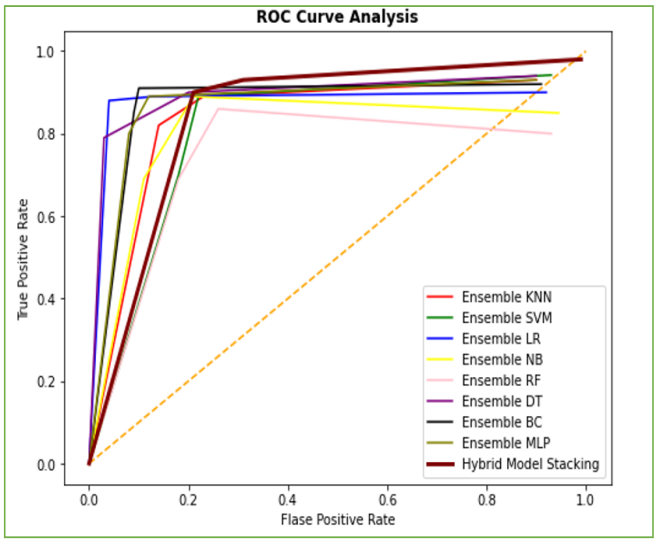

Figure 10: ROC Curve Analysis with Hybrid Model Stacking for Hinselmann

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Result of target feature: Schiller Individual Classifier Performance Analysis

Because of NGBFA-RF, age, harmonal_contraceptives (years), first_sexual_intercourse, num_of_pregnancies, IUD (years), smokes (years), smokes (packs/year), STDs (number), STDs: HIV, STDs: vulvo_perineal_condylomatosis, STDs: Hepatitis B 11 optimal features were selected as important features for the Schiller target feature. We tabulated the experimental results in Table 3 for the Schiller target variable with an accuracy of 96% with the proposed model. Tables 5 and 6 and Figures 11 and 12 show the outcomes of various specific homo & heterogeneous classifiers in contrast with the proposed NGBFA - hybrid model stacking. Figure 13 shows the ROC-AUC curve values.

Table 5: Comparison Chart of Individual Classifier Evaluation Metrics for Schiller

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	80	70	70	69
SVM	77	69	65	64
LR	64	66	65	65
NB	65	70	64	63
RF	93	88	87	88
DT	71	64	63	64
BC	94	91	90	90
MLP	92	81	80	81

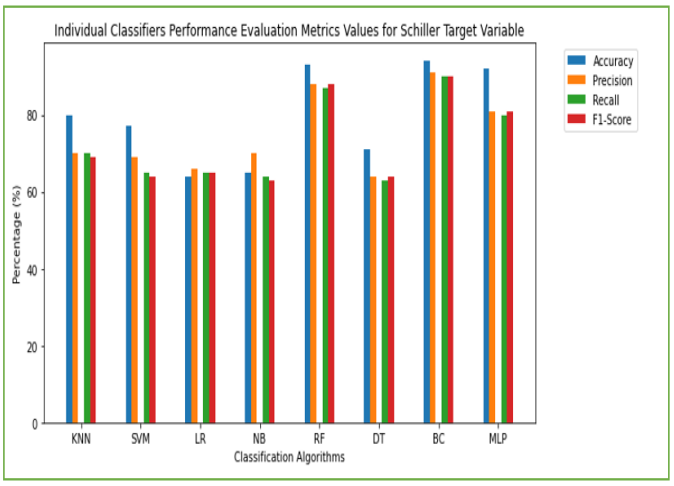

Figure 11: Comparison Chart of Individual Classifier Evaluation Metrics for Schiller

Comparison of Proposed Hybrid Model Stacking with Traditional Ensemble Classifier Models

Table 6: Comparison of various Homo & Heterogenous Classifiers with Hybrid Model Stacking for Schiller

Classifier	Accuracy	Precision	Recall	F1-Score
Ensemble KNN	89.68	91	90	90
Ensemble SVM	92.85	93	93	93
Ensemble LR	69.84	70	70	69
Ensemble NB	69.04	72	69	67
Ensemble RF	94.84	95	95	95
Ensemble DT	88.88	89	89	89
Ensemble BC	95.63	96	96	96
Ensemble MLP	94.84	95	95	95
NGBFA - Hybrid Model Stacking	**96.03**	**95**	**96**	**96**

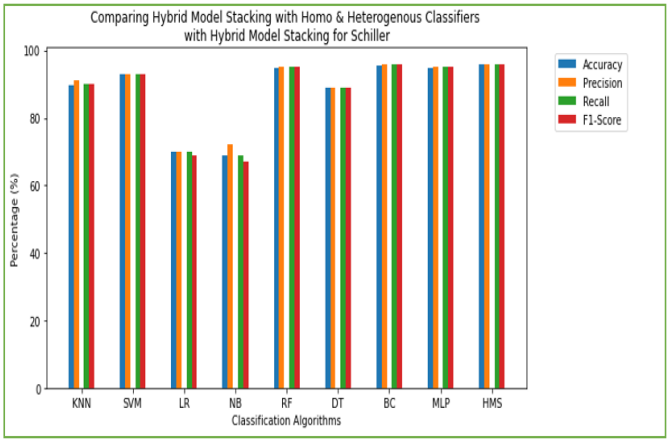

Figure 12: Graph Comparing Hybrid Model Stacking with Homo & Heterogenous Classifiers with Hybrid Model Stacking for Schiller

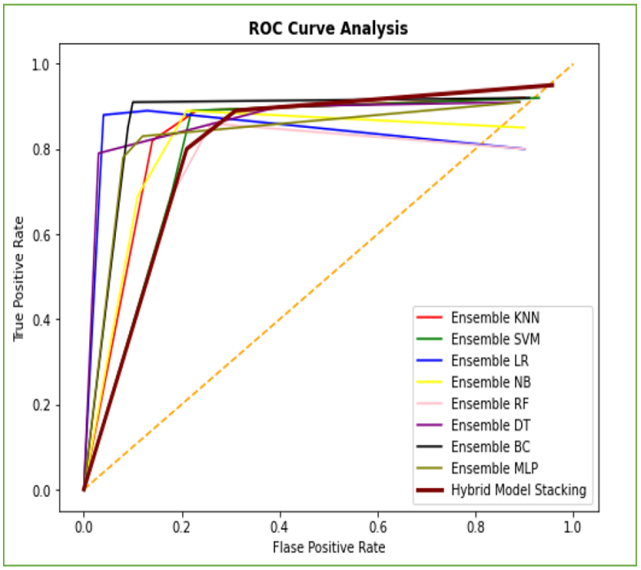

Figure 13: ROC Curve Analysis with Hybrid Model Stacking for Schiller

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Result of target feature: Citology Individual Classifier Performance Analysis

Because of NGBFA-RF, age, number_of_sexual_partners, first_sexual_intercourse, num_of_pregnancies, and IUD (years), we selected five optimal features as important features for the Citology target feature. We tabulated the experimental results in Table 4 for the Citology target variable with an accuracy of 93% with the proposed model. Tables 7 and 8 and Figures 14 and 15 show the outcomes of various specific homo & heterogeneous classifiers, contrasting with the proposed NGBFA - hybrid model stacking. Figure 16 shows the ROC-AUC curve values.

Table 7: Comparison Chart of Individual Classifier Evaluation Metrics for Citology

Classifier	Accuracy	Precision	Recall	F1-Score
KNN	75	61	63	61
SVM	77	70	68	68
LR	61	57	56	56
NB	58.9	59	58	59
RF	90	80	79	79
DT	78.6	76	76	75
BC	87	88	83	84
MLP	89	78	79	78

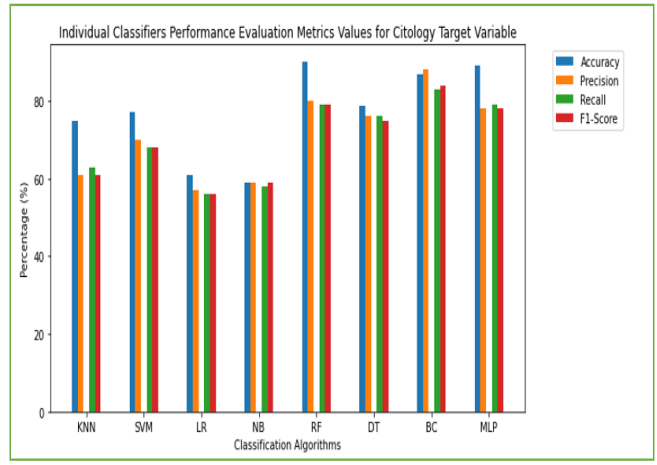

Figure 14: Comparison Chart of Individual Classifier Evaluation Metrics for Citology

Comparison of Proposed Hybrid Model Stacking with Traditional Ensemble Classifier Models

Table 8: Comparison of various Homo & Heterogenous Classifiers with Hybrid Model Stacking for Citology

Classifier/Metric	Accuracy	Precision	Recall	F1-Score
Ensemble KNN	86.59	89	87	86
Ensemble SVM	84.67	85	85	85
Ensemble LR	59.38	60	59	59
Ensemble NB	56.32	58	56	56
Ensemble RF	92.25	94	94	94
Ensemble DT	92.72	93	93	93
Ensemble BC	91.18	91	91	91
Ensemble MLP	90.42	91	90	90
NGBFA - Hybrid Model Stacking	**93**	**92**	**93**	**93**

Table 9: Comparison of various Homo & Heterogenous Classifiers and Hybrid Model Stacking with MCC and Cohen's kappa

Classifier	Biopsy		Hinselmann		Schiller		Citology	
	CK	MCC	CK	MCC	CK	MCC	CK	MCC
Ensemble KNN	0.485	0.498	0.489	0.588	0.444	0.510	0.499	0.587
Ensemble SVM	0.512	0.523	0.547	0.547	0.554	0.574	0.574	0.546
Ensemble LR	0.521	0.524	0.578	0.547	0.524	0.574	0.596	0.569
Ensemble NB	0.523	0.547	0.654	0.541	0.574	0.585	0.665	0.588
Ensemble RF	0.564	0.561	0.541	0.569	0.574	0.514	0.554	0.556
Ensemble DT	0.510	0.531	0.578	0.548	0.565	0.585	0.577	0.512
Ensemble BC	0.498	0.514	0.547	0.611	0.445	0.569	0.536	0.636

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Ensemble MLP	0.523	0.587	0.576	0.654	0.566	0.536	0.574	0.674
NGBFA - Hybrid Model Stacking	**0.710**	**0.720**	**0.724**	**0.777**	**0.754**	**0.765**	**0.785**	**0.784**

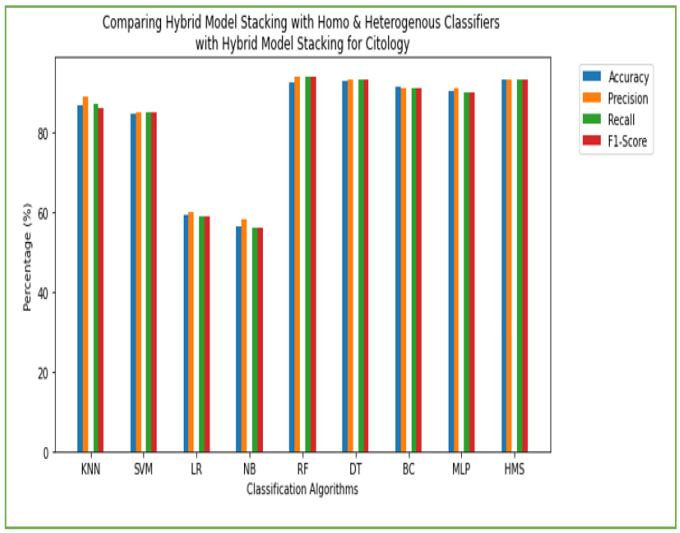

Figure 15: Graph Comparing Hybrid Model Stacking with Homo & Heterogenous Classifiers with Hybrid Model Stacking

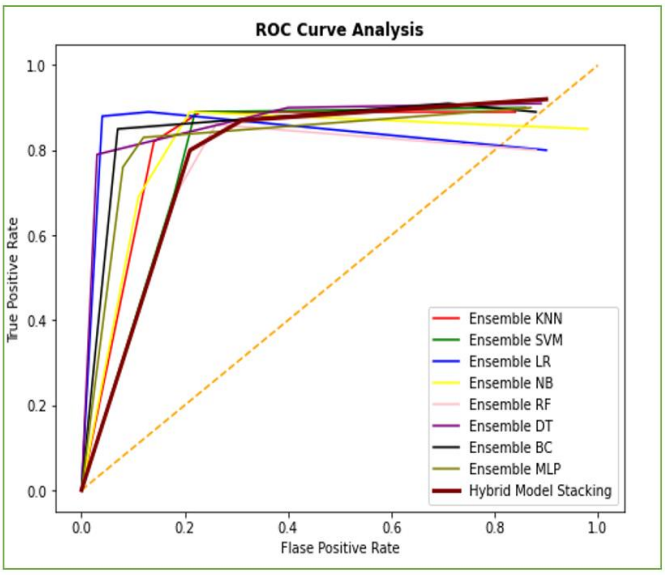

Figure 16: ROC Curve Analysis with Hybrid Model Stacking for Citology

The Cohen's kappa values of Hybrid Model Stacking are the greatest in the four target or independent characteristics (Biopsy, Hinselmann, Schiller, Citology), 0.710, 0.724, 0.754, and 0.785, respectively, according to Table 9 and Figure 17. Hybrid Model Stacking achieves a significant degree, based to the previous explanation of Cohen's kappa. Furthermore, as demonstrated in Table 9, Hybrid Model Stacking performs well regarding MCC in the four target or independent qualities.

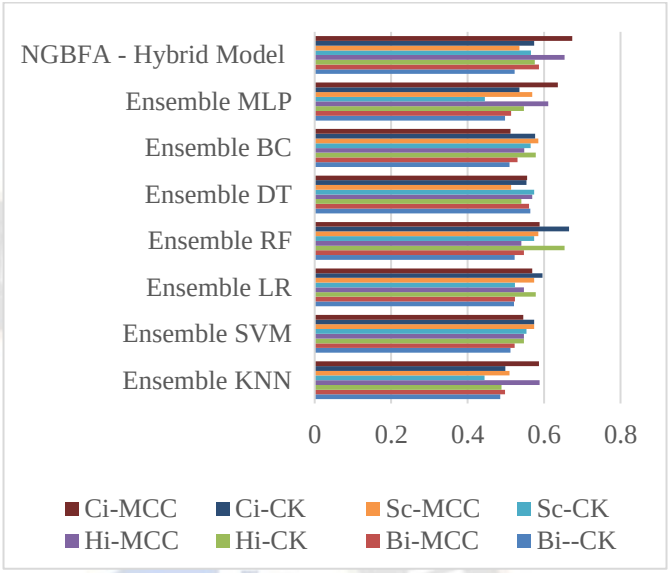

Figure 17: Comparison of various Homo & Heterogenous

V. CONCLUSION

To analyze numerous cervical cancer possibilities, a mathematical machine learning-based model was developed in this study. A review of statistical methods, machine learning, and methodologies that can assist in detecting cervical cancer is presented after the authors identify the areas of research that need more attention. In addition, the study has included eight classification algorithms to form a hybrid ensemble classifier to predict CC accurately. To develop and evaluate all modeling techniques, we have examined optimal prospects. According to the collected dataset, the proposed methodology's accuracy and other evaluation metrics have been analyzed for all four target variables by the proposed algorithm, a robust model stacking: a hybrid approach classifier consisting of two levels. Level 1 consists of homogenous ensembles of various classifiers, and level 2 consists of heterogenous classifiers using a soft voting algorithm for predicting cervical cancer through the risk factors composed of the UCI repository. SMOTE methodology has been used to address the imbalance problem of the dataset. The NGBFA-HMS has been applied to all four target variables and got better outcomes compared to other individual classification algorithms. Cohen's kappa values of Hybrid Model Stacking are the highest in the four target or independent attributes (Biopsy, Hinselmann, Schiller, Citology), 0.710, 0.724, 0.754 and 0.785. As a future enhancement, it is important to stimulate research into cervical cancer prediction using emerging technologies and methods. Several socio-demographic factors can be considered, such as the level of education in the region where the sample data was collected. Schools and educational institutions can be important in extending awareness of better healthcare to their students' families.

REFERENCES

[1] Organization CDC, Centers for Disease Control and Prevention retrieved from https://www.cdc.gov/cancer/cervical/basic_info/index.htm#:~:text=When%20cancer%20starts%20in%20the,at%20risk%20for%20cervical%20cancer. Last accessed on 06/04/2022.

[2]. W. H. O. R. Health, W. H. O. C. Diseases, and H. Promotion, Comprehensive Cervical Cancer Control: A Guide to Essential Practice, World Health Organization, Geneva, Switzerland, 2006.

[3] Organization W H 2022 retrieved from https://www.who.int/health-topics/cervicalcancer#tab=tab_1 last accessed 06-04-2022.

[4] Organization W H 2021 retrieved from https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

[5] Bosch FX, de Sanjosé S. Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality. J Natl Cancer Inst Monogr. 2003;(31):3-13. doi: 10.1093/oxfordjournals.jncimonographs.a003479. PMID: 12807939.

[6] Ch. Bhavani, Dr. A. Govardhan. Supervised Algorithms of Machine Learning in the Prediction of Cervical Cancer: A Comparative Analysis. Annals of the Romanian Society for Cell Biology, 1380–1393. 2021, Retrieved from https:// www.annalsofrscb.ro/index.php/journal/article/view/4580

[7] Organization ASCO retrieved from https://www.cancer.net/cancer-types/cervical-cancer/symptoms-and-signs last accessed 06/04/2022.

[8] Jaswinder Singh & Sandeep Shrama. (2019). Prediction of Cervical Cancer Using Machine Learning Techniques. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11 (2019) pp. 2570-2577.

[9] Akter, L., Ferdib-Al-Islam, Islam, M.M. *et al.* Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques. *SN COMPUT. SCI.* **2,** 177 (2021). https://doi.org/10.1007/s42979-021-00551-6.

[10] Mohamed Hosni, Ibtissam Abnane, Ali Idri, Juan M. Carrillo de Gea, José Luis Fernández Alemán,Reviewing ensemble classification methods in breast cancer, Computer Methods and Programs in Biomedicine, Volume 177, 2019, Pages 89-112, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2019.05.019. (https://www.sciencedirect.com/science/article/pii/S0169260719301907)

[11] A. Idri, M. Hosni, A. Abran, Systematic Mapping Study of Ensemble Effort Estimation, in: Proc. 11th Int. Conf. Eval. Nov. Softw. Approaches to Softw. Eng., 2016: pp. 132–139. doi:10.5220/0005822701320139.

[12] R.E. Schapire, Measures of Diversity in Classifier Ensembles, Mach. Learn. 51 (2003) 181–207. doi:10.1049/ic:20010105.

[13] Pintelas P, Livieris IE. Special Issue on Ensemble Learning and Applications. *Algorithms*. 2020; 13(6):140. https://doi.org/10.3390/a13060140.

[14] Karlos, S.; Kostopoulos, G.; Kotsiantis, S. A Soft-Voting Ensemble Based Co-Training Scheme Using Static Selection for Binary Classification Problems. Algorithms 2020, 13, 26.

[15] Sabzevari, M., Martínez-Muñoz, G. & Suárez, A. Building heterogeneous ensembles by pooling homogeneous ensembles. *Int. J. Mach. Learn. & Cyber.* **13,** 551–558 (2022). https://doi.org/10.1007/s13042-021-01442-1.

[16] J. L. Fernández-Alemán, J. M. Carrillo-de-Gea, M. Hosni, A. Idri and G. García-Mateos, "Homogeneous and heterogeneous ensemble classification methods in diabetes disease: a review," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 3956-3959, doi: 10.1109/EMBC.2019.8856341.

[17] Esra'a Alshdaifat, Malak Al-hassan, Ahmad Aloqaily, Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers, ICT Express, Volume 7, Issue 3, 2021, Pages 342-349, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2020.11.005. (https://www.sciencedirect.com/science/article/pii/S2405959520304872).

[18] de Oliveira JM, dos Santos EM, Carvalho JRH, de Vasconcelos Marques LA (2013) Ensemble of heterogeneous classifiers applied to lithofacies classification using logs from different wells. International joint conference on neural networks. IJCNN, Dallas, TX, USA, pp 1–6.

[19] Nanni L, Brahnam S, Ghidoni S, Lumini A (2015) Toward a general-purpose heterogeneous ensemble for pattern classifcation. Comput Intell Neurosci 2015:85.

[20] Haque MN, Noman N, Berretta R, Moscato P (2016b) Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classifcation. PloS One 11(1):e0146116.

[21] Alshdaifat E, Al-hassan M, Aloqaily A (2020) Effective heterogeneous ensemble classification: an alternative approach for selecting base classifiers. ICT Express, South Korea.

[22] Luong, Anh Vu, Vu, Trung Hieu, Nguyen, Phuong Minh, Van Pham, Nang, McCall, John, Liew, Alan Wee-Chung, Nguyen, Tien Thanh, "A homogeneous-heterogeneous ensemble of classifiers**.**", , 2020.

[23] Nanglia, S., Ahmad, M., Khan, F. A., & Jhanjhi, N. Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing and Control*, *72*, 103279.

[24] Feng Y, Wang X, Zhang J. A heterogeneous ensemble learning method for neuroblastoma survival prediction. IEEE J Biomed Health Inform. 2021 Apr 13;PP. doi: 10.1109/JBHI.2021.3073056. Epub ahead of print. PMID: 33848254.

[25] Fernandes, K., Chicco, D. and Cardoso, J.S. (2018) Supervised deep learning embeddings for the prediction of cervical cancer diagnosis. PeerJ, 4, 1–20.

[26] Mehmood M, Rizwan M, Gregus ml M and Abbas S (2021) Machine Learning Assisted Cervical Cancer Detection. *Front. Public Health* 9:788376. doi: 10.3389/fpubh.2021.788376.

[27] Pinar Yildirim. 2018. Classification with incomplete data and ensemble learners for the prediction of cervical cancer risk. In Proceedings of the International Conference on Intelligent Science and Technology (ICIST '18). Association for Computing Machinery, New York, NY, USA, 1–5. DOI:https://doi.org/10.1145/3233740.3233741.

[28] Barua, Lipika, M. Sharif, and Tania Akter. "Analyzing Cervical Cancer by using an Ensemble Learning Approach based on Meta Classifier." *International Journal of Computer Applications* 182 (2019): 29-33.

[29] Gupta, S., & Gupta, M. K. (2021). Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm. *The Computer Journal*.

[30] CH. Bhavani, A. Govardhan, Cervical cancer prediction using stacked ensemble algorithm with SMOTE and RFERF, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3451-3457, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.269. (https://www.sciencedirect.com/science/article/pii/S2214785321051282)

[31] CH Bhavani, A. Govardhan, A Novel Genetic Inspired Binary Firefly Algorithm For Feature Selection in the Prediction of Cervical Cancer, Advances in Data Science and Adaptive Analysis, doi: 10.1142/S2424922X23500031.

[32] Peng L, Zhang H, Zhang H, Yang B (2017) A fast feature weighting algorithm of data gravitation classification. Inf Sci 375:54–78. https://doi.org/10.1016/j.ins.2016.09.044

[33] Hsu W-Y (2018) A decision-making mechanism for assessing risk factor significance in cardiovascular diseases. Decis Support Syst 115:64–77. https://doi.org/10.1016/j.dss.2018.09.004