ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2024.1631.311

BILEVEL OPTIMIZED RECURSIVE FEATURE ELIMINATOR FOR CERVICAL CANCER FEATURE SELECTION PROCESS

S. Nandhinieswari[1], A. Indumathi[2]

[1]Research Scholar, Kongunadu Arts and Science College, Coimbatore, India;
Assistant Professor, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India. email: nandhinics@srcw.ac.in,
orcid: https://orcid.org/0009-0002-7395-9311
[2]Associate Professor, Kongunadu Arts and Science College (Autonomous), Coimbatore, India. email: indumathia_ca@kongunaducollege.ac.in,
orcid: https://orcid.org/0000-0001-6783-8922

SUMMARY

An immense need has emerged in several areas of biological area for the development of prediction algorithms capable of managing the increasing complexity of high-dimensional information. In developing countries, Cervical Cancer (CC) kills more women than any other disease or accident, and it's the top cause of death among women worldwide. Early detection and treatment lead to improved results and longer patient survival, which in turn reduces cancer mortality. For the majority of real-world data science problems, not all dataset variables are useful for building models. The accuracy of a classifier and the model's ability to generalize are both reduced by repeated variables. Furthermore, adding more variables increases the overall complexity of a model. Deep learning's feature selection approach is a good fit for this issue. When it comes to selecting features for linear regression, our novel Bilevel Optimized Recursive Feature Eliminator (BORFE) method represents a revolutionary development. Finding the optimal fit for a model and eliminating its most undesirable features is the objective of this innovative feature selection method. This study presents a novel cross-validation approach using a Bilevel Optimization–Based -Recursive Feature Extractor (BORFE) to perform an in-depth analysis of the hyper-parameters. When used with cross-validation, RFE finds the optimal number of features and the optimum selection of ranking features. According to the evaluation metrics, BORFE performs better than the other conventional algorithms when it comes to FS on cervical cancer datasets. For the cervical cancer dataset, this shows that BORFE can solve FS problems successfully.

Key words: *cervical cancer, feature selection, feature elimination, recursive feature extractor, optimization.*

Received: August 22, 2024; Revised: October 25, 2024; Accepted: November 18, 2024; Published: December 24, 2024

INTRODUCTION

The second-most frequent kind of cancer in women, after breast cancer, is gynaecological cancer. It starts in a woman's reproductive system. Women diagnosed with gynaecological cancers have a significantly reduced lifetime due to the seriousness of this disease. Gynecological cancers include cervical cancer among others, including ovarian, uterine, vaginal, and vulvar cancers. Each subtype of gynaecological cancer is associated with its own unique set of risk factors. Worldwide, 7.5% of female

cancer fatalities are attributable to cervical cancer, making it the second most prevalent cancer in women [1]. The tumour known as cervical cancer develops when cells in the cervix tissue start to divide and multiply uncontrollably, without proper regulation. Early identification may prevent most occurrences of cancer, but in those where the tumour is malignant, the cells can travel via the bloodstream and infect other areas of the body. More characteristics and missing values are often included in medical datasets [3][6][33]. By optimization, it is essential to identify the relevant and significant characteristics for statistical model development [29]. Many different kinds of cancer research have made heavy use of Machine Learning (ML) techniques because of how much better they are at making predictions and doing optimization-related investigations. Using ML approaches, the research [4] showed accurate findings in cancer prediction and diagnosis [36]. The study included a variety of relevant works. When it comes to data mining, machine learning, and statistics, R is among the most utilized and well-known software frameworks. Machine learning experiments may benefit from the creative, user-friendly, and extensible domain-specific functions provided by the R packages [5][27][39]. The accuracy of the model can then be assessed using a variety of assessment standards, leading to improved performance efficiency.

When cross-validating, the hyper-parameters are chosen so that the out-of-sample generalization error is as small as possible. The usual procedure involves defining a grid across the relevant hyperparameters along with doing cross-validation of 10-fold on all values of the grid [2][44]. Parameter issues are common in the analysis of data and can come about in many different ways. For example, they are common in feature selection, kernel creation, parallel learning, and many others. Greedy procedures including genetic algorithms, filter methods, stepwise regression, and backward elimination are used for such high-dimensional problems [9][21][31][35][37][38]. Along with their practical inefficiency, these heuristic models, like grid searching, also possess the fundamental defect of being unable to ensure the quality of the "solution" generated. This discretization in grid search also ignores the continuous nature of the model parameters, which is another limitation. It has been recently shown that the parameter for normalization is continuous in research on finding the complete regularization route of SVM. Specifically, this paper [13] states that selecting a single regularization coefficient C is crucial as well as it demonstrates that computing the SVM solution for any conceivable value of the variable is quite tractable [8] [34]. The selection of suitable priors becomes difficult once Bayesian approaches have treated model parameters as random variables. Ultimately, when it comes to choosing values for parameters, out-of-sample testing remains standard. Improving approaches that integrate strong theoretical grounding with efficient computing is an immediate need from the perspective of "optimizing model selection" by out-of-sample estimations. An approach based on bilevel optimization methods is proposed in this research.

One feature selection strategy that has been developed recently for small sample classification issues is Recursive Feature Elimination (RFE) [14]. RFE was first used to classify cancer using microarrays, where there are fewer than 100 training samples and many thousands of features. It has since become a useful method for choosing features from small samples. Using RFE, we can increase the accuracy of generalization by removing features that aren't essential and will not significantly affect training errors. Furthermore, Support Vector Machines (SVMs) are closely connected to RFE and have shown good generalizability, especially for small sample classifications [16]. Despite its promising results in small-sample feature selection situations, RFE has a tendency to keep independent features while removing weak or redundant ones. First, as mentioned in [15], features that are likely redundant may enhance classification [7]. Second, combining two inadequate features that aren't useful on their own might lead to a noticeable performance improvement. Therefore, it is possible to reduce classification performance by only eliminating characteristics that are weak or redundant. SVM the ability to solve convex optimization problems using hyperparameters that the user may choose is a fundamental component of several Deep Learning statistical methodologies. When deciding on these criteria, many individuals employ and agree upon cross-validation. As a result, it reduces the number of useful hyper-parameters in the model due to it employs a grid-search method to find them. This occurs because of the large number of grid points in the high-dimensional space [10, 11, 12].

This study presents a novel cross-validation approach based on a Recursive Feature Extractor (RFE) that is based on bilevel optimization. It does a thorough analysis of the hyper-parameters. When used with

cross-validation, RFE finds the optimal number of features and the optimum selection of ranking features. According to the evaluation metrics, BORFE performs better than the other conventional algorithms when it comes to FS on cervical cancer datasets. The features selected by BORFE play an important part in the differential deep learning network, as shown by certain biological activities. The conventional procedure for feature removal for each state is reconsidered by BORFE. The original RFE has zero bearing on the next stage, but BORFE will keep weak or redundant characteristics that might be valuable when paired with others, which is the fundamental distinction between the two methods.

The main contributions of this research are given below:

- We introduced a novel RFE based on bilevel optimization for the important feature selection process.
- Bilevel optimization is used to fine-tune the REF. The primary difference between the two approaches is that BORFE will retain weak or redundant qualities that might be useful when combined with others, whereas the initial RFE has no impact on the subsequent stage.
- For the cervical dataset, we eliminated minority classifications during data pre-processing. To further guarantee parity between the normal and abnormal classes, we used a resampling approach.
- The experimental results obtained on the cervical cancer dataset demonstrated that our proposed approach can reduce feature dimension when compared to existing algorithms. This can help avoid a different kind of overfitting problem that is common with classes with a limited number of training samples.

The remainder of the paper was structured as follows: We reviewed previous research on feature selection methods for cervical cancer datasets in the "Related works" section II. Section III consists of the "Proposed method" of BORFE feature selection algorithms. We detailed our experiments and offered our findings in section IV under "Experiments and results". Section V of "Conclusion & future development" included the results and next initiatives.

RELATED WORK

The prediction model for breast cancer prognosis and diagnosis was greatly improved by integrating classifier approaches and feature selection techniques, according to the research on women's cancer [17,18,22,28]. Aiming to identify the most significant risk factors, the research on staging predictions in cervical cancer [19] retrieved rules from the dataset based on signs and symptoms and used decision tree classifiers. In order to achieve data equality, cervical cancer research [20] used RUS and ROS techniques. For the purpose of feature selection, the Stability Selection (SS) approach was used. The original dataset was cleaned up for 190 occurrences of missing values ('?', Null) in the current research. A total of 668 entries made up the raw dataset. In this case, the SS technique and the RF algorithm were combined to form the learning model. The RUS and ROS approaches were used to test the success of this model. This study attained 98% accuracy using a ROS-based SS approach, which was more effective than an RUS-based SS method on this dataset. Using cervical cancer data, another study [32] chose 25 features for prediction using the KNN method, 17 features using the decision tree classifier, and 11 features using the random forest algorithm. Based on the results, the KNN method seems to be the superior model over the decision tree and random forest algorithms, with an Area Under the Curve (AUC) of 0.822. However, the algorithms examined here used different numbers of samples for training and testing, all taken from the cervical cancer dataset. In order to deal with the unbalanced data, the research [40] used oversampling, under-sampling, and mixed-sampling approaches to categorize cervical cancer data. This approach achieved 97% accuracy using a decision tree classifier after prioritizing six features. Research-based on observation has shown that the cervical cancer dataset used in a number of publications has had missing value cases eliminated and that finding important features has received less weight. Handling missing values in the dataset, identifying exact qualities, and achieving the outcomes of improved prediction accuracy via optimization are all subsequent challenges. Consequently, the goal of this research is to overcome these obstacles.

PROPOSED METHODOLOGY

Here we presented the high-level framework of our proposed model, as shown in Figure 1. A training set and a test set are provided by the cervical dataset, which comprises 32 risk variables for 858 samples. The dataset is encoded via data pre-processing as it cannot be utilized directly in the Deep Convolutional Neural Network (DCNN) model. We cleaned the data, removed minorities, oversampled, encoded, and normalized the dataset as part of the data pre-processing. A training set, a validation set, and a test set were created from the dataset after data preparation. The feature selection and training procedures make use of the training and validation sets, while the test set is used to ensure the model's ultimate performance is verified. Following this, we optimized the feature subset even more by using BORFE feature removal to the reduced features. We trained the DCNN model using the optimum feature subset that was produced after feature selection. Our proposed model's effectiveness was demonstrated by its final performance on the test set.

Preprocessing the data

An essential component of effective model learning is data processing, also called data engineering. Data processing includes cleaning the columns and rows, encoding the features, and normalizing the data. In order to get the data ready for analysis, this section goes over the steps used in the preprocessing phase.

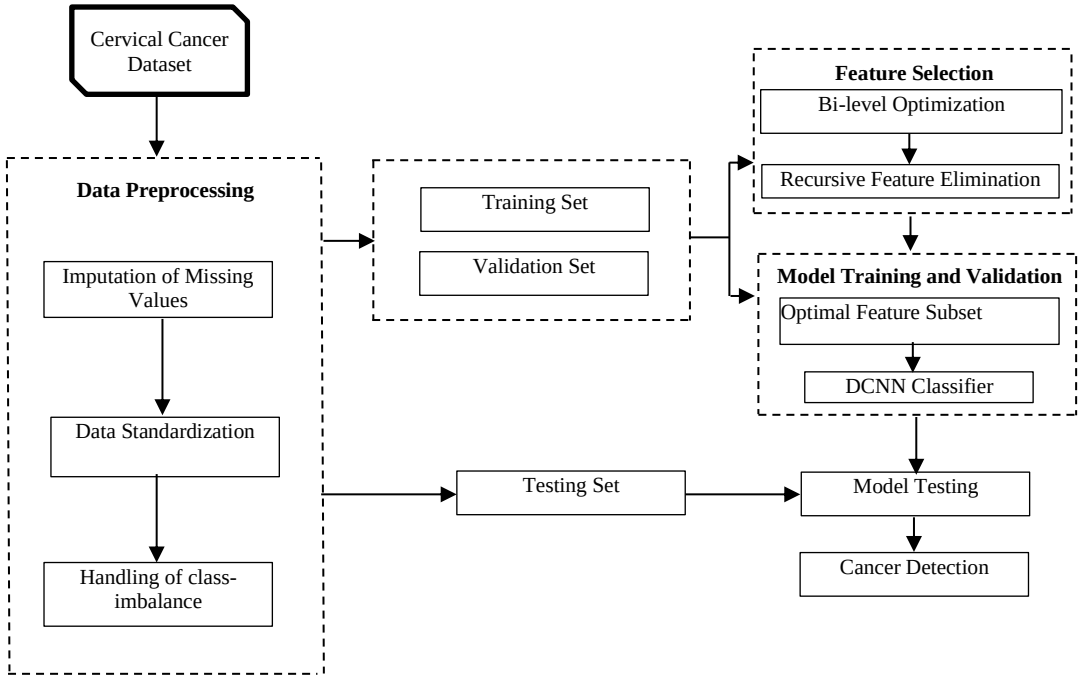

Figure 1. BORFE-DCNN proposed model

Imputation of Missing Values

Owing to the high number of missing values in the dataset labelled "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis," we eliminated these attributes from the dataset that included sexually transmitted infections. Therefore, these features were not included in the subsequent analysis. Utilizing the k-Nearest-Neighbours (KNN) method, missing values have been allocated in order to make predictions about the standards of the incoming data points using the concept of "feature similarity." It follows that the new point's assignment is based on its similarity to the points allocated to the training set. Identifying the nearest neighbours of the k, is particularly helpful for making predictions about the missing values. The final count was 734 occurrences from 858 records; 124 entries were removed and their null values were filled using KNN imputations.

Data Standardization

Through the process of data standardization, the dataset's independent features/attributes (columns) are transformed into the interval [0, 1] [23]. It eliminates data skewness by moving individual characteristics to have a zero average and unit variance. To get the standardization (Z_x), we use the z-score formula stated in Equation (1).

$$Z_x = \frac{x - \mu}{\sigma} \qquad (1)$$

In this case, x stands for the data instance, while μ and σ denote the feature's mean and standard deviation, respectively.

Handling of Class-imbalance

The results were based on the dataset's unevenly distributed ratio, which was 95% healthy class data and 5% cancer cases. There is bias in the predictions since DL classifiers choose to train on data from the class with the most occurrences. The sensitivity prediction will be unsuccessful due to the unbalanced dataset, even if the specificity (actual negative rate) may be high. A classifier that predicts 100% specificity and 0% sensitivity may be very accurate, but it's only on the surface. This is because classifying cervical cancer cases (sensitivity) and healthy people (specificity) are both important tasks, but the previous one is given more weight because it correctly finds the cases that need immediate medical attention [24, 26]. Serious health consequences might result from a wrong diagnosis of cancer cells. Because samples are only partially distributed across classes, addressing class imbalance is crucial for producing accurate outcomes [23] [25]. Due to the small size of the minority class in this instance and the fact that the under-sampling class balancing approach ignores a substantial portion of the data [26], it is not used in the research.

Feature Selection

Feature selection seeks to isolate a problem domain's most salient features. Increases in both computing speed and prediction accuracy are made. To find the most important features, we use the novel feature selection approach known as the Bilevel-Optimized Recursive Feature Elimination (BORFE). Figure 2 shows the feature extractor procedure of BORFE.

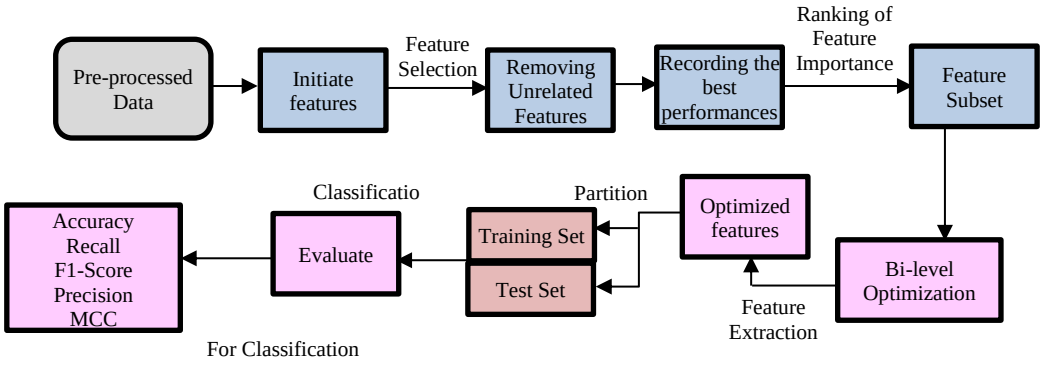

Figure 2. Feature selection process of BORFE

As a wrapper feature selection approach, RFE iteratively eliminates features based on deep learning performance, evaluating their value in the process [38]. Every time you run RFE, features that are not essential are removed. This process continues until you achieve the maximum performance or the amount of features the model sets. The input training set and validation set of our RFE method (see Algorithm 2) only include the category features and reduced numeric features from the previous stage. In addition to the features chosen in the first step, the method also takes a positive integer patient p and a list $of\ init\_features$ as inputs. The introduction of patient p ensures that RFE is stopped promptly if greater performance cannot be achieved after several iterations. $Init\_features$ may be used to decrease

the search area of RFE without having to start from the beginning with all features. It is necessary to initialize variables before recursive feature removal. In the worst-case scenario, the number of RFE iterations is determined by f_len, which reflects the number of starting features. Recording the best performances during RFE is done using $best\_performance$. Removing numerical attributes during RFE is done in three places: rm_list, which stores the subset of features chosen for optimal performance, and retain features, which maintains the features chosen after each RFE iteration. Each iteration of recursive feature elimination begins with the initialization of a dictionary performance dict, which will include the validation efficiency with MLP after each feature is eliminated. A total of ten separate tests, each with its own unique random seed, are averaged to get the score in the evaluate elimination function. After then, patient p decides whether to keep taking RFE. A single RFE iteration is carried out and the local highest efficiency is reached if patient p is greater than 0. Results from a comparison between global and local best performances are used to update both the global highest scores and the features that have been chosen.

Bi-level Optimization

All of the model parameters are found by minimizing the outer objective function in the bilevel optimization problem (1):

The input data is represented by $x^{(k)} \in X$ and the targets/labels are supplied by $y^k \in Y$. We are given m sample pairs (x^k, y^k) where $k = 1, \ldots, m$. For any positive integer d, we represent a model that relies on parameters $\theta \in R^d$ as $\varphi_\theta : X \to Y_a$. This model is a neural network, and θ gathers all of its parameters in all of our tests. We implement a loss function $L : Y \times Y 7 \to R$ for each sample in order to evaluate the model's performance. We redefine the loss as L divided by b mini-batches $B_i \subset \{1, .., m\}$, where $i = 1, .., b$ and $B_i \cap B_j = \Phi$ for $i \neq j$, which is given in Equation (2).

$$l_i(\theta) = \sum_{k \epsilon B_i} L(\varphi_\theta(x^k, y^k) \qquad (2)$$

In each cycle, we take a portion of the mini-batches $U^t \subset \{1, .., b\}$ and divide it into two groups: one for training, denoted as $T^t \subset U^t$, and another for validation, denoted as $V^t \subset U^t$. Here, $T^t \cap V^t = \Phi$ and $T^t \, UV^t = U^t$, respectively. It follows that mini-batches Bi in the training set have an $i \in T^t$, whereas those in the validation set have an $i \in V^t$. Every one of our trials uses a singleton (one mini-batch) validation set, denoted as V^t.

DCNN Classification

In several applications of image processing, including medical image analysis, CNN models have shown to be rather popular. An apparent computer vision difficulty is cervical cancer detection in cervical feature datasets. This is a moderate-level test of deep convolutional neural networks for diagnosis of cervical lesions. By freezing the top layers, the proposed VGG 19 (TL) model may be modified to identify cervical cancer. It is then evaluated using the cervical dataset. To maximize the extraction of targeted cervical cancer characteristics from cervical feature dataset, this proposed a BORFE architecture that combines fundamental benefits of parallel and depth convolutional filter. Two kinds of convolution layers make up the proposed model: one kind is used to extract features from the same input, and the other type is used to replace typical convolution layers at the network's beginning with a single convolution filter. Reducing the overfitting impact involves using several convolutional filters to eliminate the biased sections. Twelve activation layers, five max pooling layers, four cross channel normalization layers, and 15 convolutional layers make up the BORFE-based DCNN model. After this test data have been fed into the training phase, these parameters of the output are calculated. Figure 3 shows two completely linked layers with Softmax classification layers, one of which is based on the design of Google Nets. The other two levels are for altering functionality. Table 1 provides the network description of the BOREF-DCNN model, which includes the convolutional and max-pooling layers.

Table 1. DCNN network description for cancer data classification

Layer No.	**Layer type**	**Stride**	**Filter size**	**FC units**	**No. of filters**	**Output**	**Input**
1	Convolution 1	2 x 2	5 x 5	-	64	64 x 112 x 112	3 x 227 x 227
2	Max-pool_1	2 x 2	3 x 3	-	-	64 x 56 x 56	64 x 112 x 112
3	Convolution 2	1 x 1	1 x 1	-	64	64 x 56 x 56	64 x 56 x 56
4	Convolution 3	1x1	3 x 3	-	128	128 x 56 x 56	64 x 56 x 56
5	Max-pool_2	2 x 2	3 x 3	-	-	128 x 28 x 28	128 x 56 x 56
6	Parallel convolution 1	1 x 1	1x1.3x3.5x5	-	32 064 128	224 x 28 x 28	128 x 28 x 28
7	Max-pool_3	2 x 2	3 x 3	-	-	224 x 14 x 14	224 x 28 x 28
8	Parallel convolution 2	1 x 1	1x1.3x3.5x5	-	32 64 128	224 x 14 x 14	224 x 14 x 14
9	Parallel convolution 3	1 x 1	1x1.3x3.5x5	-	32 64 128	224 x 14 x 14	224 x 14 x 14
10	Max-pool_4	2 x 2	3 x 3	-	-	224 x 7 x 7	224 x 14 x 14
11	Parallel convolution 4	1 x 1	x1.3x3.5x5	-		224 x 7 x 7	224x7x7
12	Max-pool_5	1x1	5 x 5	-	-	224 x 2 x 2	224 x 7 x 7
13	Fully connected 1		-	512			
14	Fully connected 2	-	-	3	-		

Model Parameters

Using an optimized feature dataset, this study employs a two-deep learning model for cervical cancer prediction. A custom-built DCNN architecture and a transfer-learning VGG_19 have been modified for this proposed framework. A single CNN filter type and input data sizes ranging from 1 x 1 to 5 x 5 are used in the conventional neural network architecture. The input data is used by the filter to create a discriminating feature map, which then uses the input data as input. Integrating multiple convolutional filters for extracting discrimination-based multilayer features is the underlying idea behind the construction of multilayer convolutional filters. It uses the same data to expand clusters even farther. In order to extract optimized features, the training period incorporates three distinct kernel sizes: 1x1, 3x3, and 5x 5. A fixed epoch counts of 50 for 64 batch size, BORFE optimization method with a learning rate of 0.0001, and a declining learning rate of 0.01 employing a piecewise approach every ten epochs make up this proposed DCNN design and model parameters. Data is shuffled at each stage before to training in order to provide a normalization impact during training. With each convolutional layer, more discriminative characteristics are retrieved, giving the prediction an additional edge.

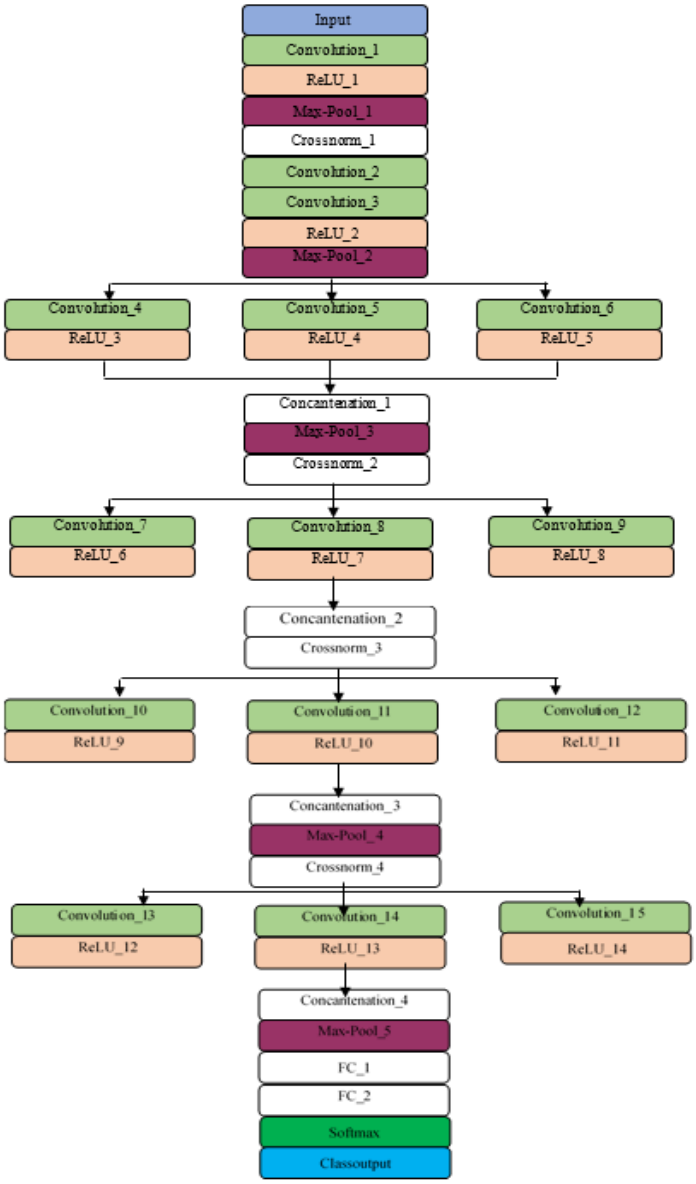

Figure 3. DCNN model for classification

Activation functions are the mathematical equations that determine how well a neural network performs. In order to determine if a neuron in the network should be active ("fired"), its functioning is related to the input relevance with the model prediction. Piecewise linear function $ReLU$ takes an integer as input and returns zero if the input is negative. Because it converges quicker and prevents saturation easily, the $ReLU$ activation is employed. It gets around the issue that logistic regression and the tan hyperbolic function can't provide results for values higher than 1. Every one of the hidden layers makes use of the $ReLU$ activation function. A definition of it is given in Equation (3).

$$f(x) = (0, x) \tag{3}$$

where the neuron receives its input from x. In order to leave the infinite activation function, the ReLU activation function is designed. The features offered by the two kernels are combined using the concatenation layer. To address the issue of overfitting in the model, channel-wise normalization of the activation function is performed using local response normalization after each concatenation layer. One option is to normalize the local response inside the channel, while the other is to do so across the channel. The DCNN based BORFE model normalizes the local responses in a given layer pixel-wise using cross-channel normalization. Equation (4) gives it.

$$x_i = \frac{x_i}{\left(k + \left(\alpha \sum_j x_j^2\right)\right)^{\beta}} \quad (4)$$

The input pixel value is x_i and the hyperparameters k, α and $\beta \in R$ are defined in Equation (2). Following normalizing, the pooling layer A_x is utilized to reduce size. We may reduce the model's computational expenses with the total number of dimensions associated with the attributes retrieved from the convolutional layer of the model by restricting the maximum pooling layer to that channel's highest pixel values using the provided 2 × 2 kernel size. The FC1 layer is followed by a fully connected layer 1 with 128 output nodes and a drop out ratio of 0.5, which is linked after the max-pooling layer 5. Layer 2, which is completely connected and has three output nodes, is linked to a dropout ratio of 0.3% in order to solve the overfitting issue. The probability of each class with respect to the training and validation performance data are produced by the softmax layer. In order to decrease the computational complexity of the model, instead of having 100 to 1000 nodes, the features are output into two classes: Normal, Abnormal. The activation function known as softmax is shown as in Equation (5).

$$f_i(z) = \frac{e^{z_i}}{\sum_g e^{z_g}} \quad (5)$$

This prediction rate is used to compress a vector of randomly assigned real-valued scores from 0 to 1 in Equation (5), which f_i is the i^{th} component of the class scores f and z vector. To find the discrepancy between the expected and actual classes, the classification cross entropy factor is utilized as the cost function. Equation (6) gives categorical cross entropy function.

$$H_p(q) = -\sum_{i=1}^{N} y_i \, .log \; log \; (\hat{y}_i) \quad (6)$$

The y_i appropriate target value and N the number class label are represented by the$(\hat{y}_i)$, i^{th} scalar value in Equation (6), where 0 represents Normal, 1 represents Abnormal. Within the framework of the proposed procedure, we examined the DCNN and VGG 19 models. By modifying the transfer learning procedure, we investigate the VGG 19 model, and we build the BORFE -DCNN from beginning to end up.

Algorithm -BORFE based DCNN Model

Input: Training set T, **and Validation Set** V
Patient p
Initi_features = $\{f1, f2, \ldots fn\}$
Output: Selected Features

Begin
For $initi_{features}$ do
  If $p > 0$ then
    For $Keep_{features}$ do
      $temp_{rm}.append(keep_{feature(j)})$

  $minimize\ w \epsilon u\ \ l_{val(\theta^*(w))}$, $Sparsity\ constraint\ u, selected\ data\ point\ \theta$
      $subject\ to\ \theta^*(w) = arg_\theta min\ l_{tr}(\theta, w)$
      $Score = evaluate_{elimination}(T, V, temp_{rm})$
      $performance_{dick(kepp_{feature(j)}]} = score$
    End

$max_{key} = get_{max_{key}(Performce_{dict})}$
$rm_{list}.appand(max_{key})$
$Keep_{features}.remove\ (max_{key})$
$if\ performance_{dict}(max_{key}) > best_{performance}\ then$
$best_{performamnce} = Performamnce_{dic}(max_{key})$
$Selected_{features} = Keep_{features}$
$increase\ patient\ counter$
Else
$Reduce\ patient\ count$
End
Else
Return $Selected_{features}$
End
End
End

RESULT AND DISCUSSION

This section compares the inferences and simulation outcomes of DCNN-BORFE model with state of art algorithms. Given that the dataset contains four target variables, the reported result is the mean of all classes.

Dataset Details

The 2016 UCI Machine Learning Repository made the cervical cancer (risk factors) data set [33] freely accessible (University of California, Irvine). Patients' actual electronic health records are included in this collection. Eight hundred fifty-eight (858) cases involving patient records were evaluated and found to be positive. Biopsy, Hinselmann, Schlier and Citology are the four target variables. Every instance in the dataset has 32 characteristics that outperform other feature subsets. "Missing values" are present in a few of the cases. It is the decision-makers represented by the four target variables who decide whether the patient has a valid illness diagnosis (1, 'unhealthy' tag) or not (0, 'healthy' tag). A total of 734 samples are included in the UCI data set, with healthy cases making up the majority class and cancer patients creating the minority class. Table 2 shows the dataset with target diseases.

Python 3.11, with its support for scientific computing, simplicity of use, and flexibility, is used for model implementation in this study. The setup is Windows 11, with an i7 processor. Data processing and analysis are made efficient with the help of Python's strong libraries, such as NumPy, Pandas, Matplotlib, and Seaborn. Advanced machine learning libraries such as Scikit-learn, TensorFlow, and Keras are also used for model implementation and analysis.

Table 2. Cervical cancer dataset details

Target	**Cancer Data**	**Healthy**
Hinselmann	**34**	**700**
Biopsy	**51**	**683**
Schiller	**71**	**663**
Citology	**41**	**693**

Evaluation Metrics

A true negative (TNx) is the percentage of cases where the test came back negative, whereas a true positive (TPx) is the number of instances that were properly categorized into the right class. False negatives (FNx) are cases with a well-known positive condition for which the outcome of the test is negative, while false positives (FPx) are the number of instances that were incorrectly predicted under a certain class. Here, x is the size of the dataset.

(I) AUC-ROC: For the purpose of this research, AUC-ROC was used as the performance metric to assess the proposed models. Since there is a problem with uneven grouping, precision is not a good metric to use. Instead, we will evaluate the ROC AUC, which ranges from 0 (very bad) to 1 (excellent), with a 0.5 for arbitrary theory.

(II) Accuracy: the accuracy, denoted as Acc_x , is the percentage of samples that were correctly distributed over the whole class, as given in Equation (7). Acc_x is equal to TPx plus TNx plus FPx plus FNx.

$$Acc_x = \frac{TP_x + TN_x}{TP_x + TN_x + FP_x + FN_x} \tag{7}$$

(III) Recall (Rx): As seen in Equation (8), recall is defined as the proportion of correctly identified instances relative to all instances in that class. The formula in Equation (8) is used.

$$R_x = \frac{TP_x}{TP_x + FN_x} \tag{8}$$

(IV) Precision (Px): A classifier's accuracy may be measured by its precision, which is defined as the percentage of true positives relative to the total number of false positives. Equation (9) shows it mathematically. Three, P_x equals TP_x plus FP_x.

$$P_x = \frac{TP_x}{TP_x + FP_x} \tag{9}$$

(V) F1-Score (F1): The results of a test are shown by the F1 score, which is also called the F-score or the F-measure. The accuracy (P_x) and recall (R_x) of the test are both taken into account when calculating the score. The F1 score is at its peak at 1 (perfect accuracy) and at its worst at 0. Here is the formula for the F1-score in Equation (10). The calculation for F1 is 2 times the product of P_x and R_x, or P_x plus R_x multiplied by 4.

$$F1 = \frac{2 \times P_x \times R_x}{(P_x + R_x)} \tag{10}$$

(VI) Matthew's Correlation Coefficient (MCC): One way to measure how well predictions and observations matchup is using the Matthews correlation coefficient (MCC). The formula for MCC is given by Equation (11). The result of the following equation is the maximum common factor (MCF):

$$MCC = \frac{(TP_x * TN_x) - (FP_x * FN_x)}{(TP_x + FP_x)(TP_x + FN_x)(TN_x + FP_x)(TN_x + FN_x)} \tag{11}$$

Performance Comparison of Feature Extraction

The features included in the dataset utilized for analysis are comprehensively summarized in Table 3. The columns in the table provide information about the features, and each row represents an attribute or variable in the dataset.

Table 3. Actual features in dataset

1	Age
2	Number of sexual partners
3	First sexual intercourse
4	Num of pregnancies
5	Smokes
6	Smokes (years)
7	Smokes (packs/year)
8	Hormonal Contraceptives
9	Hormonal Contraceptives (years)
10	IUD
11	IUD (years)
12	STDs
13	STDs (number)
14	STDs: condylomatosis
15	STDs: cervical condylomatosis
16	STDs: vaginal condylomatosis
17	STDs: vulvo-perineal condylomatosis
18	STDs: syphilis
19	STDs: pelvic inflammatory disease
20	STDs: genital herpes
21	STDs: molluscum contagiosum
22	STDs: AIDS
23	STDs: HIV
24	STDs: Hepatitis B
25	STDs: HPV
26	STDs: Number of diagnoses
27	STDs: Time since first diagnosis
28	STDs: Time since last diagnosis
29	Dx: Cancer
30	Dx: CIN
31	Dx: HPV
32	Dx
33	Hinselmann
34	Schiller
35	Citology
36	Biopsy

Bilevel Optimized Recursive Feature Elimination (RFE) yields features that are most relevant to highlighting important health indicators, especially those related to sexual and reproductive health is illustrated in Table 4. The recognized effects of smoking and hormonal contraceptives on cervical cell alterations and cancer risk prompted their inclusion. Since HPV is a well-established risk factor for cervical cancer, Sexually Transmitted Diseases (STDs) serve as a major factor. Particular illnesses like syphilis, condylomatosis, genital herpes, and HPV are particularly important. It is possible that time-related characteristics, such as the interval between first and final STD diagnosis, indicate the persistence and recurrence of infections, which in turn affect the course of disease. The presence of precancerous or cancerous conditions can be indicated by diagnostic indicators such as cancer diagnosis (Dx), Cervical Intraepithelial Neoplasia (CIN), and Human Papilloma Virus (HPV). Finally, we have provided the results of the routine diagnostic tests for cervical anomalies, which include the Hinselmann, Schiller, and cytology exams. The combination of these elements creates a thorough set of diagnostic, lifestyle, and clinical variables that are essential for health risk assessment predictive modelling, especially in relation to cervical cancer and related disorders.

Table 4. Top selected features from bilevel optimized RFE

1	Smokes
2	Hormonal Contraceptives
3	STDs
4	STDs: condylomatosis
5	STDs: syphilis
6	STDs: genital herpes
7	STDs: HIV
8	STDs: HPV
9	STDs: Time since first diagnosis
10	STDs: Time since last diagnosis
11	Dx: Cancer
12	Dx: CIN
13	Dx: HPV
14	Dx
15	Hinselmann
16	Schiller
17	Citology

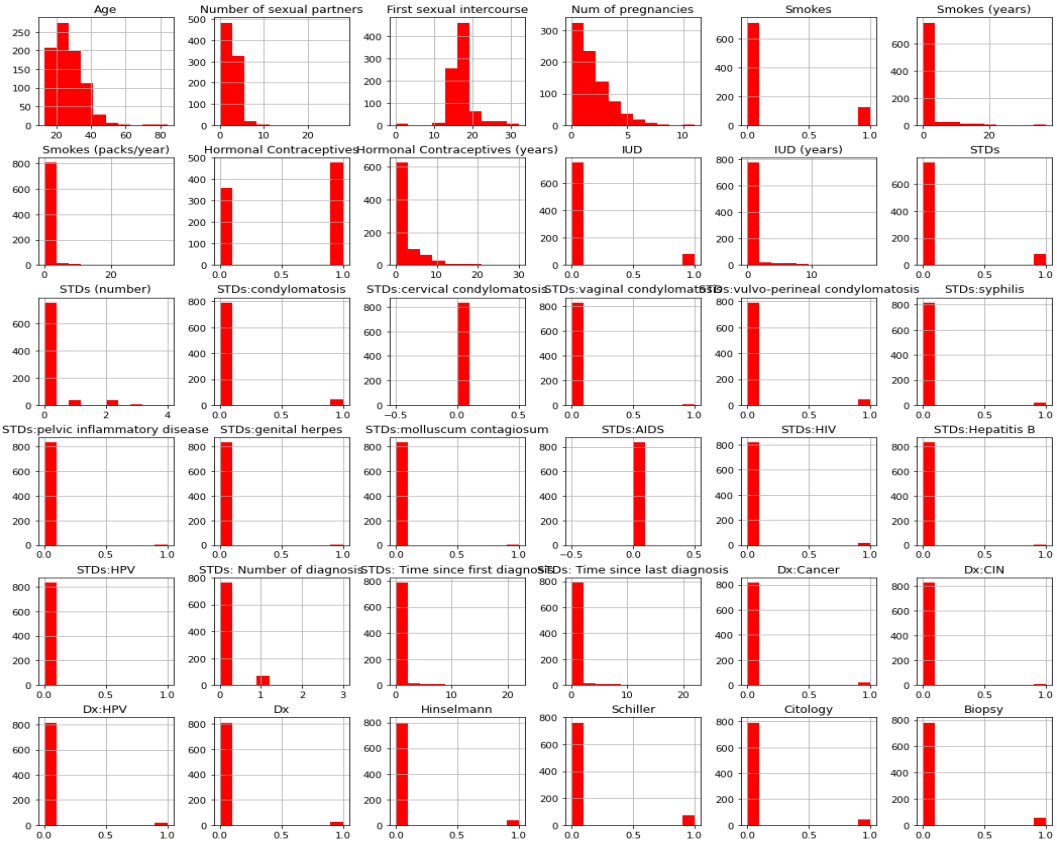

Figure 4. Histogram visualization of cervical cancer dataset description

Insights into the distribution and characteristics of major features in the cervical cancer dataset are provided by the histogram. Figure 4 displays the distribution of all features.

To evaluate BORFE, we used the same methodology as Hastie et al. 2017 [41], which examined several feature selection algorithms, such as Best Subset [42], Forward Stepwise [30], Lasso [8], and Relaxed Lasso [43]. Compared to Relaxed Lasso and Lasso, BOSO produces a sparser model. Part of the reason for this is because we used an information criterion (eBIC) to determine the size of the model. Therefore, compared to Lasso and Relaxed Lasso, BOSO produces regression models with a much lower number of false positives and similar numbers of false negatives (Refer to Figures 5 and 6).

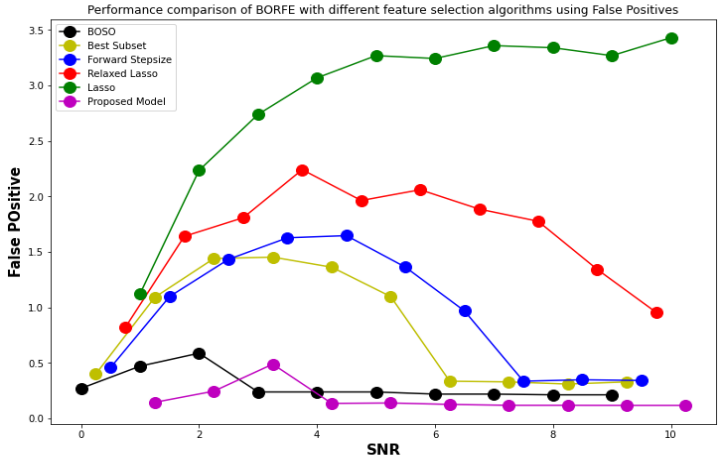

Figure 5. False positive comparison of BORFE with different feature selection methods

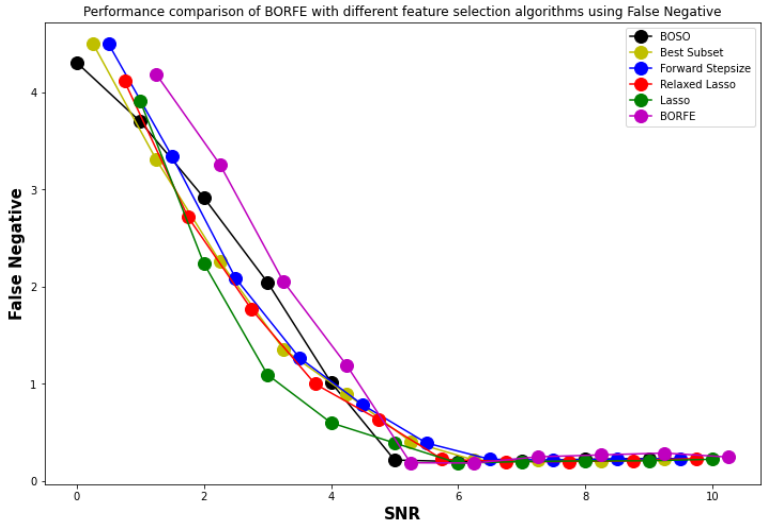

Figure 6. False negative comparison of BORFE with different feature selection methods

Performance comparison of classification

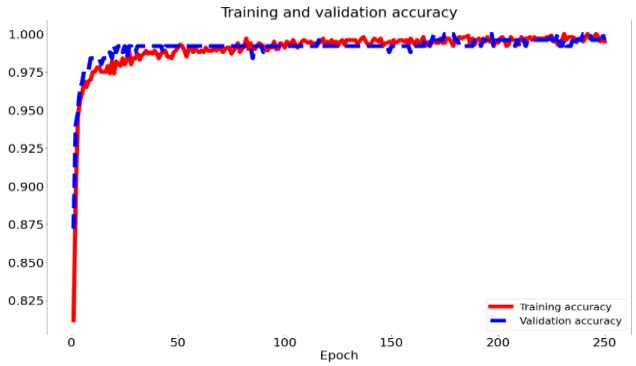

Figure 7. Training and validation accuracy of the proposed model

The accuracy of the proposed model during training and validation is shown in Figure 7, which spans many epochs. Model learning, generalizability, and possible problems like overfitting and underfitting are shown by the two curves that commonly accompany the training accuracy curve and the validation accuracy curve on the graph. Early on in the training process, while the model is still learning the basics

of the data, the training accuracy rises sharply. At the same time, the model may become better at generalizing to new data, which would mean that the validation accuracy will keep increasing up. There is a general trend toward curve stabilization as training advances. A well-generalized model is one in which the difference between the two sets of accuracy, training and validation, remains small.

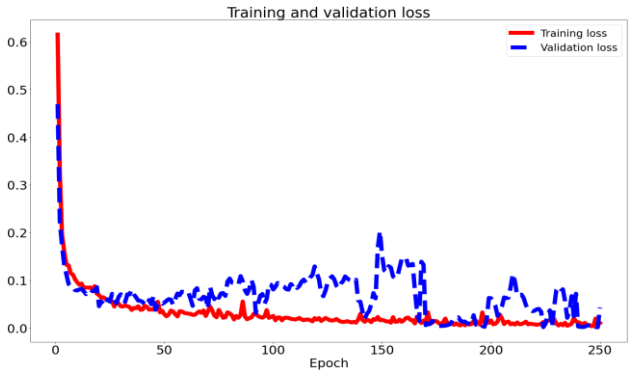

Figure 8. Training and validation loss of the proposed model

Over the course of many iterations, the proposed model's training and validation losses are shown in Figure 8. A smaller loss number indicates greater performance, since it measures the model's prediction inaccuracy. As the model gains knowledge from the data, the loss should go down during training. The ideal situation is for the training and validation loss curves to both decrease and settle at low values, indicating that convergence has taken place. Overfitting occurs when a model has good performance on training data but poor results on unknown data; this happens when the training loss keeps going down while the validation loss stays the same or begins to increase up. On the other side, underfitting occurs when a model fails to capture important patterns in the data, which may be seen when both losses are substantial. When the loss values for the training and validation sets are low and comparable, it means that the model has been well-trained and can generalize well.

Table 5 displays the outcomes that were achieved by applying all 30 characteristics to thoroughly unbalanced data, where the number of normal cases exceeds that of cancer patients. Since even the most precise DL algorithms failed to achieve acceptable results in terms of recall, F1_Score, precision, and MCC when faced with imbalanced data, it is clear that accuracy is irrelevant in such a situation. When analysing the performance of DL algorithms, we only employed balanced data after considering the prediction results achieved on the imbalanced data.

Table 5. Performance comparison of various classifier with full feature data

	AUC	**Accuracy**	**F1**	**Prediction**	**MCC**	**Recall**
RF	0.52	0.94	0.07	0.29	0.09	0.19
KNN	0.5	0.94	0.03	0.07	0.03	0.02
GBC	0.57	0.9	0.19	0.18	0.14	0.21
MLP	0.52	0.93	0.08	0.15	0.07	0.06
Proposed Model	**0.5**	**0.97**	**0.02**	**0.03**	**0.02**	**0.02**

Table 6. Performance comparison of various classifier with BORFE feature selection

Algorithms	**AUC**	**Accx**	**FI**	**Px**	**MCC**	**Rx**
KNN	0.919	0.917	0.922	0.856	0.847	0.772
MLP	0.819	0.82	0.82	0.799	0.642	0.844
RF	0.991	0.991	0.991	0.982	0.982	1
GBC	0.988	0.987	0.987	0.975	0.975	1
Proposed Model	0.996	0.995	0.996	0.994	0.994	1

Table 6 displays the area under the curve (AUC) scores obtained by the DCNN model and other ML methods on two distinct feature sets, one chosen using BORFE and the other using the whole feature set. Boldfaced in each feature set is the model that achieved the highest AUC score. Table 4 shows that the BORFE feature set is having highest performance metric when compared with complete feature set.

Table 7. Proposed model comparison full features vs BORFE feature

Proposed Model	Full Features	BORFE
Accuracy (%)	98.54	99.87
Prediction (%)	98.85	99.88
Recall (%)	99.54	99.67
AUC (%)	99.12	99.18
F1 (%)	99.41	99.51
MCC (%)	99.31	99.78
Time elapse (s)	4.685	3.875

Table 7 details the time it took to train and test the proposed DCNN architecture on both the entire feature set and BORFE feature sets, as well as the prediction scores for each. In comparison to forecasts based on the whole feature set, the BORFE feature set performs better, according to the overall findings. Time is of crucial importance when dealing with huge feature datasets, as opposed to smaller ones.

CONCLUSION AND FUTURE ENHANCEMENT

In recent years, cervical cancer has become a major cause of cancer-related mortality among women. However, by using Deep learning, we can identify the variables that increase the probability of this malignancy developing in females. The objective of this research was to create a new deep learning model for cervical cancer classification utilizing a Deep convolutional neural network (DCNN) and a bilevel optimized recursive feature elimination (BORFE) algorithm for the purpose of cervical cancer prediction. BORFE based feature 11selection aims to eliminate irrelevant or unnecessary features by using a number of criteria. By calculating the relative importance of each element to the desired result, the bilevel optimization use this information to identify the most important features. DCNN classifiers have shown a reasonable level of performance in detecting women showing clinical signs of cervical cancer. They have shown to be exceptionally precise and reliable in their findings. Based on the results of this research, it is clear that using the BORFE method to build classifier models and integrating an ideal feature subset via improved feature selection methodologies may improve the accuracy of cervical cancer detection predictions. These results can be used to make more accurate predictions about other types of gynaecological cancer. In the future, these enhanced features will be used by an efficient classifier to classify cervical cancer.

REFERENCES

[1] World Health Organization (2019) Fact sheet: human-papillomavirus - (hpv)-and-cervical-cancer.
[2] Mustapa M, Rahmah U, Cakranegara PA, Firdaus W, Pratama D, Rahim R. Implementation of Feature Selection and Data Split using Brute Force to Improve Accuracy. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2023;14(1):50-9.
[3] Abdoh SF, Rizka MA, Maghraby FA. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access. 2018 Oct 5;6: 59475-85. https://doi.org/10.1109/ACCESS.2018.2874063
[4] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1; 13:8-17. https://doi.org/10.1016/j.csbj.2014.11.005
[5] Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. Machine learning in RJ Mach. Learn. Res. 2016; 17:5938-42.
[6] Mohandas DR, Veena DS, Kirubasri G, Mary IT, Udayakumar DR. Federated Learning with Homomorphic Encryption for Ensuring Privacy in Medical Data. Indian Journal of Information Sources and Services. 2024;14(2):17-23.
[7] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002 Jan; 46:389-422. https://doi.org/10.1023/A:1012487302797
[8] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996 Jan; 58(1):267-88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[9] Elshrkawey M, Alalfi M, Al-Mahdi H. An enhanced intrusion detection system based on multi-layer feature reduction for probe and dos attacks. Journal of Internet Services and Information Security (JISIS). 2021 Nov;11(4):40-57.
[10] Ong CS, Smola A, Williamson R. Learning the kernel with hyperkernels.

[11] Caruana R. Multitask learning. Machine learning. 1997 Jul; 28:41-75. https://doi.org/10.1023/A:1007379606734
[12] Evgeniou T, Pontil M. Regularized multi--task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining 2004 Aug 22 (pp. 109-117). https://doi.org/10.1145/1014052.1014067
[13] Hastie T, Rosset S, Tibshirani R, Zhu J. The entire regularization path for the support vector machine. Journal of Machine Learning Research. 2004;5(Oct):1391-415.
[14] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002 Jan; 46:389-422. https://doi.org/10.1023/A:1012487302797
[15] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research. 2003;3(Mar):1157-82. https://doi.org/10.1023/A:1012487302797
[16] Vapnik V. Statistical learning theory. John Wiley & Sons google schola. 1998; 2:831-42.
[17] Trivedi J, Devi MS, Solanki B. Step towards intelligent transportation system with vehicle classification and recognition using speeded-up robust features. Archives for Technical Sciences/Arhiv za Tehnicke Nauke. 2023 Jan 1(28). http://dx.doi.org/10.59456/afts.2023.1528.039J
[18] Lavanya D, Rani DK. Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE). 2011 Oct;2(5):756-63.
[19] Latha DS, Lakshmi PV, Fathima S. Staging prediction in cervical cancer patients–a machine learning approach. International Journal of Innovative Research and Practices. 2014;2(2):14-23.
[20] Akyol K. A Study on Test Variable Selection and Balanced Data for Cervical Cancer Disease. International Journal of Information Engineering & Electronic Business. 2018 Sep 1;10(5). https://doi.org/10.5815/ijieeb.2018.05.01
[21] Devi G, Kavitha M, Surendar A. High Speed Image Searching for Human Gait Feature Selection. International Journal of communication and computer Technologies. 2016;4(2):88-95.
[22] Arora G. Desing of VLSI Architecture for a flexible testbed of Artificial Neural Network for training and testing on FPGA. Journal of VLSI circuits and systems. 2024;6(1):30-5. https://doi.org/10.31838/jvcs/06.01.05
[23] Chicco D. Ten quick tips for machine learning in computational biology. Bio Data mining. 2017 Dec 8;10(1):35. https://doi.org/10.1186/s13040-017-0155-3
[24] Chicco D, Rovelli C. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. PloS one. 2019 Jan 10;14(1). https://doi.org/10.1371/journal.pone.0208737
[25] Rekha G, Tyagi AK, Reddy VK. A wide scale classification of class imbalance problem and its solutions: a systematic literature review. Journal of Computer Science. 2019;15(7):886-929.
[26] Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of biomedical informatics. 2019 Feb 1; 90:103089. https://doi.org/10.1016/j.jbi.2018.12.003
[27] Kashif R. A Compact Circular Polarized Antenna for Fixed Communication Applications. National Journal of Antennas and Propagation. 2019 Mar 3;1(1):1-4. https://doi.org/10.31838/NJAP/01.01.01
[28] Karegowda AG, Jayaram MA, Manjunath AS. Feature subset selection problem using wrapper approach in supervised learning. International journal of Computer applications. 2010 Feb;1(7):13-7.
[29] Asl TM, Asl TS. Strategy optimization for responding to primary, secondary and residual risks considering cost and time dimensions in petrochemical projects. Archives for Technical Sciences/Arhiv za Tehnicke Nauke. 2022 Jul 1(27). https://doi.org/10.59456/afts.2022.0227.033t
[30] Efroymson M. Stepwise regression–a backward and forward look. In Eastern Regional Meetings of the Institute of Mathematical Statistics 1966 Apr 27 (pp. 27-9).
[31] Abinaya R, Vidhya S, Vadivel S. Latent Palm Print Matching Based on Minutiae Features for Forensic Applications. International Journal of communication and computer Technologies. 2014;2(2):85-7.
[32] Parikh D, Menon V. Machine learning applied to cervical cancer data. Int. J. Math. Sci. Comput. 2019 Jan 5;5(1):53-64. https://doi.org/10.5815/ijmsc.2019.01.05
[33] Fernandes K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. InPattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8 2017 (pp. 243-250). Springer International Publishing. https://doi.org/10.1007/978-3-319-58838-4_27
[34] Bi J, Bennett K, Embrechts M, Breneman C, Song M. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research. 2003;3(Mar):1229-43.
[35] Sihag V, Vardhan M, Singh P, Choudhary G, Son S. De-LADY: Deep learning based Android malware detection using Dynamic features. J. Internet Serv. Inf. Secur. 2021 May 31;11(2):34-45. https://doi.org/10.22667/JISIS.2021.05.31.034
[36] Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semidefinite programming. Journal of Machine learning research. 2004;5(Jan):27-72.
[37] Park M, Kim S, Kim J. Research on Note-Taking Apps with Security Features. J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl. 2020 Dec;11(4):63-76. https://doi.org/10.22667/JOWUA.2020.12.31.063

[38] Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Applied Sciences. 2019 Jun; 1:1-6. https://doi.org/10.1007/s42452-019-0645-7

[39] Camgözlü Y, Kutlu Y. Leaf Image Classification Based on Pre-trained Convolutional Neural Network Models. Natural and Engineering Sciences. 2023 Dec 1;8(3):214-32. https://doi.org/10.28978/nesciences.1405175

[40] Choudhury A, Wesabi YM, Won D. Classification of cervical cancer dataset. arXiv preprint arXiv:1812.10383. 2018 Dec 11. https://doi.org/10.48550/arXiv.1812.10383

[41] Hastie T, Tibshirani R, Tibshirani RJ. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692. 2017 Jul 27. https://doi.org/10.48550/arXiv.1707.08692

[42] Bertsimas D, King A, Mazumder R. Best subset selection via a modern optimization lens. Ann Stat. 2016;44(2):813-852. https://doi.org/10.1214/15-AOS1388

[43] Meinshausen N. Relaxed lasso. Computational Statistics & Data Analysis. 2007 Sep 15;52(1):374-93. https://doi.org/10.1016/j.csda.2006.12.019

[44] Momma M, Bennett KP. A pattern search method for model selection of support vector regression. In Proceedings of the 2002 SIAM International Conference on Data Mining 2002 Apr 11 (pp. 261-274). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972726.16