International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 576-589 March 2026-Sept 2026 www.csjournals.com

Optimizing Multi-Class Cervical Cancer Diagnosis: A Hybrid Approach Integrating Filter-Based Feature Selection with Feedforward Neural Networks

Mehboob Ali[1], Marziya Arman[2]
[1]Department of Computer Science and IT, University of Ladakh, Kargil, India
[2]Department of Computer Science and IT, University of Jammu, Jammu, India

Abstract

In clinical medical diagnostics, the classification of uterine cervix cancer using morphological cell parameters presents a significant challenge due to highly overlapping feature values in morphologically similar cells. Monolithic neural networks often struggle to map this complex problem space efficiently due to the presence of irrelevant, redundant, or noisy attributes. To address this limitation and mitigate the curse of dimensionality, this study proposes a hybrid neural network architecture that integrates advanced feature selection techniques with feedforward neural network classifiers. Three prominent feature ranker algorithms—Neighborhood Component Analysis (NCA), Laplacian Score, and Relief—were implemented to reduce dimensionality while preserving critical discriminative information. By analyzing the structural intersections of these feature selectors, four optimized low-dimensional datasets (D_1, D_2, D_3, and D_4) were engineered. Ten distinct neural network training algorithms were then systematically trained and evaluated on these datasets using a robust 10-fold cross-validation framework. Empirical results demonstrate that the hybrid configuration combining the Laplacian and Relief feature selection algorithms (Dataset D_3) with the Levenberg-Marquardt optimization training algorithm yields the highest classification accuracy of 84.8%. This represents a notable 4% improvement over a conventional, monolithic Levenberg-Marquardt model. While the proposed hybrid framework significantly enhances classification potential, reduces training time, and minimizes overfitting, detailed error analysis reveals that normal-category cells (Class 2) remain the most frequently misclassified. These findings underscore a strong foundation for hybrid CAD (Computer-Aided Diagnosis) systems while highlighting critical pathways for future optimization to meet strict medical validation thresholds.

Keywords: Cervical Cancer Classification, Hybrid Neural Networks, Feature Selection, Laplacian Score, Relief Algorithm, Levenberg-Marquardt.

1. Introduction

Cervical cancer occurs when the cervix tissue cells grows and replicates abnormally because of uncontrolled cell division and cell death. This malignant tumor because of uncontrolled cell division forms the cervical cancer. In any type of cancer human body is not able to manage these affected cells for normal usual function rather these cells gets transform into tumor. The tumor if becomes malignant spreads through the blood stream to other parts of the body and thus infects

those parts also. Cervical cancer takes good number of years to develop. The infected cells are called cervical intra-epithelial neoplasia (CIN) also called cervical dysplasia. Therefore any cell on the surface of the cervix if shows any unusual change and may be precursor of cancer are called CIN. Normally in most of the cases the human immune system resolves CIN and do not causes the CIN to show any symptom. But unfortunately in some cases these lesions may develop in to cancer if not treated on time. But this progression to a full cancer from CIN takes many years. Therefore timely and precise diagnosis of cervical cancer is very important. This cancer has become one of the main causes of death among women after breast cancer worldwide and therefore it has become one of the contentious issue of research before the world's scientific community. The main aim of scientific research in cervical cancer aims to have an early diagnostic system so that the cancer may get caught at a very early stage. Among the most common cancer affecting women cervical cancer stands at the fourth place and stands at seventh place among all the common cancers worldwide. As per the reports of WHO death toll of cervical cancer in 2012 was 52800 and about 84% of these cases are reported from poor and developing countries of the world. The main reason of this huge burden of cervical cancer in developing countries is attributed to poor access to medical care and absence of screening and treatment services (Kent A, 2010). Etiology of cervical cancer is still not clear. Medical experts still believe that there is not a single dominant cause of cervical cancer. The only means of avoiding this cancer is early detection and if the cancer is detected early before spreading to other organs the survival rate of patient is more than 97%. Medical research have found that this cancer is caused by a virus called human papillomavirus (HPV) transmitted sexually. About more than 120 types of human papillomavirus (HPV) are reported in medical literature (Chaturvedi Anil et al., 2010). Out of 120 viruses 15 are classified as high risk types (16, 18, 31, 33, 35, 39, 45, 51, 52,) 3 as probably of high risk (26, 53, and 66), and 12 as low-risk (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and CP6108) (Munoz Nubia et al., 2003). After the HPV virus infects a normal cell the immune system of cell removes the virus but in some cases the virus is not done away by the immune system. The virus then affects the genetic information of the cell causing it to do uncontrolled cell division and growth which causes cervical cancer. There are many factors which causes cervical cancer intercourse and pregnancy at early age, multiple sex partners, having weak immune system, smoking, poor menstrual hygiene etc. normally the early stage of cervical cancer are asymptomatic and with progression of cancer to advanced stages the symptoms starts rapidly appearing. Some common symptoms of cervical cancer includes bleeding from vagina abnormally, severe pain during vaginal intercourse

2. Materials and Methods

2.1. Feature Selection Techniques

Medical datasets usually contains large number of attributes/features and among all these features some features do not add any weightage too the information value of the dataset. Such features are nearly irrelevant or can be considered as redundant. Sometimes such features are

considered as noisy attributes. Because of the presence of such irrelevant and noisy attributes in the dataset any algorithm applied fails to find an optimal decision boundary for classification. On such datasets the machine learning algorithms always fails. Therefore it is reported in literature that the performance of machine learning algorithms can be enhanced if an optimal set of feature is obtained for the dataset by removing the noisy and redundant features. (Cheng, Wei and Tesang) (Inbararani, Azar and Jothi, 2014). Feature selection do include a set of techniques that are applied on a dataset to produce an optimal set of features in order to enhance the classification potential of machine learning algorithms. The presence of unnecessary, redundant and noisy features makes the learning process of machine learning algorithm very slow and error prone. Before applying the machine learning algorithm feature selection algorithms are applied in order to reduce the dimensionality of dataset by removing those redundant features. After obtaining an optimal feature set a new dataset is created from the original database containing only the features as selected by the feature selection algorithm. The aim of these feature section techniques are to find the best subsets of features for making the neural network learn the problem space more efficiently. The concept of feature selection was there from 1924 as reported in the literature. During those days the concept of feature selection were applied to reducing the number of variables in a linear regression problems and mostly used the statisticians. Emergence of machine learning coupled, advancement in computer hardware and need of more complex data pushed the emergence of this field of feature selection. Feature selection technique not only reduces the number of features in a dataset but it increases the generalization ability of a base classifier by reducing the curse of dimensionality. In addition it also network training time and enhances the models' interpretability. The feature selection algorithms have the following objectives.

1. Improves the performance of a machine learning algorithm by reducing the chance of over fitting
2. Making the datasets simple and easy to visualize so that the machine learning algorithm are able to understand the hidden patterns.
3. Having a less dimensional dataset reduces the time required to train and test a model. Therefore both training and testing reduces.
4. To get rid of the curse of dimensionality problem.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 576-589 March 2026-Sept 2026 www.csjournals.com

2.2. Methods of Feature Selection.

The methods of feature selection are generally divided into the following three categories.

1. Filter Method
2. Wrapper Method
3. Embedded Method

1.Filter Method

The filter methods are applied as a preprocessing step in machine learning. The filter approaches are used independent of the learning algorithm and are applied with the application of statistical and analytical techniques. In such methods the assumptions of the learning classifier are not included, therefore these feature selection techniques usually are useful in exploring the hidden and underlying relationship between the features. The filter methods are faster in speed but the classification accuracy of such algorithms are usually are lower than wrapper methods. These filter methods are also applied as a preprocessing step for the wrapper method. Fig.1. shows the steps involved in filter method of feature selection.

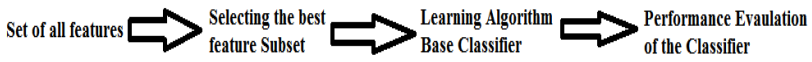

Fig.1. Steps Involved in Filter Method feature selection

2.Wrapper Method

The wrapper method is dependent on a base classifier to find the optimal subset of features. Every new subset is applied to train abase classifier. The learned classifier is then tested on the test set to find the optimality of the subset. Error rate of the classifier on the test set determines the score for the subset under test. These methods are very slow as compared to filter method because a base classier needs to be trained for determining the best subset of feature. Forward selection, backward selection and recursive elimination are best examples of wrapper method. In forward selection the feature set is started as an empty set a feature is added to this set if and only if it increases the classification accuracy. In each step of this algorithm a feature is added the classification result of the resulting feature set is noted and if the algorithm produces greater accuracy after the addition of the feature then the feature is kept in the feature set otherwise it is removed. The reverse procedure is applied for backward selection algorithm. In backward selection the feature set is started from a full set and the features are eliminated one by one from

the set and if the removal of a features enhances the classification result the feature is discarded otherwise the feature is kept and a new feature is tested. In this way after checking with all the features the final feature set is created. The recursive method is a greedy technique which aims to find the best feature subset in terms of classification accuracy. On each iteration the recursive method created models and selects aside the top and worst performing features. The next model is constructed on the left out features and the process continues until all the features are exhausted. The final feature ranking is done as per the order of their elimination.

3.Embedded methods

The embedded methods lies computationally between the filter and wrapper method. This methods uses base classifiers having their own feature selection module. The examples of embedded method includes mimetic algorithm, regularized trees, Lasso, RIDGE etc. Both LASSO and Ridge regression have built-in ability to reduce the over fitting through a penalization method. LASSO performs the L1 regularization which penalizes as per the magnitude of coefficient. Ridge regularization performs the regularization by adding penalty equal to square of the magnitude of coefficients.

In this experiment three well known feature selection techniques are applied to reduce the dimensionality of the dataset and to get the best subset of feature.

2.3. Neighborhood Component Analysis (NCA)

This algorithm uses the diagonal adaptation to determine the feature weights. The algorithm is very efficient for supervised models based on pairwise distance between observations to get the final response. This algorithm is non-parametric method for selecting the optimal set of features in order to increase the prediction accuracy of classification algorithm. This algorithm uses the regularization technique to calculate the weight of features as per their importance to minimize an objective function which calculates the average leave one out classification over the training data. This algorithm maximizes the leave one out classification accuracy by using the gradient ascent technique. The experimental analysis carried out do conclude that this algorithm is very insensitive to the increase in number of irrelevant and redundant data therefore performs better than some other algorithms with the same objective. In NCA the regularization parameter is tuned to correctly detect the relevant features in the data. When all the feature weights are close to zero this indicates that value of regularization parameter is too large and when the regularization parameter reaches to infinity all the feature weights tends to zero. Therefore in

neighborhood component analysis the tuning of regularization parameter is very important in order to detect and identify the most relevant and important features. The best value of regularization parameter will produce the minimum possible classification error/loss.

2.4.Laplican Algorithm

The Laplican algorithm selects the best features by ranking them according to their laplican score. The Laplican score is calculated from nearest neighbor similarity graph. The Similarity graph represents the local neighborhood relationships among the data points in form of an undirected graph. Nodes of the graph represents the data points and undirected edges corresponds to the connection between the data points. If the distance between two data points is positive and is above some threshold than the two points are connected using an edge in the similarity graph. This similarity graph is constructed based on the nearest neighbor. The similarity graph is finally stored in a similarity matrix and contains the pairwise similarity between two data points. This matrix is also an adjacency matrix and is also symmetric because the edges of similarity graph are undirected. A value of 0 at any index means that the two nodes are not connected in the similarity graph. This similarity matrix is used to calculate the laplican matrix (He, X., D. Cai, and P. Niyogi 2005) . The following steps are followed in to calculate the Laplican Matrix

Step 1. A local neighborhood is defined for each data point using the nearest neighbor method and the pairwise distance $Dist_{i,j}$ for all points i and j in the neighborhood is computed.

Step 2. The value of similarity matrix are converted using the kernel transformation method

$$S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$$

Where σ represents the scale factor for the kernel

Step 3. Each feature is centered by removing its mean as given by

$$\tilde{x}_r = x_r - \frac{x_r^T D_g \mathbf{1}}{\mathbf{1}^T D_g \mathbf{1}} \mathbf{1},$$

Where x_r is the r^{th} feature, D_g is the degree matrix.

Step 4. Compute the score of each feature using the equation given as

$$s_r = \frac{\tilde{x}_r^T S \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r}.$$

Step 5. From the above equation the Laplican Matrix can be computed as

$$L_r = \frac{\tilde{x}_r^T L \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r} = 1 - \frac{\tilde{x}_r^T S \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r},$$

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 576-589 March 2026-Sept 2026 www.csjournals.com

where L_r is the Laplacian matrix

This laplican matrix is used to rank the features according to their importance and from the full feature set a reduced feature subset is selected representing the ground truth in a more better way and making the classifier learn the problem space in a better and more ease way.

2.5.Relief Algorithm

This algorithms computes the weights of features in scenarios where the class is a multiclass and categorical. This algorithm penalizes the predictors which produces different values for the neighbors of a same class and rewards those predictors which results different values for the different classes**.** (Kononenko, I., E. Simec, and M. Robnik-Sikonja,1997), At the onset this algorithm sets the weight of all the predictors to zero. Then the algorithm iteratively selects any observation randomly and finds the k-nearest observations to the selected observation for each class. The weights are updated for each nearest neighbour for a given predictor as $W_j^i = W_j^{i-1} - \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq}$ if x_r and x_q do belong to the same class and given as $W_j^i = W_j^{i-1} + \frac{p_{y_q}}{1 - p_{y_r}} \cdot \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq}$ if x_r and x_q do belong to the different classes. Where W_j^i is the weight of the feature F_j at the i^{th} iteration. is Prior probability of the class to which X_r *is given by* P_{y_r} and P_{y_q} is the probability of class to which X_q belongs. The number of iterations is specified by m. The relief algorithm also works for regression problems (Robnik-Sikonja, M., and I. Kononenko, 1997).

Relief has actually a family of algorithms and are proven to be very successful feature estimator. The family of relief algorithms uses conditional dependencies among the features and produces a unified idea of feature selection in classification and regression problems. These algorithms are widely known as feature selection techniques where as they are also used in preprocessing stages before the model is trained (Marko RS and Igor K, 2003)

3. Literature review with respect of application of neural networks for detection of Cancer

As computational power increased and larger datasets became available, the literature transitioned from shallow neural networks and traditional classifiers to deep learning paradigms. Specifically, Convolutional Neural Networks (CNNs) have revolutionized image-based cancer detection.

Esteva et al. (2017) demonstrated a major milestone by training a deep CNN on a dataset of 129,450 clinical images consisting of over 2,000 different diseases. The network was tested against 21 board-certified dermatologists on biopsy-proven clinical images representing the two most common skin cancers: keratinocyte carcinomas and malignant melanomas. The CNN achieved dermatologist-level classification accuracy, proving that deep neural networks could match human expert performance in visual diagnostic tasks.

Hammond et al. (1998) applied machine learning techniques to the prognosis of oral cancer, demonstrating that these automated tools perform on par with medical specialists in screening tasks. Although the incidence rate of oral cancer is relatively low, its prognosis becomes exceptionally severe if left untreated. Because of this low incidence rate, widespread manual population screening is not logistically or financially viable. Consequently, computer-aided automated systems can serve as crucial decision-support tools for healthcare professionals, enabling them to identify high-risk cases and recall patients for targeted mucosal re-examinations and lifestyle counseling.

Sarkar et al. approached breast cancer diagnosis as a pattern recognition problem within computer science. Utilizing the publicly available *Wisconsin-Madison Breast Cancer Database* to train and evaluate various machine learning models, the authors implemented a K-Nearest Neighbor (K-NN) classifier. The K-NN algorithm demonstrated superior classification performance compared to other contemporary machine learning techniques evaluated for the same task, outperforming the leading benchmark solutions by 1.17%.

Delen et al. (2005) designed a predictive framework to estimate the survivability rates of breast cancer patients. To construct this automated system, they evaluated two prominent data mining techniques—artificial neural networks (ANNs) and decision trees—alongside traditional logistic regression. The models were trained and tested using a massive secondary dataset consisting of 200,000 cases. To guarantee an unbiased performance evaluation, the authors utilized a robust 10-fold cross-validation technique, whereby the dataset was partitioned into ten equal subsets,

with nine subsets used for training and one subset reserved for validation over ten iterative runs. The comparative analysis revealed that the C5 decision tree algorithm achieved the highest predictive accuracy, followed by artificial neural networks and logistic regression, respectively, providing valuable insights into the predictive capabilities of different algorithmic approaches.

4. Design and Implementation of Hybrid Neural Network Architecture

In machine learning algorithms the higher dimension always poses a challenge to understand and get the hidden patterns of the dataset. Therefore reduction of dimension is always an important task to make these algorithm learn the problem space. As in the first experimental study it is empirically shown that a single neural network is not able to learn the problem space at rate which the problem demand. For example in medical diagnostic tools the misclassification rate should be below 1% so that the false positive and false negative rate becomes minimum. Therefore in this experimental study hybrid neural network architectures are evaluated empirically for their classification ability. Three feature selection techniques are applied to reduce the dimensionality of the dataset in order to achieve a better classification result as compared to monolithic neural networks. The three feature selection individually works on the full dataset and tries to find the best discriminative feature set. The features which are ranked well are considered as the good features and low ranked feature mean the feature has low contribution in deciding the class of an instance. In order to get a robust and more accurate subset of feature, the feature set of all the neural network training algorithms are evaluated thoroughly. The original dataset is divided into four full datasets with less number of features. The first dataset is created by selecting features which in intersection of all the three feature selection algorithms. Those features which are ranked in top twenty by all the three algorithms are selected. The first dataset therefore contains only the top 20 ranked feature by all the three algorithms. The second database is created by including the features which are at the intersection of all the three feature section algorithms and those feature which are at the intersection of the NCA and Laplican Score algorithm.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 576-589 March 2026-Sept 2026 www.csjournals.com

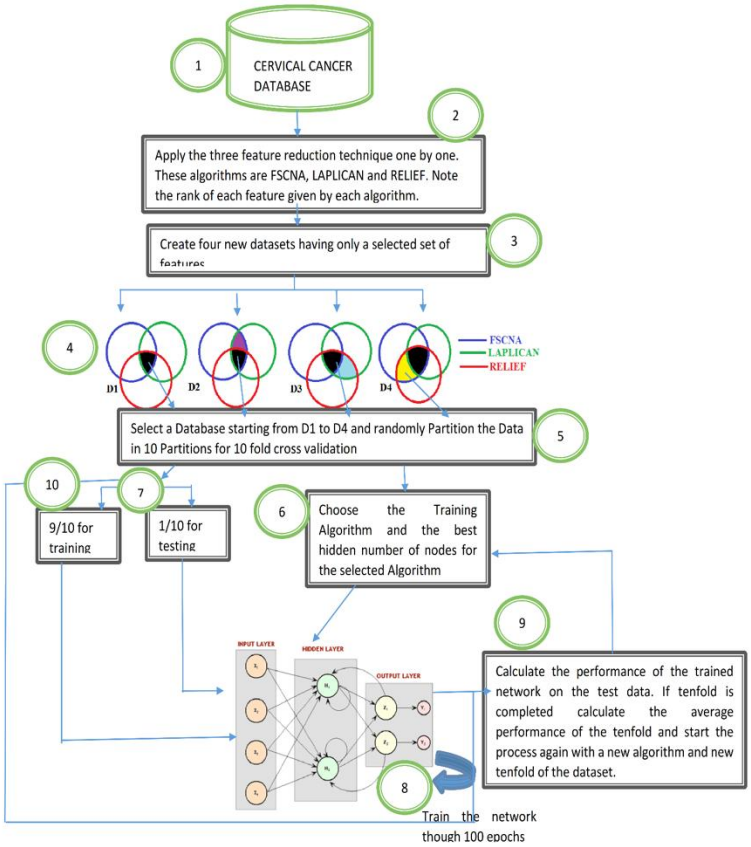

Fig.2. Hybrid Neural Network Architecture

That means this database selects the features which are ranked important by all the three algorithm and those features which are ranked top by both the laplican score and NCA algorithms. The third dataset is created by including the features lying at the intersection of all three algorithms and the features at the intersection of Laplican and Relief. Similarly in the fourth dataset the features included are those which are at the intersection of all the three and at the intersection of NCA and Relief. In all the four datasets the focus is laid to create a dataset which includes feature set from all the three algorithms. In this way the feature set so created is well crafted set and does include only the features which are having strong discriminative power. These algorithms makes it sure that the redundant and irrelevant features are discarded and are not included in the final feature set in order to reduce the curse of dimensionality problem. A Matlab code is written to implement all the three algorithms and to create all the four datasets. A separate module is included in the code which handles the ranking output of algorithm and sorts the original dataset based on the highest ranked features as per the

algorithms. After creating the four datasets these datasets are then used to train the three top feed forward neural network algorithms. The four datasets represents four feature subsets and the objective is to find the best subset among all the three. In this way this experiment is double checking the quality of feature subset. At the first level each individual feature selection technique makes the best feature selection and again the quality of feature subset is checked through the classifier. All the four datasets are used to train neural networks and the top performing algorithm and the feature subset is indicated and evaluated. In this experiment the dataset is selected a training algorithm is selected with the best number of hidden nodes. The dataset is divided into 10 partitions to have a tenfold cross validation. After checking the performance of an algorithm over the said dataset second algorithm with its reported best number of hidden nodes is applied on the said dataset. In this way all the algorithms are applied over all the datasets. A Matlab code did all these managing of applying all algorithms over all the datasets through tenfold cross validation. A hybrid architecture in which the neural network is assisted a feature selection technique is finally decided and selected based upon the duo of the best feature set and the best performing algorithm.

5. Results and Discussions.

The classification result of all the algorithms over the four datasets is shown in the fig.3. in the result it clearly depicts that dataset D3 with Levenberg Marquardt training algorithm performs the best and produces a classification result of 84.8% about 4% enhancement over the previous monolithic Levenberg marquardt algorithm thus clearly showing that this hybrid architecture consisting of feature selection algorithm Laplican + Relief and the Levenberg Marquardt algorithm produces the good result among all the others. The feature of dataset D3 is therefore selected for further analysis and all the other features are removed.

The classification accuracy also showas that all the other training algorithms also performs good over the Database 3. Therefore the feature set of Database3 is the ideal feature set among all the feature sub sets created.

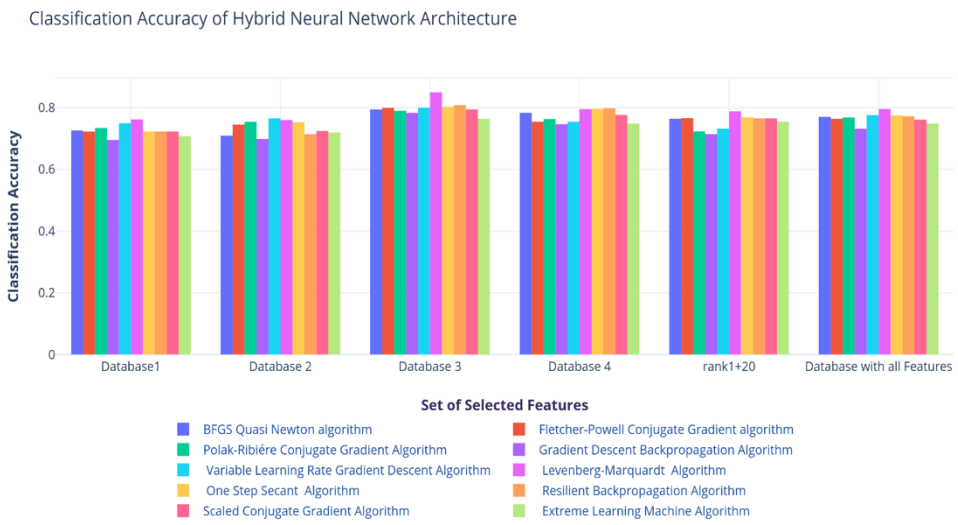

Fig.3. Result of all algorithms over the four datasets

Classification of the three performing algorithms are also shown in form of Confusion Matrix from Fig 4 to Fig 6. The confusion Matrix of Levenberg Marquardt algorithm shows that among all the classes class 2 is the most misclassified class and Class 7 is the most best classified class. The second least classified class is class 5. Out of the all the instances of class 2 16 instances are misclassified in the category of Class 5, this indicated that a normal category cell is wrongly classified as abnormal category cell. Similarly out of total instances of class 5, 19 instances are misclassified into Calss 4 cell thus increasing the false negative rate.

Both variable length learning rate gradient descent and one step secant algorithm performs very poor for the Class 2 Instances and produces good classification results for Class 7 instances. The highest classification accuracy of variable learning arte gradient descent is 81.6% whereas that of one step secant is 80.6%. these experimental results suggests that for seven class classification still there is huge scope of progress because the best performing algorithm on the best feature subset produced a classification accuracy of 84.8% which mean we are still 15.2% away to be an ideal and medically robust classifier. In the succeeding experiments are carried out to further enhance the classification of the system in order to developed a very intelligent neural network based diagnostic tool

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 576-589 March 2026-Sept 2026 www.csjournals.com

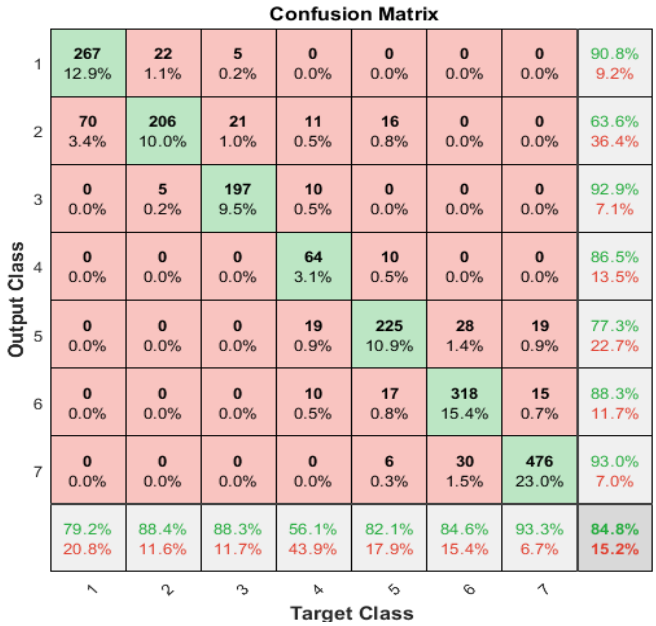

Fig.4. Shows the Confusion Matrix of Levenberg Marquardt Algorithm trained on the feature subset created by the Laplican and Relief Algorithm

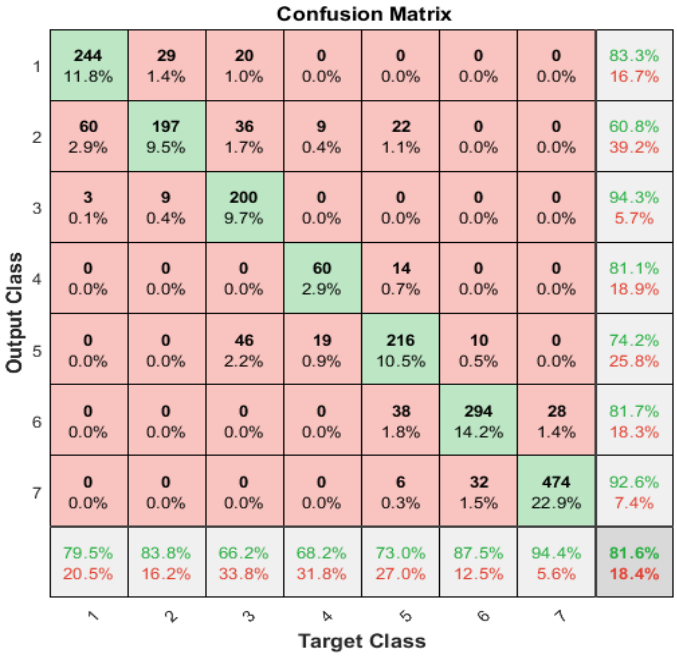

Fig.5. Shows the Confusion Matrix of Variable Learning rate Gradient Descent Algorithm trained on the feature subset created by the Laplican and Relief Algorithm

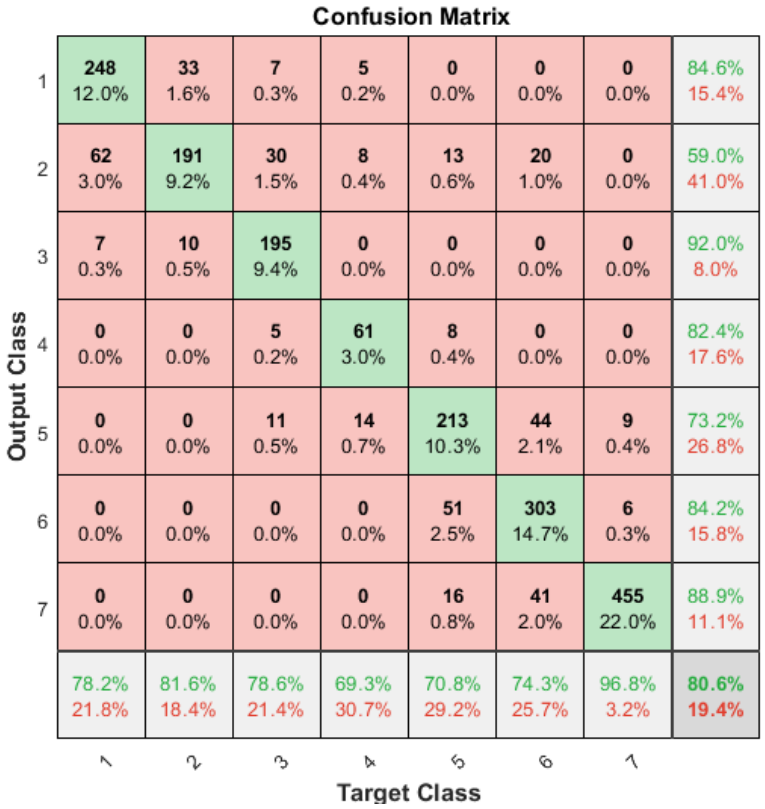

Fig.6. Shows the Confusion Matrix of One Step Secant Algorithm trained on the feature subset created by the Laplican and Relief Algorithm

References:

1. Cheng, T.H, Wei, C.P and Tesang, V.S., (2006), "Feature Selection for Medical Data Mining: Comparison of Expert Judgement and automatic approaches", 19th IEEE Symposium on Computer based Medical Systems (CBMS 06),IEEE,pp.165-170.
2. Inbarani, H. H., Azar, A.T. and Jothi, G., (2014), "Supervised Hybrid Feature Selection based on PSo and Riugh sets for medica diagnosis", Computer methods and programs in biomedicine, 113 (1),pp 175-185
3. He, X., D. Cai, and P. Niyogi. "Laplacian Score for Feature Selection." NIPS Proceedings. 2005.
4. Kononenko, I., E. Simec, and M. Robnik-Sikonja. "Overcoming the myopia of inductive learning algorithms with RELIEFF." Retrieved from CiteSeerX: (1997).
5. Robnik-Sikonja, M., and I. Kononenko. "An adaptation of Relief for attribute estimation in regression." Retrieved from CiteSeerX:, (1997).
6. Marko RS, Igor K: , "Theoretical and empirical analysis of relief and rreliefF" ,. Machine Learning Journal 2003, 53:23-69
7. Hammond, Peter, and Paul M. Speight. "Screening for Risk of Oral Cancer and Pre- cancer." Intelligent Data Analysis In Medicine and Pharmacology (1998)
8. Sarkar, Manish, and Tze-Yun Leong. "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem." Proceedings of the AMIA Symposium. American Medical Informatics Association, 2000.
9. Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34.2 (2005): 113-127.